"Person, Shoes, Tree. Is the Person Naked?" What People with Vision Impairments Want in Image Descriptions

Abigale Stangl
School of Information
University of Texas at Austin
Austin, TX USA
stangl@utexas.edu

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

Danna Gurari
School of Information
University of Texas at Austin
Austin, TX USA
danna.gurari@ischool.utexas.edu

ABSTRACT
Access to digital images is important to people who are blind or have low vision (BLV). Many contemporary image description efforts do not take into account this population's nuanced image description preferences. In this paper, we present a qualitative study that provides insight into 28 BLV people's experiences with descriptions of digital images from news websites, social networking sites/platforms, eCommerce websites, employment websites, online dating websites/platforms, productivity applications, and e-publications. Our findings reveal how image description preferences vary based on the source where digital images are encountered and the surrounding context. We provide recommendations for the development of next-generation image description technologies inspired by our empirical analysis.

Author Keywords
Image captions, alt text, accessibility, visual impairment

CCS Concepts
•**Human-centered computing** → **Empirical studies in accessibility;**

INTRODUCTION
Digital images are plentiful across the media and information landscape. Towards enabling people who are blind or have low vision (BLV) to consume such content, a variety of efforts focus on the provision of *alternative text (alt text)* that is read through a *screen reader*. A screen reader is a software application that enables people who are BLV to read the text that is displayed on the computer screen with a speech synthesizer or Braille display. Alt text image descriptions are read off by a screen reader when a content author has followed recommended protocol, e.g. [13], and created an alt text attribute within a document or website's source code.

Though provision of alt text is a best practice, most digital images lack descriptions. A 2017 study of popular websites in many categories (as ranked by alexa.com) found that between 20% and 35% of images lacked descriptions, and that many images that did contain alt text had extremely low-quality descriptions, such as the word "image" or a filename [17]. Images on social media are particularly problematic; a 2018 study found that only 0.1% of images on Twitter had alt text [16]. While the ideal is for content authors to always provide high quality image descriptions (i.e. using the alt text field) at the time of document authorship, many are not despite efforts and resources developed to scaffold content authors in producing them (e.g., [13, 26]).

The absence of alt text from content authors has motivated scholars and practitioners to innovate, by introducing a variety of more scalable image description services that are powered by *humans* [4, 5, 7, 6, 45], *computers* [14, 24, 35, 37, 38, 43], and a *mixture of their efforts* [17, 28, 32, 33]. In designing image descriptions, such services can leverage the many guidelines for how to write effective descriptions [13, 11, 26, 29, 30, 34, 39, 41, 42, 44]. However, existing guidelines are limited in that they do not clarify how to account for the finding of Petrie et al. [30] in 2005 – an interview study with five blind people that found that the most useful information to be included *"was thought to be context dependent"*, i.e. based on the source in which the image is found.

Towards the goal of closing this *description gap* between what people want and what is provided, we present a qualitative study designed to investigate the image description preferences of people who are BLV. We interviewed 28 BLV people, guided by the question: "What are BLV people's experiences with and preferences for image descriptions found in different digital sources?". We draw on the following definition of *source*: the platforms and media where one may encounter digital images. Examples of digital images found in different sources are shown in **Figure 1**. We focused our investigation on seven sources: news websites, social networking sites/platforms, eCommerce websites, employment websites, online dating websites/platforms, productivity applications, and e-publications. We conclude with recommendations regarding what is important information to incorporate into image descriptions found in different sources. These recommendations can be of great value for improving human-powered, computer-powered, and hybrid image description services for people who are BLV. More generally, our work contributes to the design of social and technical infrastructures that are

CHI'20, April 25–30, 2020, Honolulu, HI, USA

ISBN 978-1-4503-6708-0/20/04…$15.00
DOI: https://doi.org/10.1145/3313831.3376404

accessible to all and support people to engage more fully with digital media.

RELATED WORK

Our research builds on prior work including guidelines for alt text image descriptions, studies about BLV users' image description preferences, and systems for facilitating or automating image description. Of importance, throughout this paper we use the term *description* as opposed to *caption* or *alt text*. Though the terms alt text and caption are commonly used in related scholarship, they infer specific linguistic structures of description that does not take into account contemporary AI-powered approaches to description as described in [28, 35].

Guidelines for Describing Images to People Who are BLV

The task of creating image descriptions–interpreting visual information and transmuting its meaning into language–is non-trivial [20, 23, 26]. Still, numerous efforts have made authoring image descriptions more approachable. Many focus on guiding web developers [41]. For instance, the Web Content Accessibility Guidelines (WCAG) provide basic instructions for the generation of alt text. The Diagram Center [11] provides instruction on assessing whether images are functional or decorative, whether information can be gathered from surrounding text, and to provide age-appropriate descriptions. The Diagram Center also notes that effective image captions describe foreground, background, color, and directional orientation of objects [11]. Such suggestions are in line with findings from related scholarship [34].

While the aforementioned works focus on one-size-fits-all guidelines for authoring image descriptions, other efforts have noted that descriptions need to be responsive to the context in which an image is found. Petrie et al. (2005) championed this idea [30], albeit did not present findings according to individual source types. Rather, they recommended guidelines that represented description preferences commonly observed across 10 sources (10 homepages in 10 different sectors), which were that descriptions include 1) the purpose of the image, 2) what objects and people are present, 3) any activities in progress, 4) the location, 5) colors, and 6) emotion [30]. More recently, researchers have discussed the types of content that should be included in descriptions of images found on social networking sites (SNS): describe all salient objects [29]; specify who is in the image, where it was taken, and others' responses to it [39]; indicate key visual elements, people, and photo quality [44]; and when captioning people, objects, and settings, specify details including the people count, facial expression, age, inside, outdoor, nature, close-up, and selfie [42]. Our work extends these prior works by identifying preferences of people who are BLV across seven sources. Building upon our observations, we also propose recommendations for the types of content that image description technologies should deliver for people who are BLV.

Understanding Users' Experiences with Descriptions

Our work relates to the body of literature aimed at understanding how people who are BLV experience image descriptions provided by technologies. The literature shows that people who are BLV want descriptions for digital images found on websites [30], on SNS [2, 29, 39], within digital publications, and in productivity applications [15]. Like many, they place value in image descriptions to stay up to date with the news [29], to enjoy entertainment media [29], and to engage in social interactions [2, 10, 29, 39, 44]. In addition to these common uses of images, people who are BLV depend on image descriptions to avoid risk (by not sharing images deemed unprofessional or low quality, or images that contain inappropriate content) [2, 8, 44]. In addition, scholars have found that people who are BLV want descriptions for images that they take in order to learn about the content of these images [4, 44].

