Automatic Alt-text: Computer-generated Image Descriptions for Blind Users on a Social Network Service

Shaomei Wu
Facebook
shaomei@fb.com

Jeffrey Wieland
Facebook
jeffw@fb.com

Omid Farivar
Facebook
omidf@fb.com

Julie Schiller
Facebook
jschiller@fb.com

ABSTRACT
We designed and deployed automatic alt-text (AAT), a system that applies computer vision technology to identify faces, objects, and themes from photos to generate photo alt-text for screen reader users on Facebook. We designed our system through iterations of prototyping and in-lab user studies. Our lab test participants had a positive reaction to our system and an enhanced experience with Facebook photos. We also evaluated our system through a two-week field study as part of the Facebook iOS app for 9K VoiceOver users. We randomly assigned them into control and test groups and collected two weeks of activity data and their survey feedback. The test group reported that photos on Facebook were easier to interpret and more engaging, and found Facebook more useful in general. Our system demonstrates that artificial intelligence can be used to enhance the experience for visually impaired users on social networking sites (SNSs), while also revealing the challenges with designing automated assistive technology in a SNS context.

Author Keywords
User Experience; Accessibility; Artificial Intelligence; Social Networking Sites; Facebook.

ACM Classification Keywords
H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

INTRODUCTION
With more than 2 billion photos uploaded and shared across Facebook, Instagram, Messenger, and WhatsApp [29] each day, images are becoming a more prominent part of today's Internet experience, especially on SNSs [20, 30]. This increase in visual media may increase usability challenges for people with low vision or blindness who rely on screen readers. A few recent studies investigated the experience of visually impaired people with visual content on SNSs, highlighting both the desire and challenges for this group to engage with photos, and opportunities for web developers and designers [20, 30].

There have been a handful of applications that are designed specifically to help people with vision loss take or understand photos [1, 2, 8, 11, 12, 27, 34]. While the existing applications provide machine or human generated information about the images, most of them work on a small scale, have non-trivial latency, and cost money or social capital. In addition, these applications exist as independent services/apps that take ad-hoc queries from the user. Thus, these systems are suboptimal for application to SNSs, as they require users to constantly make active decisions about whether it is worthwhile to investigate a photo, which is both time-consuming and cumbersome. In this paper, we present the first real-time, large-scale, machine-generated image-to-text description system tested as part of the browsing experience of a mainstream SNS.

With automatic alt-text (AAT), screen reader users can browse their Facebook News Feed and hear a machine-generated description of each image they encounter as alt-text. The alt-text is constructed in the form of "Image may contain...", followed by a list of objects recognized by the computer vision system. The major design decisions include: the selection of object tags, the structure of information, and the integration of machine-generated descriptions with the existing Facebook photo experience.

We iterated our design through 4 rounds of formal in-lab usability study sessions and informal QA sessions with our visually impaired colleagues. We also evaluated the system with 9K screen reader users through a two-week field study, during which half of them (test group) had AAT enabled in their Facebook iOS app. All 9K users were invited to report their experiences via surveys after the study period. The majority of our test users were excited about AAT and expressed high interest in seeing it deployed more widely across Facebook and in other SNSs. Our field study showed that blind users of AAT self-reported that photos were easier to interpret and that they were more likely to "Like" photos on Facebook. We did not, however, observe a significant difference between groups in logged data for photo engagement actions, like "liking" or "commenting" on photos.

CSCW '17, February 25-March 01, 2017, Portland, OR, USA
 ISBN 978-1-4503-4335-0/17/03…$15.00
DOI: http://dx.doi.org/10.1145/2998181.2998364

Figure 1: Screenshot of Facebook News Feed with AAT.

The text in the white box is generated by AAT and read to a blind user by VoiceOver. AAT is normally invisible (since only read by a screen reader), but shown here visually for clarity.

Our research aims to leverage the power of artificial intelligence to enhance visually impaired users' experiences with photos on SNSs. The biggest challenge, as our study revealed, is serving users' desires for more information about the images, with a higher-quality and more socially-aware computer algorithm. For example, blind and visually impaired users would like AAT to provide more detailed image descriptions, including people's identities, emotions, and appearances. However, we must carefully design and evaluate the automated system to avoid social miscues and respect the privacy of those being described.

The major contribution of this work is to demonstrate that a large-scale, real-time, automatic alt-text generation system is technically feasible, as well as useful to blind users of SNSs. As part of the built-in photo infrastructure on Facebook, AAT provides free, additional information about photos that complements existing image description services [1, 8, 27], and makes people feel more included and engaged with conversations around photos on Facebook. By presenting the design and evaluation process of our system in details, we hope this work will shed light on future development of innovative accessibility features for SNSs and encourage the adoption of artificial intelligence into assistive technologies to enhance the experience of people with disabilities online and offline.

RELATED WORK

There has been a long line of research addressing web accessibility issues, mostly focusing on the usability of specific websites, products, and platforms [6, 13, 15, 16, 17, 18, 21, 23, 25]. These works, complemented by Web accessibility standards [31] and assistive technologies, have enabled more and more visually impaired people to use computers, mobile devices, and the Internet. However, the proliferation of visual content on the Internet, especially on SNSs, has introduced new challenges to visually impaired users participating in online conversations [20, 30].