While the need for image descriptions is clear, few prior studies focus on understanding BLV people's preferences for what kind of content they want described for images found on different sources. Our current understanding comes from a small body of dispersed literature. As previously noted, in 2005 Petrie et al. asked five BLV participants about the kinds of images they wanted described, what image content they wanted described, and their preferred length of description [30]. Others focused on BLV participants' experience with descriptions for images presented on social media platforms and how BLV users perceive automatically generated captions [25, 43, 44]. Finally, others have inquired into how people who are BLV want to interact with image descriptions, and how different delivery structures impact their experience [28, 35, 40]. Despite the importance of these findings, to our knowledge no prior work has explored how BLV people's preferences for the content in the image descriptions vary based on where they encounter an image description (e.g. on a social media site versus in an e-textbook). Our work fills this gap towards supporting opportunities to make image descriptions context-specific.

Image Description Technologies

Many images found on digital sources do not contain alt text or effective image descriptions [17, 16]. The low rate of manually-produced descriptions has inspired some investigations into new approaches to generate image descriptions. These approaches are often described as human-powered, automated, and hybrid approaches. *Human-powered approaches* [5] provide near-real-time descriptions of images through crowdsourcing [4, 45], friendsourcing [7], and social microvolunteering [6]. *Automated image description approaches* employ artificial intelligence models to generate image descriptions [24, 14, 38, 37, 43, 35]. *Hybrid image description technologies* create human-in-the-loop workflows that work in tandem with automated approaches [17, 28, 32, 33]. Tools also have been introduced to train non-specialists (including crowdworkers) to identify which images and diagrams in text books need alt text [12, 26]. Extending prior work, our study reveals new design opportunities for improving image description technologies by contextualizing descriptions based on the source where images are found.

STUDY DESIGN

We conducted a qualitative study guided by the following two research questions:

RQ1: What are BLV people's experiences with digital images on different sources?

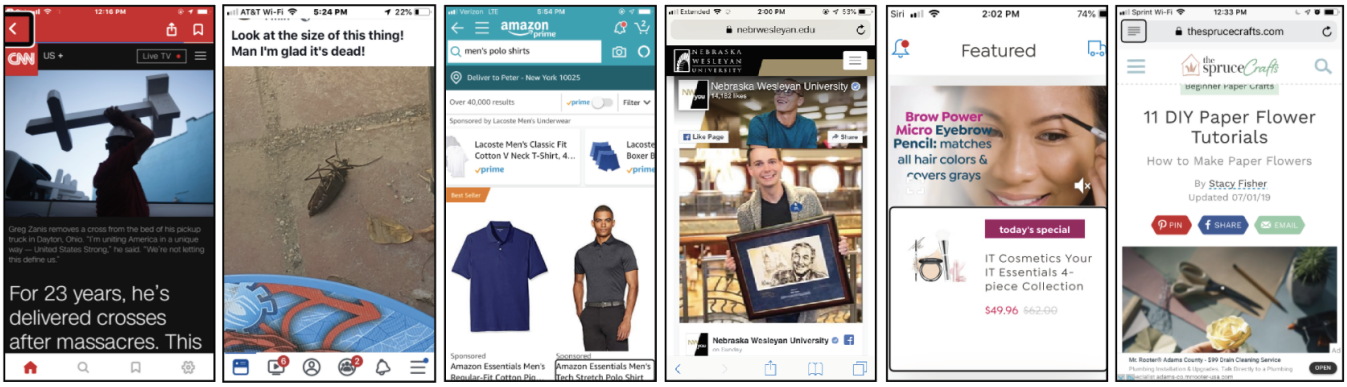

Figure 1. Examples of digital images that participants in our qualitative study encountered when browsing different sources. Participants wanted more information for all these images, particularly because none of the images had associated alt text.

RQ2: What are BLV people's description preferences for digital images in these different sources?

We formed these questions based on the understanding that source is a significant factor that impacts a person's description preferences [30]. We assumed this to be the case in order to limit the scope of this study[1].

Data Collection

To learn about BLV people's experiences with digital images that they encounter on different sources, we designed a semi-structured interview protocol that included 15 open-ended questions, 13 Likert survey statements, and a contextual inquiry. Prior to each interview, we asked each participant to bring their preferred access technology with them. We audio recorded each interview. After the interviews, we sent the audio files to be transcribed by a professional service. We also took field notes to keep track of emerging themes.

Our interview procedures are in the Supplementary Materials. In summary, for the open-ended research questions, we asked about our participants' visual impairment, access technology preferences, experience with digital images, and experience with technologies and services that provide image description (from alt text to automated image description services). For the contextual inquiry, we asked participants to open their technology and visit three to five sources where they would expect to find digital images; the number of sources varied based on how long it took for the participant to complete the task or their familiarity and interest in the source. We suggested the following options: a news website, a SNS post, an eCommerce website, an organization or employment web page, and a productivity document, e.g. Word or PowerPoint. We identified these sources based on prior work that indicated that people who are BLV want image descriptions to pursue their interests through staying up to date with the news [29], enjoying entertainment media [29], eCommerce [35], staying socially connected [39, 29, 2, 44, 10], dating [31] and performing work or academic pursuits [15].

[1]Other contextual factors that we choose to not report on in this study include: 1) image use case, 2) the topic of the source, 3) a person's visual impairment and visual literacy, 4) interaction preferences. We elaborate on these other factors in the Discussion section.

Participant Recruitment

We recruited participants by circulating an IRB-approved announcement on social media, on a listserv managed by organizations serving people who have visual impairments, and through snowball sampling at an independence training center. To be eligible, we specified that participants had to be at least 18 years old, be BLV, and use a screen reader and/or magnification. The announcement explained that participants would be compensated with an Amazon gift card, at the rate of 20 USD per hour. We aimed to have equal participation of people who have congenital blindness, acquired blindness, congenital low vision, and acquired low vision. At the onset of recruitment, we accepted all participants that met our basic criteria for inclusion. After 20 participants were recruited, we selected participants based on their visual impairment towards the goal of equal representation.

In total, 28 people participated in our study. We conducted 25 of the interviews in person in a 50 mile radius of our U.S. metro area, and another 3 over the phone with individuals in other states to achieve greater diversity of visual experience within our participant pool. The same protocol was used when conducting the interview over the phone, with the key distinction being that the researcher and the participant accessed the image sources on their own devices when conducting the contextual inquiries. We believe that we reached participant saturation based on Alroobaea et al.'s finding, which states that there is no certain number of participants for finding all usability problems (during interviews and think-aloud approaches), though the rule of 16+/-4 users gains much validity in user testing [1]. The interviews lasted between 1.25 hours and 2 hours, depending on the participant's experiences and interest in the topic. All participants used Apple or Android phones for the contextual inquiries.