While many existing systems caption images either with humans [1, 27, 34] or with computer algorithms [8, 9, 22], none of them fully satisfied the needs of visually impaired users on SNSs. The human-powered systems are constrained by scalability, latency, cost, and privacy concerns, while the automated systems have limitations on accuracy and generalizability on the variety of imagery on SNSs [20]. Our work contributes to this area of research by adapting a state-of-the-art computer vision system to the assistive technology context at scale. Working closely with visually impaired users of SNSs in our design process, we built an experience that proved to be enjoyable and useful to this community, and demonstrated the potential of artificial intelligence to assist people with disabilities.

Human-powered Photo Understanding Tools

Assistive technology researchers and developers have built several systems to help blind people understand photos [1, 8, 27, 34]. Most of these systems are human-powered, relying on friends, volunteers, or crowd workers to describe photos or answer visual-related questions submitted by blind users. Although these systems can provide high-quality content, they are not widely adopted by the blind community due to issues such as scalability, latency, cost, and privacy.

Scale. The number of photos that can be processed by these systems is constrained by the number of crowd workers and volunteers. For example, [4] reported having each volunteer

provide, on average, one high-quality answer every two days, which obviously does not scale with the rate images are being shared online today.

Latency: Crowdsourced services such as [27] take at least 30 seconds and up to a few minutes for each image request. Friendsourced services have even bigger variance and often greater latency. VizWiz, a service that utilizes both crowd workers and friends/volunteers, reported the average time of 133 seconds for a crowdsourced answer, and 58 minutes for a volunteer-provided answer [2]. Even the best case latency for such services is too slow for browsing SNSs.

Cost: Crowdsourced services, such as [27], charge a pay-per-photo or monthly subscription fee to cover their costs, while friendsourced services are free of charge. Research has found, however, that blind users often prefer crowdsourced services than friendsourced ones because of the perceived high social cost of owing favors to friends and appearing to be dependent on them [6]. Furthermore, the existence of a monetary/social cost imposes a mental cost to the user: she has to constantly make active decisions about whether or not to have an image described. This cost makes the experience more stressful and mentally taxing.

Privacy: When humans are in the loop, privacy is always a concern. Although research has shown examples of sensitive personal information inadvertently revealed in photos sent to sighted assistants, most of these systems can not detect or prevent situations like this from happening [3].

Our approach (AAT) complements existing human-powered services by addressing all of the issues above. Built on top of a state-of-the-art computer vision system [24, 26, 33], we process over **2 billion** images each day, generate descriptions at the speed of less than a second per image, with no human inspection of the images or monetary charge to users. More importantly, by using auto-generated descriptions for image alt-text, any photo a user encounters on Facebook can have the description read aloud by a screen reader automatically in real time, in addition to the existing metadata and descriptions generated by other services currently available (e.g. [2, 4]). This approach introduces the least friction and creates an experience that is most similar to sighted users' social media experience.

Automated Photo Captioning

There has been a series of work on automatic photo caption generation in the field of computer vision and AI research [9, 22]. However, all the existing algorithms for caption generation were designed and evaluated for sighted people, thus they face several challenges when applied to accessibility and assistive technology.

The direct use of caption generation for alt-text presents challenges, since there are nuanced differences between captions and alt-text. A desired caption for a sighted person can be very different from alt-text for a blind person. For example, a sighted person might not care about whether the salient objects are included in the caption (e.g., results presented in [9] do not mention salient objects such as people and tree), but a blind person would usually prefer alt-text that describes all salient objects in the image [20]. Even for systems that do try to capture most of the objects [8], they usually do not describe the theme of the overall image (such as "selfie", "landscape"), which can be important information for blind users as well.

Another challenge is to understand and mitigate the cost of algorithmic failures. While sighted users can easily ignore or correct the machine-generated caption, blind users of our system can not directly assess the quality of generated descriptions. The cost of system failure in our use case is much higher than in existing auto-captioning systems – e.g., blind users could be misled to make inappropriate comments about photos in which humans are mis-identified as "gorillas" [10]. The impact of such failure has never been studied in an accessibility context.

To address these issues, we designed our system to show only object tags with very high confidence. We conducted rounds of lab and field studies to understand the value/risk of different object tags and the potential impact to blind users' experiences.

We want to emphasize that the contribution of this paper is not to advance computer vision research for object/theme identification, but to present a useful, fast, free alt-text generation system for blind people that enhances their experience on SNSs. The most similar system to our knowledge is alt-text bot [8], which also employed AI technology to generate alt-text for web images. Its drawbacks are: (1) latency: it takes multiple seconds for alt-text bot to provide alt-text as it sends requests to the AI server on the fly, whereas alt-text in AAT is precomputed and available immediately on image focus; (2) availability: alt-text bot works only on the desktop web with a specific browser extension, whereas AAT is available on all mainstream platforms (web, mobile web, Android, iOS) without the need to install anything. This work is the first to report the design process and measure the performance (i.e., recall, accuracy, user feedback) of a system of this nature. By doing so, we aim to inspire more formal research on this topic and provide benchmarks for future development.

AUTOMATIC ALT-TEXT (AAT)

We designed two versions of automatic alt-text for our research: (I) a mock version similar to Facebook News Feed containing only stories with images, and AAT-generated descriptions added after each story; (II) an embedded experience in Facebook's News Feed with generated descriptions as alt-text for images (see Figure 1). Version I was mainly used for development and in-lab user studies while version II was used for the field study.

Each description sentence starts with "image may contain", followed by concept tags. The major design choices were:

(1) Which concept tags to include;
(2) How to organize and present the detected concept tags;

(3) How to integrate the new information with the existing Facebook experience.

We leveraged feedback from in-lab user studies and a field study to iterate on our design around these decisions. Informally, we also ran numerous quality assurance (QA) sessions within our research team to tweak the experience. We received a tremendous amount of valuable feedback from one of our blind colleagues who has been an enthusiastic test user of our system since day one.

The following subsections will discuss each of our design choices in-depth.

Selection of Tags

The object-recognition service we used was trained in a similar fashion to the models described in the works of [24, 26, 33]. While the system can be trained to recognize a wide range of objects and themes (both referred to as "concept" in the rest of this paper), we hand-selected a set of 97 concepts based on their prominence and frequency in photos, as well as the accuracy of computer vision models.

Our first selection criteria for these concepts was their prominence in Facebook photos. To measure what was "most prominent", we took a random sample of 200K public photos on Facebook and had 30 human annotators annotate 3 to 10 things in each photo. Since there is potentially an infinite number of details in a given photo, we intended to capture the most salient objects or visual characteristics by limiting the number of tags used for each photo annotation. An example annotation could be: "man, pink shirt, jeans, chair, curtain". Each annotator labeled between 1K and 15K photos, while each photo had up to 2 annotators labeling it. After simple tokenization and stemming of tags, we sorted all the tags by frequency and took the top 200 as concept candidates.

We then filtered the concept candidates that could have a disputed definition or that were hard to define visually. Those concepts mainly fell in the following 3 categories:

(1) Concepts associated with personal identity. For example, we decided to leave out gender-related concepts such as *woman*/*man*, *girl*/*boy*, as gender identification is more complex, personal, and culture-depdendent than what may appear visually on photos.

(2) Adjectives, such as *red*, *young*, *happy*. These concepts are fuzzy and context-dependent. They also tend to be attributes of other concepts rather than of the image.

(3) Concepts that are challenging for an algorithm to learn. For example, the concept *landmark* (by definition) includes any object that can be recognized from a distance and be used to establish a location. In practice, landmarks on photos vary so much in shape and size that this concept is hard to define visually and hard to distinguish from general non-landmark buildings, resulting in poor performance by computational models.

In addition, we ran a network clustering algorithm on the network of concepts. Each node represented a concept and was weighted by the number of images containing that specific concept. Edges represented the number of images co-occurring between any pair of concepts. The purpose of this was to uncover the major categories of tags (i.e., people, clothing, animal, nature, food, text/meme) and to ensure that our selected concepts covered them all.

We ended up with a list of 97 concepts that provided different sets of information about the image, including people (e.g., *people count*, *smiling*, *child*, *baby*), objects (e.g., *car*, *building*, *tree*, *cloud*, *food*), settings (e.g. *inside restaurant*, *outdoor*, *nature*), and other image properties (e.g., *close-up*, *selfie*, *drawing*). The user studies suggested that a relatively high accuracy is required for blind users to trust our system and find it useful, so we set the current confidence level to be approximately 0.8 for all concepts (a few more sensitive concepts have the confidence level as high as 0.98), which gave us at least one tag for over 90% of photos viewed by visually impaired users on their News Feeds. We used a method to identify visually impaired users similar to [32]. Our object recognition models were evaluated with human-labeled data to ensure the confidence score was calibrated and represented model precision. The recall is especially high for *people count*, which is present in about half of all photos viewed. However, the algorithm would sometimes miss people when their faces were small or side-facing. Tags about setting and themes (e.g. *indoor*, *night*, *selfie*, *close-up*) also have very high recall (usually above 90%) even with our current precision threshold, as they are prominent characteristics of the image that are easier for neural networks to detect. The recall rates for object tags vary, largely depending on the size of the object and their location in the image – the algorithm performs better with bigger objects such as *sky* and *mountain*, and worse with smaller objects such as *ice cream* and *sushi*.

Construction of Sentence

Alt-text for photos provides a description of the image, and is read back to screen reader users when the screen reader gains focus on the photo (e.g., "Windmill by the sea"). Traditionally, alt-text is provided by web administrators or mobile application developers for the images used in their websites/applications. Previously, most photos on Facebook have the default alt-text "[Name of photo owner]'s photo". With AAT, our goal is to algorithmically generate descriptive alt-text for user-generated images on Facebook. Our lab studies with the prototype showed that blind users preferred to hear the constructed alt-text in a form of a complete sentence rather than a mere list of tags, as it sounds more natural and friendlier. However, limited by the accuracy and consistency of existing natural language caption generation systems [9, 22], we decided to forgo the free-form sentence approach and construct a generic sentence with a formula for better robustness and accuracy. The sentence always starts with "Image may contain:", and followed by a list of tags. Using the word "may" was