Table 1 summarizes participants' demographic information. As shown, the participants represent a diversity of backgrounds in terms of gender (16 women, 12 men), age (range is 18 to 67 with a mean of 39.05), education (from people who had not

ID	G.	Age	Edu.	Occupation	Diagnosis	Onset	Vis. Exp.	Access Tech
P01	M	28	B.A.	Videographer	Myopia	17	ALV	iOS Mag.
P02	F	39	J.D.	Art teacher	Malignant pseudotumor cerebri	34	ATB	iOS V.O
P03	M	39	<B.A.	Student	Meningioma tumor	36	AB-LC	iOS V.O
P04	M	67	<B.A.	Coffee roaster	Retinal detachment	0	CTB	iOS V.O
P05	F	29	B.A.	Student	Retinal detachment	0	CB-LC	iOS V.O
P06	M	30	<M.S.	Unemployed	Retinal detachment	19	ATB	iOS V.O
P07	M	23	H.S.	Student	Retinal detachment	17	AB-LC	iOS V.O
P08	F	21	<B.A.	Unemployed	Leber's congenital amaurosis	0	CTB	iOS V.O
P09	F	34	AA	Unemployed	Uveitis due to rheumatoid arthritis	32	AB-LC	iOS V.O
P10	M	64	M.A.	Access consultant	Retinitis pigmentosa	0	AB-LC	iOS V.O
P11	F	41	<B.A.	Real estate manager	Retinal scarring	34	ALV	iOS V.O/ Mag.
P12	F	58	B.A.	Housing manager	Retinal detachment	47	ALV	iOS V.O/ Mag.
P13	M	53	<B.A.	Clinical facilitator	Retinitis pigmentosa	46	AB-LC	iOS V.O
P14	F	34	AA	Student	Retinal detachment	9	AB-LC	Android S2S
P15	M	51	<H.S.	Music producer	Retinitis pigmentosa	26	ALV	iOS V.O
P16	M	39	<PhD	Student	Leber hereditary optic neuropathy	16	ALV	iOS V.O
P17	F	24	H.S.	Unemployed	Small eyes that never grew	0	CTB	iOS V.O
P18	F	19	<B.A.	Student	Septo-optic dysplasia	0	CLV	iOS V.O/ Mag.
P19	M	27	<B.A.	Student	Retinopathy of prematurity	0	CTB	iOS V.O
P20	M	37	B.A.	Minister	Relative lens position	0	CTB	iOS V.O
P21	F	37	B.A.	Teacher	Glaucoma	0	CTB	iOS V.O
P22	F	65	M.Ed	Rehab. specialist	Rieger's anomaly	10	AB-LC	iOS V.O
P23	F	33	<B.A.	Politician	Retinitis pigmentosa	0	CTB	iOS V.O
P24	M	29	B.A.	Student	Leber hereditary optic neuropathy	23	ATB	iOS V.O
P25	F	47	M.Ed	Educator	Unknown	0	CLV	iOS V.O/ Mag.
P26	F	18	H.S.	Student	Optic nerve dysplasia	0	CTB	iOS V.O
P27	F	27	B.A.	Student	Laser eye surgery gone wrong	13	ATB	iOS V.O
P28	F	20	B.A.	Student	Unknown	0	CB-LC	iOS V.O

Table 1. Demographics of study participants. *G=Gender* (M=Male; F=Female). *Edu=Education* (< H.S.= Some High School; H.S.=High School; AA=Associates; < B.A.=Some Bachelors of Arts; B.A.=Bachelors of Arts; < M.S.=Some Masters in Science; M.S.=Masters in Arts;M.Ed.=Masters in Education; < PhD=Some Doctorate JD-Law=Doctor of Jurisprudence). *Vis Exp=Visual Experience* (CTB = Congenital Total Blindness: No visual cues or direct visual experience with images; CB-LC = Congenital Blindness with some light/color perception: No direct experience with images; ATB = Acquired Total Blindness: Prior experience with images; AB-LC = Acquired Blindness with some light/color perception: Prior direct experience with images; CLV = Congenital Low-Vision: Limited prior experience with images; ALV = Acquired Low-Vision: Prior experience with images). *Access Tech* (iOS Mag=iOS Magnification Tools; iOS V.O.=iOS Voice Over; Android S2S=Android Select to Speak).

completed high school to people who have a doctorate), and occupation (from people who are unemployed or retired, to those who are students, DJs, lawyers, and educators). These participants had a range of visual impairments (from unformed retinas, to myopia, to blindness acquired due to laser surgery) as well as varied experiences with visual information.

Data Analysis

After conducting all 28 interviews, we performed a qualitative analysis of the transcribed data. We then performed *axial coding*, a process of identifying and relating codes to each other, via a combination of inductive and deductive thinking [36]. We used deductive reasoning to identify the sources of interest based on the literature, and then inductive reasoning to attribute the content patterns to these sources. To prepare the data for the axial coding, two team members cleaned up major errors in the transcript by reviewing the audio, all the while taking analytical memos to record emergent themes.

At the onset we established the seven sources (*news websites, social networking sites/platforms, eCommerce websites, employment websites, online dating websites/platforms, productivity applications, and e-publications*) plus an *other* category to account for emergent sources as parent codes (or primary phenomena orienting the study). We then used a semantic analysis technique to identify and code text segments according to the parent codes. Braun and Clark explain that to perform semantic analysis one should "not [be] looking for anything beyond what a participant has said or what has been written [9]." While doing this, we dynamically identified and refined a set of child codes. Child codes that we identified include: *Image Access Behavior:* statements about how one approaches consuming the media; *Image Access Experience:* statements related to one's exposure or interaction with content in digital images; *Description Content Wants:* statements about the features, attributes, or details that should be included in an image description, and *Description Considerations:* statements related to the factors that impact image access or content preferences. Within each subset of data we made note of common and unique themes amongst all participants' responses. For instance, under the child code *Description Content Wants* we noticed often times participants talked about the *level of detail* they wanted for an image on a source or their need to understand the *purpose* of an image.

Performing the qualitative data analysis with the sources as the parent codes enabled us to perform a cross-source analysis that highlights how our participants' image experiences and description content preferences differ based on source. Of note, we present the variety of perspectives shared by our participants, as opposed to a quantitative analysis of how many people in our participant group shared the same experience, because the aim of this work was to understand the range of experiences and content wants as opposed to the frequency.

FINDINGS

We present our findings for our research questions: **RQ1:** What are BLV people's experiences with digital images on different sources? **RQ2:** What are BLV people's description preferences for digital images in these different sources?[2]

Source 1: News Articles

RQ1: While many participants shared that they read news articles, we observed in the contextual interviews that none of the image descriptions encountered provided the participants with the information they needed in order to understand what was in the image.[3] One reason was because the alt text was uninformative; e.g., it simply states *".jpeg"* or a long file name. Participants' responses to uninformative alt text was similar to the sentiment shared by P28: *"I don't know what the heck it was, but I'm sure the article will tell me what the image was."* Reinforcing this perspective, we heard P26 share what she thought was in the image based on the article headers: *"Honestly, I don't really know. Okay. Okay. They're explaining there, going into all these, this detail about like one of these explosions are happening on the earth's inner core and that gives me like no information on what's going on in the image."* Another common reason participants did not engage with images was that they were unaware an image was present on the web page, but some participants acknowledged this could be happening. In P17's words, *"That's annoying. I'm sorry. So the way that this website is structured is unclear. It has a heading that shows the title of the article and then it has a bunch of other headings to other news articles. So it's difficult to tell. There is no image as far as I can tell."*

RQ2: Participants shared that they want image descriptions that clarify the purpose of the image in the news sources. As P28 noted, *"So usually if there is an image attached to an article, there's a reason for that image. They may take 1500 pictures of a protest, but only choose two [to] be on the website. Why did those two pictures get chosen?"* In P16's words, *"I think it's [images are] just information to tell the story. But, just saying 'image' does nothing. If there's an image, tell me why it's important, I guess."*

Regarding the type of description content that participants want, we heard a variety of preferences based on the news story. For example, P05 noted, *"[My preferences] depend on the article, but I would say the scene [of the event]. What like is it a politician in the, like a person of interest? Um, is it a sports team? What are they doing? So like what's the scene, what's happening, what are the actions?"* This variety reinforces participants' stated preference for descriptions to clarify the purpose of images. We elaborate about some of the variability we heard below.