Pseudonym	Gender	Age	Level of Vision Loss	Occupation	Screen Reader Usage	Facebook Usage
Michael	M	~30	Blind since 2 years ago	Animator	VoiceOver on Mac/iOS, Jaws on Windows	m-site or iOS app
Daisy	F	18	Some vision	Student	VoiceOver on Mac/iOS	iOS app
Christy	F	~35	Blind	Dorm Counselor	VoiceOver on Mac/iOS, Jaws on Windows	m-site, iOS app
John	M	~40	Blind since 30	Assistive Technology Trainer	VoiceOver on Mac/iOS, Jaws on Windows	m-site, iOS app, desktop web

Table 1. Lab study participants and their usage of Facebook and screen reader.

intended to convey uncertainty, since the underlying object recognition algorithm can not guarantee 100% accuracy. The formulated structure also provided another advantage of our system: the sentence can be easily constructed in any written language as long as the concept tags are translated[1]. This can be especially meaningful for small languages that lack ovolunteers or crowd workers for human-powered image description services.

One of the biggest challenges for our system was communicating uncertainty. We experimented with different approaches, such as showing the confidence level (e.g. 50%, 75%, "low", "medium") for each tag. However, as some of our test users noted, having an accuracy number/level per tag was cumbersome and hard to interpret. Consequently, we decided to only show tags that were above a relatively high confidence threshold (minimal 0.8) for a smoother experience.

Another design decision we iterated on was how to properly organize concept tags. After experimenting with different tag ordering methods (e.g., ordering tags by confidence level) with in-lab study participants and both our sighted and blind colleagues, we ended up ordering tags by categories: people first, followed by objects, then finally settings/themes. That is, when faces are detected in the image, the list of objects will always start with the count of people, followed by whether they are smiling or not[2]. The reason we put people first is because almost all participants indicated that people are the most interesting part of the image, especially their mood and what they were doing (action). We do not yet have action tags besides *smiling*, but once the model is trained to detect other actions those tags will appear in this first group. Given that people are described first, we put tags about the settings/themes after tags about physical objects in the photo so that we do not jump between things and themes conceptually. Within each category, the tags were ordered by the confidence level, as it also reflected the prominence of each object/theme. In this way, the constructed sentence is consistent, engaging and easy to understand.

Build a Seamless Experience

One advantage to using machine-generated descriptions is that a user does not need to take any action to have a photo described. While "on demand" requests for information about a photo or user supplied alt-text can be useful, machine-generated descriptions provide unmatched coverage and convenience for photos on large-scale services such as Facebook. Also, since this is the first time the machine-generated automatic photo descriptions are built into a mainstream SNS, we want to make sure that it is as lightweight and unobtrusive as possible, so that it enhances blind users' existing Facebook experience without a learning curve.

IN-LAB USABILITY STUDIES

We ran user studies with 4 participants using version I of our system in a lab setting (see Table 1), spanning two months. Our major research questions for the in-lab studies were: (1) what is the added value of computer-generate photo descriptions to blind users of Facebook; (2) what is the risk of providing incorrect descriptions and how to mitigate such risk. We used the first study session as a proof of concept, and used the later 3 sessions to fine-tune the user experience and design. We stopped after the 4th session for two major reasons: (1) the feedback we received started to become repetitive; (2) some research questions are hard to answer due to the limitation on space and time during in-lab studies. In the end, since we recruited all our participants through *LightHouse for the Blind and Visually Impaired*, a San Francisco based organization that provides rehabilitation services and technical training for the blind, we want to point out that our interview study participants may be more tech-savvy than an average screen reader user.

Interview Protocol

The participants were interviewed in 90-minute, one-on-one sessions, which were transcribed and coded. We asked participants how they currently use technology and social media, especially their experience with photos on Facebook. In particular, they were asked to describe situations where encountering images improved or hindered their experience on Facebook.

Next, the participants evaluated 5-8 photos in their mocked News Feed. After hearing the photo as presented on Facebook, we asked what they thought was in the photo,

[1] AAT became available in 20 langauges within months after we launched the English version. See all lanugages at https://www.facebook.com/help/216219865403298

[2] We describe "people smiling" when we detect smile in at least one of the faces.

and to rate their confidence on a scale of 1 to 5. Subsequently, we let participants hear the generated alt-text and asked them again to rate their confidence given the new information.

We tested in two scenarios, one with concept tags with confidence level 90%, and then again with a lower confidence level threshold (either 50% or 75%). These experiment thresholds (90%, 75%, 50%) provided qualitatively different levels of recalls (low, medium, high) and precisions (high, medium, low, respectively). These conditions were counterbalanced between participants so that some received the high confidence treatment first and others the low confidence treatment.

Finally, the participants were asked for their comprehension (i.e. "how would you describe what you've experienced today to another friend who uses a screen reader?") and general feedback on AAT.

In-Lab Usability Findings

Across the semi-structured interviews, the overall reaction to AAT was positive: a frequent response migrated from disinterest in photos to intrigue. For example, in the end of her interview, Christy said, "*I would definitely pay more attention because I'd be like 'oh, interesting.' Especially if you guys made some changes and added some more information […] I would definitely look at the photos a little more.*" Although participants did sometimes run into confusion and second-guessed themselves when the computer vision algorithm made mistakes, they were still interested at the system and willing to accept imperfections. This positive experience was rooted in an increased awareness of photographic content, increased ability to interact with their social network, and a pleasant surprise at the technical possibilities.