For images where people are the central focus, the key content they want centers on the identifiable characteristics of the person or group and important details about their interactions or experiences. As P19 shared, *"There are times when all you want to know is that there's a group of people sitting on a bus, other times you want to know that all the people are all smiling or had tears rolling down their faces and they look really sad– especially in like any kind of a news article."* In addition, we observed that in some cases our participants want to know about a person's race or ethnicity. As P9 noted, *"I would put that in there to make sure that everybody is represented equally. For me, it helps keep the news accountable and aware of their racial profiling."* Others noted that knowing whether a person is a celebrity or not matters within the news context because they have cultural influence.

For images where the focus is on the events and scenes, participants want to know about the central people or objects, the activity that they are engaged in, how they are interacting with one another, and pertinent information about the setting.

For images that highlight objects or landmarks, they want to hear about the unique characteristics or features. As P4 noted, *"If I was reading about a new airplane I might want to know how big it is, how many people it carries."* As P11 pointed out *"sometimes you want to know what is written on the protest sign behind the person standing at the podium."*

Source 2: Social Networking Sites (SNS)

RQ1: Most participants shared that they use SNS.[4] While participants reported a high engagement with image descriptions in SNS, this only was with respect to Facebook. As P08 shared, *"In comparison to other places, Facebook app is honestly pretty accessible. If there's someone in the images it will tell me the name of the person and something about the setting."* Despite the positive responses, participants also readily described limitations. For example, P02 shared, *"It's very hit or miss. Sometimes it'll just tell me 'person, shoes, and trees.' Is the person naked? Does that mean that there's a person only wearing shoes and they're standing on top of a tree? It's not specific enough content."* Participants reported frustrations with other SNS that do not provide image descriptions. For example, P07 noted, *"Twitter–they are not accessible at all. I think that Facebook gives their images alt text but Twitter does not."* Similarly, P14 shared, *"I do use Twitter once in a while. Last time I was there I noticed the pictures and images alt text weren't there."*

[2]We chose not to report the Likert data, because there was great variance in the rationale participants assigned each of the scores. For instance, one person might have given a 3 (undecided) because they did not care, where another gave a 3 because the situation seemed ethically complicated. Accordingly, we found the scores to be less meaningful, but opted to include the participants' rationale for their preferences in the overall thematic analysis.

[3]During the contextual inquiry, 12 participants demonstrated how they access news articles from CNN, BBC, Fox News, New York Time, Wired Tech News, or other local news websites. The other 16 participants accessed a news article through the Apple News App.

[4]Twenty-six participants indicated that they have Facebook profiles, and two people reported using Twitter. During the contextual interviews, nineteen people brought us to Facebook.

RQ2: Participants shared that they want image descriptions in SNS that help them understand the purpose of the image. As P16 noted, *"People share a lot of personal images. You have to infer why they're sharing it based on their strange texts. More detail is necessary."* We learned that purpose is especially important when the person posting the image does not provide a comment or the comment did not directly reference the image content.

For images on Facebook, the type of description content our participants wanted centered on descriptions of people. For instance, P09 noted, *"I like to have them include more like facial expression...were they smiling or smirking? Was it a mischievous look?"* More generally, participants want to recognize facial expressions or body language to help them decide how to respond to the image. Notably, participants want more content described when they or people they know were in the image. As P12 shared, *"I want it pretty detailed especially if it's somebody I know...what's going on and why they're in the picture and what else is happening in the picture."*

Other description wants center on the elements in an image that help them understand what the person is doing or their environment. In the context describing a family portrait, P05 asked for *"Something like five family members standing in front of a Christmas tree or something like that. The number of people and who's in the picture. Who's in the picture and their actions–what they're doing."* In a different context, P09 shared, *"If my friend was showing off her engagement ring...I would want to know if it was a princess cut. Just giving that it is a ring...that is not enough."* Additionally, if a person's attire is remarkable, our participants wanted that information.

Source 3: eCommerce Apps and Websites

RQ1: Many participants indicated that they shop online.[5] Amazon was the primary website of choice. Participants shared that they shop online for clothing, household items, electronics, entertainment media, gift items, as well as to do research about new products and as a hobby.

Overall, we learned that our participants have very low expectations for image descriptions on eCommerce sources. We repetitively heard frustration and apathy from many participants. For example, P2 exclaimed, *"Amazon gives really poor descriptions honestly...I mean it really is all you get is Ding, Ding, Ding, Ding, Ding [the screen reader issuing a tone for an empty image description]. Amazon can really piss me off. I was buying an ottoman. There was no description in a picture. I had to lurk in the comments looking for somebody who finally said what the hell it looked like. They encourage people to put photos in the comment section; none of those photos are described."* P19 noted, *"Amazon...So one kind of pet peeve I have of pictures is that a lot of times since there isn't any alt texts or anything, some of the screen readers will tend to think that the path of the picture should be read. So you'll have this entire, 5,000 character long path name...you have to read a page of these stupid identifiers."* In addition to the frustration for missing information or nonsensical descriptions, our participants expressed concern that they do not have equal access to image content on eCommerce websites. P5 shared, *"Amazon is the least accessible. Accessibility for me, it means being able to get the information from an image comparable to how a sighted person would get that information. I don't get that."*

RQ2: The type of description content participants want for eCommerce centers on descriptions of objects. This is unsurprising given that many images on eCommerce sources contain one product on a clean, solid-colored background. The specific descriptive details participants wanted varied based on the type of product.

For clothing, they first wanted to know color and then attributes such as the general style of the garment (formal, professional, athletic, casual), stylistic details on the garment (zippers, pockets, thick hemlines, sleeve length, material, pattern), and how it fits on the model's body as well as the model's body shape. For example, P6 noted, *"Color is interesting, so is the length of the sleeves. Maybe the cut, zip front, the hem line, how it fits, are the selves big or tight? Does it have drawstrings?"* P9 shared, "*I would say the model is the model really skinny? Is the model more of a plus size is, because to me the models really help paint the picture of how the shirt [is] gonna fit."*

When it came to household items and electronics, participants' description wants centered on the unique attributes of the object's form or materials, as well as text, symbols, or logos on the item. For instance, when P18 was learning about an image of a mug, she asked very specific questions like *"How much of the cup is covered by it [the pattern]? How and where does the handle attached? Is there anything about it that makes it look good to use while traveling?"* When discussing purchasing items on eBay, P8 reflected, *"It would be great if they described any scratches or dents or cracks on an item."* The participants also want any text or logos described for products. For example, P19 brought up a picture of a computer adapter with a lightening bolt port and had the interviewer describe the picture to him. He responded, *"I didn't know that the lightning bolt was actually a picture of a lightning bolt on it [the computer adapter]. I definitely want that detail"*.