Increased interest on photos

Participants had a perception that their News Feed contained 40-60% photos and most participants reported they ignored these photos unless they were particularly well described in the user supplied description or came with additional metadata such as place check-ins or tags of people. Daisy for example said, "*If I can't tell what it is, I usually ignore it*". Similarly John reported, "*If I can't glean something from either the check in or their caption I'm likely just to go by it.*" They mentioned feelings of isolation and frustration. Christy said, "*Sometimes people post an image and it's really significant to them and everybody's commenting on it […] and I have no idea what it is.*" However, after using the system over the course of the interview they reported increased awareness of photos. John said, "*I probably haven't put this much thoughts into […] images on Facebook until today.*"

Low-cost information gain

The additional information provided by AAT appeared to reduce the time and social cost for our participants to interact with photos. Currently, the participants used all available information to help them better understand a photo, including related user comments for context. However, this is often prohibitively time-consuming. For photos without descriptions or comments, asking sighted people for help is not always viable or socially acceptable. Michael mentioned that he would only ask his girlfriend to describe photos to him. Christy reported, "*I've asked my friends if they can please describe their images but I'm not always - not everybody is going to remember and I'm not going to ask all the time because I would feel like a jerk.*"

Before hearing the generated alt-text, participants rated at most two photos with high confidence for what the photo was about, about 10-20% of the photos they inspected during the lab session. This ratio aligns with what is observed on Twitter [20]. With AAT, they reported higher confidence in some photos, especially the poorly-described ones with uninformative or missing captions. However, AAT also provides value for photos with good descriptions and metadata as it can corroborate this existing information.

Areas of improvement

Basing on both explicit feedback from the participants and our observations of them interacting with the prototype, the major area of improvement for the prototype was descriptive detail. Christy said, "*I'd rather have additional information and take a chance that it would be wrong than have no information at all.*" All participants wanted more tags. Participants specifically wanted more tags related to the main actors and key pieces of context such as actions, emotions, clothing, and overlaid text. When asked about the optimal confidence level for tags, participants were willing to forgive the algorithm and take the generated tags with "*a grain of salt*". John said, "*You might be wrong. But I could say that I'm willing to live with you being wrong 40 percent of the time.*" However, a lab setting is not ideal for assessing the impact of an errant tag.

Overall, we found lab studies most useful for concept testing and identifying usability issues in the existing design. We consistently observed that computer-generated alt-text provided valuable context when accuracy was high. We also fixed the issues that surfaced. For example, some tags provided by the computer are odd-sounding (e.g. "has person") or confusing (e.g. "synthetic" for images with added graphics or visible edits); we dropped some of these and rephrased others: "has person" was rephrased as "one or more people", and "synthetic" was dropped. We also originally presented the algorithm confidence level for each tag, but removed these numbers per Michael's suggestion, as "*it draws attention to the technology, not the photo*".

Limitations of in-lab studies

There were two important questions that remained hard to answer in our lab studies, they were:

(1) How can we balance the amount of information provided by the computer algorithm against the risk that such information might be wrong or even offensive?

All participants answered that they would like to have more tags, but had trouble anticipating the consequence with wrong or potentially offensive tags in the lab setting (a distinct possibility as we lower precision of the system to increase the recall). Sometimes the tag itself is not offensive but the connection between an object and a tag is (e.g. labeling a "person" as "horse"). Also, since blind users cannot easily verify the accuracy of the tags on the spot, the effect may not be immediate or even within Facebook but might take place in the future and/or in other settings. As a result, we were not able to determine the optimal thresholds for concepts through in-lab studies: in most cases, our participants were not able to articulate what changed after we adjusted the thresholds, or decide on which threshold setting they preferred. We thus designed the system with easy-to-tune threshold parameters for each tag, with initial settings (confidence = 0.8) that achieved both high precision and good coverage. In this way, we can bootstrap the optimal thresholds and tune the parameters reactively based on user feedback.

(2) How would the new information change the way that people with vision loss interact with photos on Facebook?

Since our interview protocol was designed to emphasize blind users' understanding of the **content** of photos, participants were probed to spend more time going over the entire photo story and determine what the image was about. This is very different from how they interact with photos on Facebook normally, as many participants reported that they routinely skipped photos in their News Feed. As a result, it is difficult for us to assess how blind users' interactions with photos would change in a natural setting: would they spend more time on photos? Would they like or comment on photos more often, given the extra information from AAT? To answer these questions, we designed and conducted a large-scale field study with 9K visually impaired Facebook users. We will describe the field study and its results in detail in the next section.

FIELD STUDY

To better understand the effectiveness of AAT beyond lab interview sessions, we also designed a two-week field study in which we piloted this feature for thousands of visually impaired Facebook users. The goals of our field study were:

(1) Evaluate the user experience of automatic alt-text in a more naturalistic setting;

(2) Assess the effect on photo interactions over a longer time period and in a larger scale.

Study design

We took a sample of 9K visually impaired users using a similar method as in [31] but with further restrictions:

- Used Facebook iOS App with VoiceOver at least once a week for 4 consecutive weeks prior to the study;
- Have viewed at least 10 photos through Facebook iOS App in the 4 weeks period prior to the study;
- Have their Facebook account language setting set to English (either US, UK or India English);
- Currently living in US, UK, Canada, Australia, or New Zealand according to their Facebook profiles;
- Had their Facebook iOS App updated in the past 3 months prior to the study.