Source 4: Employers'/Employment Websites

RQ1: Our participants' familiarity with employment websites varied from current and active use to no familiarity.[6] For some people, an employment website meant a specific employer's web page, for others it meant a potential client's website, for others it meant job boards (USAJobs or Indeed). None of our participants recalled encountering image descriptions on employment websites, job boards, or the like. Several people were surprised that we would ask about this source. As P02 shared, *"I feel like you're on level 5,000; I'm still trying to figure out if there is a picture on any page. Am I missing functional content or is this just like decorative? I just assumed that all the images are not described on those job sites because they are decorative."*

[5]During the contextual inquiry, 21 participants opted to shop on Amazon. Others visited B&H Photography, CapHillStyle.com, Forever 21, Hot Topic, HSN, Starbucks, Target, and Walmart.

[6]During the contextual inquiries, 10 people chose to show us an employer's website; four of these sites were university websites. (We do not report on the sources to maintain anonymity of our participants.)

RQ2: For type of description content, participants primarily want to learn about people in the images and the work environment. Most prominently, participants want to know the facial expressions. As P23 put it, *"If they all look like they're miserable, you're probably don't want to work there or help them."* Participants also want to learning about people's attire; As P05 explained, *"I want to know how a person is dressed and looking...first impression is important in the music industry; you judge people by how they're dressed."* Some want content that would help them learn about the diversity of the people working at the company. As P10 shared, *"If there are photos of board members, I want to know if they were a bunch of white guys, if there is racial diversity."* Others expressed interest in getting information about the types of work tasks people engaged in and the work setting. P21, *"Whether the office looks busy. Are they sitting around or at a desk? Is it like a party that they're having?"* P25 anticipated wanting to know, *"Is it cluttered? User-friendly? Does it have dark walls, light walls? That might not be directly relevant to me, but it will give me a lot of information about the overall work environment and people's attitudes."*

Source 5: Online Dating Websites/Applications

RQ1: All participants reported they had never visited an online dating website. Additionally, none provided the name of a dating website and none suggested going to a dating website during the contextual interview. The reasons reported for not using them centered on the sources' overall inaccessibility, that it is preferable to meet others in person, or that they were not in need. This said, all participants provided meaningful answers to our questions about their description preferences and expressed interest in this source.

RQ2: The types of description content participants want centers on describing physical characteristics of a person, with specific interest in the color of a person's hair, the style of their hair (and/or the style of a man's facial hair, if applicable), the body type, and/or weight. Some people indicated that they would want to know somebody's eye color, race or skin tone, facial expression, and/or if the person had a defining physical characteristic. For instance, P03 noted, *"I'd say that would probably be the one defining feature. Like in any extreme irregular irregularities... things that are obvious to a person that is sighted."* Other attributes that emerged as important include: the person's attire, how well kept or clean they appear, as well as the presence of any tattoos (and what they depict). Some of our participants also indicated wanting details about the setting of the photo *"because that gives me information about their interests"* (P25). Other content wants centered on knowing whether a pet is in the image, and the composition of the photo, e.g. whether all of the photos were selfies, candid photos, or of larger group shots to know more about how subjects presented themselves.

Our participants noted that if they knew the person describing the image to them, they would be more inclined to ask for a subjective evaluation of the way a person looked. Importantly, we heard a variety of concerns related to whether an image description of a person can be objective or unbiased. For instance, P02 noted, *"How do you really describe a person? Isn't that judgment call? Even if it is as objective as you can, there's still going to be different things that people like [...] That [diversity] should be the beauty as opposed to losing it to norms."* P10 shared, *"I don't want a third party telling me they think someone is handsome or beautiful."* P16, *"It's going to be very subjective. I mean I guess you could comment on some things that are not a judgement."*

Source 6: Productivity Applications

RQ1: Participants' engagement with productivity applications varied greatly.[7] For those who use them, they reported low engagement with image descriptions. Comments about this include *"I have encountered images in only a few cases, but I don't feel like it was intentional. I'm going to say they're almost nonexistent because they're not there"* (P03), *"I don't encounter images on Word. As far as I know, no one sent me anything in Word that had a picture"* (P02), and *"I use PowerPoint, but I have to have help. I basically create slides and then have somebody help me find and paste pictures in. But reading PowerPoints is even worse. When a professor gives me them, they're not described."* (P16). Still, several participants reported using features to add alt text to images. As P08 shared, *"PowerPoint has started doing this thing where when you create a PowerPoint, you can actually go into the settings and put alt texts on the images and I love it!"*

RQ2: For text editing documents, the primary concern we heard centers on whether an image is decorative or functional. For example, P11 noted, *"If it's just like a placeholder image that is not relevant to the text, I don't really care if it's described."* P10 shared, *"Hopefully I can figure it out if it's something that would be important, and then I can figure what to pay attention to."*

The type of content participants want varies based on the image's purpose. In the words of P09 *"It would depend on the context of the document and what it was about. I would want enough information to give myself as close of a representation of that experience as I could to recreate that."* In reference to an image on a PowerPoint presentation, P11 noted, *"If it's like a biology presentation or document about a molecule and there's a picture of a molecule, I want to know like what does the molecule look like, like what are the bonds and the atoms and stuff like that."*

Source 7: E-Publications

RQ1: The participants in our study had a range of experiences using e-publications, which they understood to be digital textbooks, PDFs, and materials found on audio book platforms like Bard and Bookshare. The participants who had experience using digital textbooks noted that the images presented within this source were not accessible to them. P08 noted, *"Last year I used an online textbook. They didn't have any way of describing for you what pictures actually were."* P17 shared, *"When I came across an image it would just say image."* P28 expressed, *"The problem is that if there are images*

[7] 20 people in our study indicated that they had previously used Microsoft Word; one participant reported using Apple Pages, and two people reported on using Google Docs. Only nine people reported ever having used a presentation application, i.e. Microsoft PowerPoint, Apple Keynote, or Google Slides.

that are often not described, so this is particularly unhelpful." The participants who mentioned encountering and/or using PDFs shared that digital images within this format are almost always inaccessible. P06 noted, *"Occasionally I might get an email with a PDF attachment. The images in them are mostly not accessible."* P13 shared, *"PDFs read funny a lot."* The participants who mentioned accessing materials through audio output (e.g. Bard, Bookshare) reported a similar dearth of images being described. Importantly, when speaking to participants about digital images in e-publications, they often spoke about diagrams, charts, or maps.

RQ2: In terms of type of description content, many participants simply said *"same as for Productivity Applications"* or *"it depends on the context of the image."* The lack of depth in their answers may be attributed to the fact that we did not vary the order of our questions for each participant (i.e. since this category was last, participants may have been fatigued) or that participants had less experience with images in this source.

Emergent Sources

While we designed our study to focus on seven sources, additional sources of interest emerged. We describe these below.

Web Browsers: Participants reported encountering inaccessible images when searching on web browsers. An example is advertisements that pop-up when searching. While P10 noted that advertisements are generally a nuisance, he also shared *"I'd want the option to have it described."*

Instructional Websites/Blogs: Participants reported encountering images on blogs or websites that contain instructions for how to accomplish a task, such as crafting or cooking. The expressed description wants focused on details about the objects being made, and if there *"are more than one picture, what difference there is between the photos so I can follow along with the instructions."*

Hotel websites: Participants reported coming across inaccessible images on hotel websites. One participant (P10) provided a list of his content interests, *"[For example, in an image description I would want something like], 'Our front desk clerk stands behind the podium so they can step out easily to work directly with someone in a wheelchair. Our lobbies are covered in plush carpets, or we have tactile different floor surfaces'."*

Personal Photo Gallery: Participants wanted descriptions for images they had taken to help them know what they were sharing with friends or to organize their albums. These participants did not provide explicit description wants; we attribute this omission to us not speaking about a specific purpose for the images.

Public GUIs: Participants noted they encounter digital images on public devices or interfaces in libraries or airports, but did not specify content they would want in descriptions.

Cross-Source Analysis

Level of Experience with Digital Images: During our analysis of the data with respect to RQ1, "*What are BLV people's experiences with digital images on different sources?*", we observed that people who are BLV generally have low engagement with digital images. In some instances the low level of engagement was linked to their familiarity and use of such sources. For instance, none of our participants had direct experience with images on online dating websites as a factor of not using them. When discussing images found on employment websites and productivity applications we learned that it may be difficult for people who are BLV to discern whether an image is present, in part because they do not use these sources as often as others and/or that they do not anticipate a strong purpose for the images on these sources. For other sources, low engagement stemmed from inadequate descriptions of images on the websites (e.g., news and shopping websites). This latter class of sources are where our content preference findings can have immediate impact. Interestingly, we observed one outlier where participants reported high engagement with images: for SNS (specifically, with Facebook). Still, our findings illustrate that participants are seeking more from the image descriptions than is provided to them today and our findings offer insight in how to make such improvements.

Image In(Dependence): During our analysis of the data with respect to RQ2, *"What are BLV peoples' description preferences for digital images in these different sources?*", we made the general observation that the source informs what one expects from a description of an image on that source. For some sources (e.g. dating websites), participants expressed interest in learning about the image as-is without taking additional information from the source (e.g. text) into consideration. For other sources, participants want the description to be based on additional information beyond just the image. For example, participants expected the information surrounding the image to drive what content would be described in an image for news sites, productivity documents, e-publications, and SNS. Accordingly, when developing processes to generate meaningful image descriptions it is important to be discerning about when and how to use the content surrounding the images.

Amount of Content: Also with respect to RQ2, we observed considerable diversity across sources in terms of participants' desires for the amount of content and level of detail they want in a description. We offer a nuanced view of how participants' content wants vary around source types in **Table 2**. For each source, we specify all the types of content from a lengthy list of options that at least one of our participants thought was important content to describe. We group these findings around three key themes that are commonly the central focuses of an image composition: event/scene, people, and objects/landmarks.

Notably, for some sources, the amount of content desired in an image description was greater than on other sources. For instance, we noted that participants want to have the most content available to them for images found on SNS, dating sites, and news websites, whereas there were fewer description content wants for images found on productivity applications and e-publications. We attribute this to the fact that our participants viewed images as a central focus of SNS, whereas images on productivity applications and e-publications were viewed as more decorative (which may not necessarily be an accurate assessment of the role of imagery in these sources).

Content Area	N	SN	eC	E	D	P	EP
Event/Scene							
People Present	x	x	x	x	x	x	x
Text	x	x	x	x	x	x	x
Activity	x	x		x	x	x	x
Interaction	x	x		x		x	x
Landmarks	x	x			x	x	x
Building Features	x	x		x		x	x
Weather	x	x				x	x
Lighting				x			
People							
Text	x	x	x	x	x	x	x
Salient Objects	x	x		x	x	x	x
Activity	x	x		x	x	x	x
Gender	x	x	x	x	x		
Race/Diversity	x	x		x		x	x
Name of Person	x	x				x	x
Celebrity Name	x	x	x				x
Expression	x	x		x	x		
Attire/Clean		x		x	x		
Body Shape/Size			x		x		
Pets		x			x		
Hair Color					x		
Hair Style					x		
Eye Color					x		
Unique Physical					x		
Tattoos					x		
Object							
Text	x	x	x	x	x	x	x
Name	x	x	x	x	x	x	x
Form	x	x	x				
Fit		x	x		x		
Color		x	x		x		
Overall Style		x	x		x		
Material	x	x	x				
Logos/Symbols		x	x				
Damage			x				
Unique Features			x				

Table 2. Results of cross-source analysis where x specifies description content want. (N=News. SN=SNS, eC=eCommerce, E=Employment, D=Dating, P=Productivity, EP=E-Publication).

For other sources, the content focus was highly variable; e.g. for news websites, it was dependent on the news story.

Amount of Detail: As noted above, some images found on some sources may require more content than on others. That said, during our analysis we also noted that there are other factors that may impact the amount of content and/or the level of detail that is included in a description. For instance, we noted that the task one is involved in or the amount of time they have influence the amount of content they want. In P26's words about news sources, *"When I was younger I really loved it when people went all details...now I'm older I don't really have time for that. It's really nice when I know what's going on, but I don't have to know that a bird was flying over the people."* We also noted that the level of detail one wants may be dependent on whether they previously had vision. In contrast to P26, we heard from P09 that in almost every circumstance that they wanted as many details as possible "because to me that helps paint the picture [...] I'd rather be on sensory overload." We also learned that for some, having all content available to them is an issue of equity/justice and/or personal interest, whereas others find too much information can be distracting, unhelpful, or boring.

DISCUSSION

While it is already known that image descriptions are imperfect, our findings offer promising evidence that part of the reason may be because the one-size-fits-all approach that is widely-used today is inadequate. In what follows, we discuss how our findings relate to contemporary research followed by design recommendations for how to improve image captioning services and future research directions.

Comparison of Our Findings with Prior Work. Our findings provide new insight into BLV people's description preferences for images found in and across seven source types. This work builds on Petrie et al.'s [30] claim that the description preferences of people who are BLV vary based on source, as well as existing image description guidelines, e.g. [13].

Our findings underscore types of description content that may be desired *universally,* across different sources. For instance, our participants consistently wanted to learn about people and objects across all sources (**Table 2**). This aligns with well-established guidelines [13, 11] and prior findings for images found on SNS [29, 39, 44]. Extending prior findings, our work also reveals that participants consistently wanted a description of text that is present in images across all sources.