These criteria were not meant to exclude users from different regions or backgrounds. They were intended to provide some control for various dimensions of our sample, such as Facebook activity, language comprehension, and tech-savvy level, thus help us to reduce the possible variance introduced by these dimension and achieve more statistically meaningful results.

Once sample was identified, we then randomly divided the sampled users into two groups (control and test), and enabled AAT for the test group for two weeks. During these two weeks, users in the test group would hear the machine-generated photo description when encountering posts containing images(s) in the Facebook iOS App (e.g., photo uploaded by friends, status update with a photo attached, friend's profile picture change); whereas the users in the control group experienced no alt text changes (i.e., getting the default alt-text "someone's photo"). To ensure a fair comparison between the control and the test groups, users were not notified for the study during the study period or informed of which condition they were assigned.

After the two-week blind test period, we emailed users from both groups with an identical solicitation to complete an online survey [3] to provide feedback on their experience viewing and interacting with photos on Facebook. We informed all users that their responses to the survey would be anonymized and potentially published for academic research. We incentivized the completion of the survey by a raffle of ten $100 online retail gift cards.

We designed the survey to understand both the general experience of photos on Facebook and specifics about AAT. The survey can be broken down into 4 parts:

First, it started with simple screening questions about the participants' vision and assistive technology use.

Second, for those who answered that they were blind or visually impaired and used a screen reader, we asked about their current experience on Facebook regarding photos, including: how easy is it to interact with photos on Facebook; how useful is Facebook overall; as well as an open-ended question inviting them to input free text about their photo experience on Facebook. Those questions allow us to control on the general sentiment towards Facebook in

[3] We ensured that the survey is fully accessible with major screen reader software on most mainstream browsers and platforms. Both the email and the survey were in English.

the survey analysis, and benchmark the impact of future accessibility features on Facebook.

Third, for those who said that they were blind, we asked them to recall the last photos they encountered on Facebook and answer additional questions about their interactions with those photos, including: how easy or hard it is to tell what the photos are about; whether they have "liked" any of those photos; whether they commented on any of those photos; and whether they noticed their screen reader read back information starting with "Image may contain…".

Fourth, for those who indicated that they had noticed the new information associated with photos, we asked them to pick word(s) to describe how they felt after hearing the new information, and to answer an open-ended question about what they liked or disliked about the photo experience with the new information.

In addition to the self-reported survey data, we also extracted behavioral logs of photo interactions for sample users and compared the test and the control groups (all data was anonymized and de-identified).

Field Study Findings

In the survey feedback, we saw statistically significant differences in both sentiment and engagement with photos across the control and test groups. These differences were not born out in behavioral log data, however.

It is worth noting that although both groups actively responded to the survey, the response rate for the test groups was *significantly higher* than the control group. We received 250 and 298 completed responses from the control and test group, respectively[4]. The response rates were 5.9% for control and 7.26% for test, with two-sided z-test rejecting the null hypothesis at a 98% confidence level.

The answers to the screening questions confirmed that most of the users in our sample were indeed visually impaired, with only 7 out of all 556 survey respondents who self-reported as fully sighted but still used a screen reader for various reasons such as "*it is a lot less tiring for my eyes*".

AAT enhanced people's overall experience with Facebook: the test users indicated that Facebook was more useful to them in general versus the control group.

In terms of photo experiences, the test group reported photos were easier to interpret and that they were more likely to "like" them.

Within the test group, people responded with overwhelmingly positive sentiment about their experience with AAT. Over 90% of the write-in feedback from the test group acknowledged that AAT is "*useful*", "*a step in the right direction*", and makes Facebook "*more inclusive*". Users in the test group were also very enthusiastic about providing feedback and suggestions to further improve the feature – about 64% of them provided write-in feedback on what they liked and/or did not like about AAT. While most of the suggestions were for more detailed descriptions, the test users also raised questions about the accuracy of the algorithm and potential privacy issues.

General Facebook Experience

The test group found Facebook to be more useful overall. In Figure 2, we see that the distribution of answers from the test group is shifted to the right, with more people in the test group reporting Facebook to be more useful to them. More specifically, over 76% of test users found Facebook "very" or "extremely" useful, whereas only 63% of the control group felt the same. And the difference in ratio between control and test group is statistically significant ($z=3.6$, p-value < 0.0005).