Our work reinforces the importance of [31] by reporting that BLV people want dating platforms to be accessible. Further, in alignment with [31], we heard some of our participants express concern that a description of a person's physical appearance can be very subjective. Our study enriches our understanding of this issue, highlighting how desired description properties can even be controversial. Take a person's race as an example. Some participants noted that including information about a person's race or ethnicity in an image description would be necessary when the image is paired with a story or post related to social justice or cultural interest. Other participants noted that it is important to have access to all of the same information a person who is sighted has–which would require this information to be disclosed. Still, others expressed concern about whether race or ethnicity can be accurately determined by a picture alone, where accuracy may only be determined by the person who is being represented. Our findings underscore the importance of connecting efforts on generating image descriptions to contemporary literature (e.g. on race and gender studies) to address how to handle some of the content areas that could be considered subjective or sensitive (e.g. race, gender, body shape, disability).

Design Recommendations for Next-Generation Image Description Services. We offer our findings about description content preferences of BLV people with respect to different sources as a valuable starting point to designing improved image description services for this population. This is rele-

vant whether training professionals, training crowd workers, or creating large-scale datasets to train AI algorithms.

Our findings offer a tangible guide regarding what information is preferred for seven sources. Developers could use the taxonomy in **Table 2** to support source-specific description guidelines or templates for human-authored descriptions. Already, Morash et al. [26] have found STEM-specific templates improve alt text in textbooks; creating templates for other domains may therefore be useful. Alternatively, our taxonomy could be used to redesign instructions given to crowd workers, when authoring image descriptions that are used for training AI models, and to support inclusion of relevant details depending on image context.

In addition, our findings reveal that some description wants are more general, meaning they can be applied to all sources, whereas others only apply to a few or one source. For instance, all image descriptions should include text and identification of people and objects, whereas information about tattoos, lighting, hair style, and damage were only wanted on one source (and these sources varied). This knowledge could be useful in prioritizing the relative importance of gathering data or training models to include certain categories of information.

Our findings also reveal that the description wants our participants specified often go well beyond capabilities of current vision-to-language AI systems. For instance, we found that multimodal analysis of all of the media on the source surrounding the image (e.g. text, video) is necessary, in some cases, to devise a meaningful image description. In particular, appropriate descriptions for images found on news sites, SNS, productivity applications, and e-publications greatly depend on the surrounding content. Yet, today's AI systems only observe the image when generating a description. Future algorithms should be able to identify when and what surrounding content is needed to create a description.

Notably, many people who are BLV do not trust the descriptions provided by today's AI systems [25]. Inclusion of this population at the early stages of the innovation of technologies is one step towards ensuring trustworthiness of image descriptions. This aligns with contemporary discussion that emphasizes the need for "protecting people who fall outside of the 'norms' reflected and constructed by AI systems" [21], ensuring fairness [18] or justice [22] in AI for people with disabilities, and aligning with other ethical considerations [27].

Future Work. Despite our guide for what content is preferred when, we still believe ongoing, larger-scale analysis is important. We note a few ways in which future work could extend our study. One valuable direction is to examine participants' diversity in perspectives based on how much exposure they had to visual information, whether that is based on the level and time of onset of their vision loss, their training in visual literacy, and their direct experience with the objects or phenomena represented in an image. A further factor that may bear influence pertains to the use-case in which a person intends to use the image. Additionally, it's clear that some people prefer more detail while others prefer less. For some, having all content available to them is an issue of equity/justice and/or personal interest, whereas others find too much information can be distracting, unhelpful, or boring. In addition, when a person does not have prior experience with the content area or a similar cultural reference point, a higher degree of detail (and/or additional modes of representation) may be needed for a person to create a mental image or approximate reference, as noted in [12, 26]. Valuable future research includes personalizing descriptions so that in addition to consideration of source, there also is consideration of each person's preference for the level and type of description for each source.

Relating to our above study suggestions, we believe that next-generation image description systems might also benefit from including features that: 1) enable the user to specify the quantity of content described; 2) enable users to decide what level or precision of language that they want in a description (e.g. dark blue, vs. space blue 1C2951–RGB 294181, HEX 294181) or domain-specific language (e.g. the architectural style of buildings on a college campus). Based on this research we also hypothesize that the following features might assist people who are BLV to locate images and engage in determining the right description for them. These ideas include: 1) providing the option to read a series of descriptions written for the same image to empower an individual to learn about different description styles and assess the accuracy of a description to the surrounding context; and 2) presenting image descriptions before or after the main body of text.

We also heard from participants that they would like to be able to ask for descriptions on demand (as opposed to depending solely on alt text or existing descriptions); such an opportunity would address a series of underlying concerns about descriptions, including the inequity faced by not getting the same information as others and discomfort about receiving incomplete descriptions. These findings affirm the need for further research related to next-generation, interactive technologies for describing images [3, 19, 28].

CONCLUSION

In this study, we took a holistic approach to examining BLV people's experiences with digital images found on different sources, and the variance of their description preferences across sources. The findings we present in this paper may be used by scholars and practitioners who are working to refine the ways in which image descriptions are generated by human-powered services, AI-powered services, and hybrid services for generating image descriptions. Ensuring image accessibility for people who are BLV is particularly important given the widespread proliferation of visual media. Such descriptions may also benefit sighted users, such as when accessing media eyes-free (i.e., via a voice agent such as Alexa or Cortana), and by providing additional metadata that can support information retrieval. Developing and evaluating source-dependent image descriptions based on the guidelines presented herein is a promising area for future study.

Acknowledgements

We thank the anonymous reviewers for their valuable input, study participants for their involvement, and Meng Zhang for his assistance with editing transcribed files. This work is supported by Microsoft gift funding.

REFERENCES

[1] Roobaea Alroobaea and Pam J Mayhew. 2014. How many participants are really enough for usability studies?. In *2014 Science and Information Conference*. IEEE, 48–56.

[2] Cynthia L Bennett, Martez E Mott, Edward Cutrell, Meredith Ringel Morris, and others. 2018. How Teens with Visual Impairments Take, Edit, and Share Photos on Social Media. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 76.

[3] Nilavra Bhattacharya, Qing Li, and Danna Gurari. 2019. Why Does a Visual Question Have Different Answers?. In *Proceedings of the IEEE International Conference on Computer Vision*. 4271–4280.

[4] Jeffrey P Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, and others. 2010. VizWiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*. ACM, 333–342.

[5] Jeffrey P Bigham, Richard E Ladner, and Yevgen Borodin. 2011. The design of human-powered access technology. In *The proceedings of the 13th international ACM SIGACCESS conference on Computers and accessibility*. ACM, 3–10.

[6] Erin Brady, Meredith Ringel Morris, and Jeffrey P Bigham. 2015. Gauging receptiveness to social microvolunteering. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 1055–1064.

[7] Erin Brady, Meredith Ringel Morris, Yu Zhong, Samuel White, and Jeffrey P. Bigham. 2013. Visual Challenges in the Everyday Lives of Blind People. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. ACM, New York, NY, USA, 2117–2126.

[8] Stacy M Branham, Ali Abdolrahmani, William Easley, Morgan Scheuerman, Erick Ronquillo, and Amy Hurst. 2017. Is Someone There? Do They Have a Gun: How Visual Information about Others Can Improve Personal Safety Management for Blind Individuals. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility*. ACM, 260–269.

[9] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[10] Michele A Burton, Erin Brady, Robin Brewer, Callie Neylan, Jeffrey P Bigham, and Amy Hurst. 2012. Crowdsourcing subjective fashion advice using VizWiz: challenges and opportunities. In *Proceedings of the 14th international ACM SIGACCESS conference on Computers and accessibility*. ACM, 135–142.

[11] Diagram Center. Specific Guidelines: Art, Photos & Cartoons. `http://diagramcenter.org/specific-guidelines-final-draft.html#20`. (No Date).

[12] Diagram Center and Touch Graphics. No Date. Decision Tree. `http://diagramcenter.org/decision-tree.html`. (No Date). (Accessed on 12/30/2019).

[13] World Wide Web Consortium. 2019. How to Meet WCAG (Quickref Reference). `https://www.w3.org/WAI/WCAG21/quickref/`. (October 2019). (Accessed on 01/02/2019).

[14] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, and others. 2015. From captions to visual concepts and back. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1473–1482.

[15] Kay Alicyn Ferrell, Silvia M Correa-Torres, Jennifer Johnson Howell, Robert Pearson, Wendy Morrow Carver, Amy Spencer Groll, Tanni L Anthony, Deborah Matthews, Bryan Gould, Trisha O'Connell, and others. 2017. Audible Image Description as an Accommodation in Statewide Assessments for Students with Visual and Print Disabilities. *Journal of Visual Impairment & Blindness* 111, 4 (2017), 325–339.

[16] Cole Gleason, Patrick Carrington, Cameron Cassidy, Meredith Ringel Morris, Kris M. Kitani, and Jeffrey P. Bigham. 2019. &Ldquo;It's Almost Like They'Re Trying to Hide It&Rdquo;: How User-Provided Image Descriptions Have Failed to Make Twitter Accessible. In *The World Wide Web Conference (WWW '19)*. ACM, New York, NY, USA, 549–559.

[17] 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*. ACM, 518.

[18] Anhong Guo, Ece Kamar, Jennifer Wortman Vaughan, Hanna Wallach, and Meredith Ringel Morris. 2019. Toward Fairness in AI for People with Disabilities: A Research Roadmap. *arXiv preprint arXiv:1907.02227* (2019).

[19] Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. 2018. VizWiz Grand Challenge: Answering Visual Questions From Blind People. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[20] Simon Harper and Alex Q Chen. 2012. Web accessibility guidelines. *World Wide Web* 15, 1 (2012), 61–88.

[21] AI Now Institute. 2019. Disability, Bias, and AI. `https://ainowinstitute.org/disabilitybiasai-2019.pdf`. (November 2019). (Accessed on 01/02/2020).

[22] Os Keyes and Cynthia L. Bennett. 2019. What Is the Point of Fairness? Disability, AI and The Complexity of Justice. *arXiv preprint arXiv:1908.01024* (2019).

[23] Jonathan Lazar, Alfreda Dudley-Sponaugle, and Kisha-Dawn Greenidge. 2004. Improving web accessibility: a study of webmaster perceptions. *Computers in human behavior* 20, 2 (2004), 269–288.

[24] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.

[25] 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. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 5988–5999.

[26] Valerie S Morash, Yue-Ting Siu, Joshua A Miele, Lucia Hasty, and Steven Landau. 2015. Guiding novice web workers in making image descriptions using templates. *ACM Transactions on Accessible Computing (TACCESS)* 7, 4 (2015), 12.

[27] Meredith Ringel Morris. 2019. AI and Accessibility: A Discussion of Ethical Considerations. *arXiv preprint arXiv:1908.08939* (2019).

[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*. ACM, 59.

[29] Meredith Ringel Morris, Annuska Zolyomi, Catherine Yao, Sina Bahram, Jeffrey P Bigham, and Shaun K Kane. 2016. With most of it being pictures now, I rarely use it: Understanding Twitter's Evolving Accessibility to Blind Users. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 5506–5516.

[30] Helen Petrie, 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)* 71 (2005).

[31] John R Porter, Kiley Sobel, Sarah E Fox, Cynthia L Bennett, and Julie A Kientz. 2017. Filtered out: Disability disclosure practices in online dating communities. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 87.

[32] Elliot Salisbury, Ece Kamar, and Meredith Ringel Morris. 2017. Toward scalable social alt text: Conversational crowdsourcing as a tool for refining vision-to-language technology for the blind. In *Fifth AAAI Conference on Human Computation and Crowdsourcing*.

[33] Elliot Salisbury, Ece Kamar, and Meredith Ringel Morris. 2018. Evaluating and Complementing Vision-to-Language Technology for People who are Blind with Conversational Crowdsourcing.. In *IJCAI*. 5349–5353.

[34] John M Slatin and Sharron Rush. 2002. *Maximum accessibility: Making your web site more usable for everyone*. Addison-Wesley Longman Publishing Co., Inc.

[35] Abigale J Stangl, Esha Kothari, Suyog D Jain, Tom Yeh, Kristen Grauman, and Danna Gurari. 2018. BrowseWithMe: An Online Clothes Shopping Assistant for People with Visual Impairments. In *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility*. ACM, 107–118.

[36] Anselm Strauss and Juliet Corbin. 1998. *Basics of qualitative research techniques*. Sage publications Thousand Oaks, CA.

[37] Kenneth Tran, Xiaodong He, Lei Zhang, Jian Sun, Cornelia Carapcea, Chris Thrasher, Chris Buehler, and Chris Sienkiewicz. 2016. Rich image captioning in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 49–56.

[38] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3156–3164.

[39] Violeta Voykinska, Shiri Azenkot, Shaomei Wu, and Gilly Leshed. 2016. How blind people interact with visual content on social networking services. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. ACM, 1584–1595.

[40] Alexandra Vtyurina and Adam Fourney. 2018. Exploring the role of conversational cues in guided task support with virtual assistants. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 208.

[41] Web Accessibility Initiative (WAI) W3C. Date. Web Content Accessibility Guidelines (WCAG) Overview. `https://www.w3.org/WAI/standards-guidelines/wcag/`. (No Date). (Accessed on 06/20/2019).

[42] Qi Wu, Peng Wang, Chunhua Shen, Anthony Dick, and Anton van den Hengel. 2016. Ask Me Anything: Free-Form Visual Question Answering Based on Knowledge from External Sources. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4622–4630.

[43] 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*. ACM, 1180–1192.

[44] Yuhang Zhao, Shaomei Wu, Lindsay Reynolds, and Shiri Azenkot. 2017. The Effect of Computer-Generated Descriptions on Photo-Sharing Experiences of People With Visual Impairments. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 121.

[45] Yu Zhong, Walter S Lasecki, Erin Brady, and Jeffrey P Bigham. 2015. Regionspeak: Quick comprehensive spatial descriptions of complex images for blind users. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 2353–2362.