The write-in answers to the open-ended question about current experiences with photos on Facebook also confirmed this appreciation of the system and its added value to visually impaired SNS users. For example, one test user wrote, "*The effort, from Facebook, to make photos better for vision impaired people, is greatly appreciated.*" Another user experienced the difficulty of having friends provide descriptions to their photos in a consistent manner and mentioned the benefits of having automatic descriptions, "*I am glad to see the effort Facebook is applying. I was surprised when I first noticed that Facebook was providing alt tag-like information about pictures. I have been working for some time with my Facebook friends to get them to give a little description about their photos for my benefit. Sometimes it works, sometimes it is not.*"

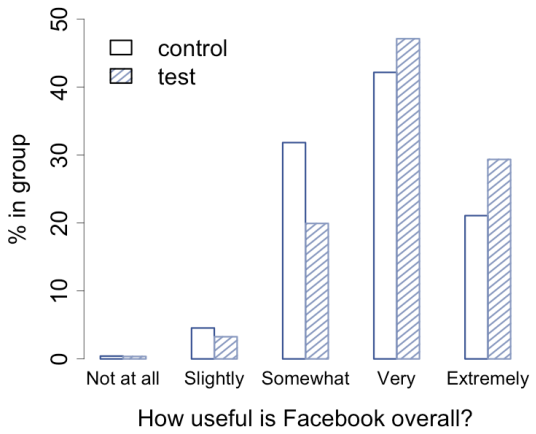

Figure 2. Distribution of survey responses on how useful Facebook is overall

[4] Note that we asked the survey respondents to leave their names and email addresses upon completing the survey to enter the raffle, but 8 people filled neither of these two fields therefore we could not identify their group assignment. We dropped their responses from our analysis.

The responses from the control group on the same question expressed their frustration when photos are not described (consistent with findings from [30]), but also mentioned several existing features that already made it easy for people with visual impairments to interact with photos. One participant in the control group mentioned, "(*I enjoy the) ease on uploading, tagging and sharing. Tagging feature especially useful for VI people who cannot identify themselves or others in photographs.*"

Engagement with Photos

As described before, survey participants who self-reported as blind in the screening question were also asked to recall and report their recent engagement with photos on Facebook. This was intended to control for visual impairment conditions and reduce variance, as the way people interact with photos can vary drastically depending on the types of visual impairment they have. There were around 75% of users in both groups who identified themselves as blind and users of screen readers. We further filtered out users in the test group who answered "No" or "I don't remember" to the question "whether you have noticed the new information about photos starting with 'Image may contain…'", as well as users in the control group who answered "Yes" to this question (possibly due to sharing devices) to ensure that the effect of AAT is present for the test group but not for the control group in our comparisons. As a result, we have 172 users in the control group and 203 in the test group (roughly 68% for both groups). All the results in this subsection were based on responses from these 375 users.

The blind users in the test group found it easier to understand the content of photos. They also self-reported as much more likely to "like" photos, and slightly more likely to "comment" on photos. However, we did not observe this reported increase in photo likes or comments by the test group as compared to the control group in the logged data.

Figure 3 shows the comparison between the two groups on how hard it is to interpret photos. Comparing to the test group, users in the control group found it much harder to interpret photos they encountered on Facebook: we see that over 75% of users in the control group found it extremely hard to understand what is in the photos, while less than 40% of the test group are in this category. Moreover, nearly a third (27%) of test group reported that it was easy or somewhat easy to tell what the photos were, whereas almost none of the control group users felt this way.

Knowing that the test group had an easier time understanding the content of photos, it is not surprising that they also reported a significantly higher likelihood of "liking" photos (see Figure 4). However, we do not see a statistically significant difference in the self-reported likelihood of commenting on photos (see Figure 5), possibly due to the rarity of commenting events and the effort involved with commenting.

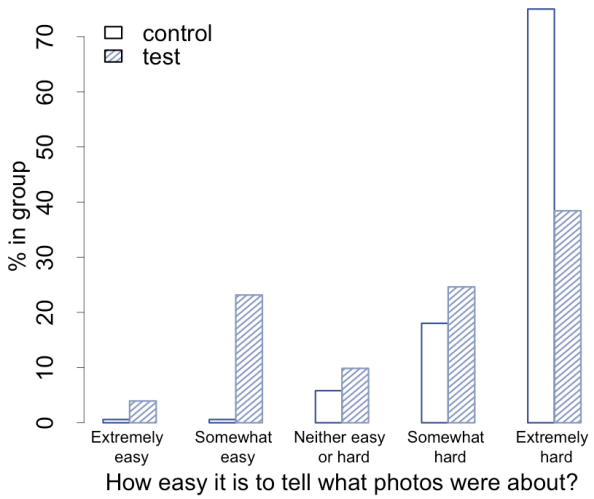

Figure 3 Distribution of survey responses on the level of difficulty to understand the content of photos on Facebook

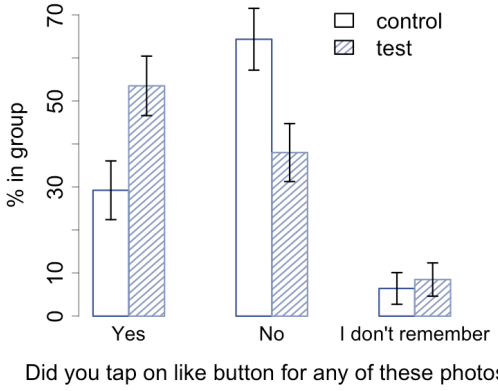

Figure 4 Self-reported data on the occurrence of photo "like" on Facebook. Error bars show binomial distribution on probability with 95% confidence interval.

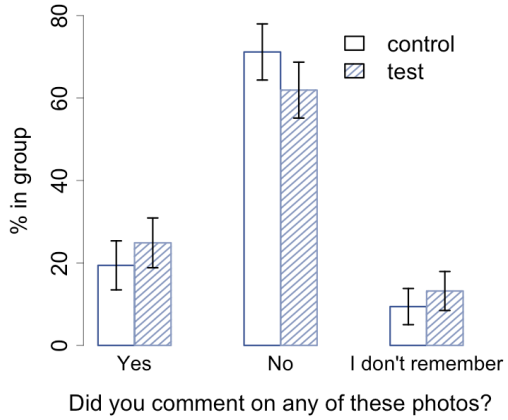

Figure 5 Self-reported data on the occurrence of commenting photos on Facebook. Error bars show binomial distribution on probability with 95% confidence interval.

To complement the self-reported photo interaction data, we also looked at the logged photo-related activities for all field study users. We anonymized the logged activity data and only analyzed in aggregation at the group level. Contrary to the self-reported data, we did not see a significant difference in the number of photos liked by users from the test group as compared to the control group. Similar null results were also found for the number of photos viewed, commented, and shared during the two-week study period. The discrepancy in the self-reported data and the logged data is interesting, and potentially due to several reasons: (1) there is a limitation on both the number of Facebook updates with photos and the amount of time a person spends on Facebook each day; those limits may largely determine the number of photos views for users in both groups; (2) our user sample is too small to observe the actual effect: a power analysis shows that our intervention (i.e., introduction of AAT on the test group) needs to increase test users' likelihood of liking a photo by more than twofold to generate statistically significant results given our sample size; (3) the amount of additional information provided by AAT is not sufficient to significantly influence a user's decision on whether to like or comment on a photo, as many other social factors are involved when making these decisions [30].

Self-reported Experience and Feedback for AAT

We provided 9 candidate feeling words (3 positive, 5 negative, 1 neutral) that were randomly ordered for each survey respondent and asked them to pick one or more words that described their feelings after experiencing AAT. The majority of the test users chose words that were positive, such as "happy", "impressed", and "surprised" (see Table 2). Survey participants could also submit words that are not in the provided list, and the top submitted word was "hopeful" (6 people), followed by "curious" (3) and "disappointed" (3).

Word	% in test group
happy	29.4
surprised	26.3
impressed	24.5
confused	6.9
other	4.9
annoyed	4.3
anxious	2.2
sad	1.4
afraid	0.2

Table 2. Distribution of chosen words for "How did you feel after hearing this alt-text (check all that apply)?"

We also had 3 independent raters code the write-in feedback about AAT from the test users into 3 themes: *useful*, *not useful*, and *improvements*. We understand that the write-in feedback may be biased since this user community might provide more positive subjective feedback about the products/features to show appreciation for accessibility efforts [28]. This feedback is nevertheless valuable as it provided deeper insights on what people liked about this feature and how they wanted it to improve.

Useful: over 90% of the write-in feedback confirmed that AAT is useful. The major reasons were:

(1) better understanding of images: "*I like how it describes what is in the picture and how many people and what the people are doing if they are smiling etc.*"

(2) make people feel included: "*For the first time, I feel like I can enjoy FB like my sighted friends.... Like I am really part of the community.*"

(3) show SNSs' efforts on improving accessibility: "*I really appreciate that Facebook is making an effort to make photos more accessible to those using screen readers*".

Not useful: a few people also expressed their disappointment in this feature, citing a lack of descriptiveness and uncertainty on how much they can trust the algorithm. For example, one user wrote, "*The descriptions are incredibly vague, and don't really give any information. I still have to ask friends what the photo actually is*". And another user asked for *"[…] more accuracy. I saw one photo that only stated it was indoors when in fact the person that added a description said it was some guide dogs out on the front porch of a restaurant. I do appreciate that FB is trying to work out this type of thing; I just hope it gets a little more reliable.*"

Improvement: acknowledging that this feature is only in its infancy, almost all users offered suggestions on how AAT can be improved in their write-in feedback. These suggestions concentrated on the following two categories (number after each category indicates the percentage of users requested for this improvement):

(1) extract and recognize **text** from the images (29%);

(2) provide more details about **people**, including their identity, age, gender, clothing, action, and emotional state (26%).

Other requests covered a wide variety, such as expanding the vocabulary of the algorithm, increasing the recall for existing tags, and making AAT available in more languages and platforms.

DISCUSSION

Our findings showed that blind users on a social networking site would benefit from a large-scale, real-time system that automatically generates image alt-text using machine learning. As one of the first such systems embedded in a mainstream social networking site, this feature proved to be

of value to blind users and received very positive feedback, while also surfacing several design challenges.

The first challenge for AAT is the trade-off between descriptive quality and algorithmic accuracy. As one of the lab study participants (John) suggested, one possible solution to this dilemma is to outsource the decision power to the user. For example, users could set a parameter that decides whether they want to see more tags with lower accuracy or fewer tags with higher accuracy. This is particularly nuanced, since blind users often cannot judge the accuracy of machine-generated descriptions, thus might not be able to adjust the setting for the desired experience robustly. Also, while lab study participants expressed their willingness to see more, but less accurate tags, feedback from our field study suggested that our users would not trust our system when the accuracy is poor. We also explored the possibility of first presenting the candidate photo description to the photo owner. However, since not all uploaded photos are viewed, there could be a significant amount of work wasted on photos that are not consumed by blind users. The photo owner also needs to understand the difference between alt-text for blind users versus captions for sighted people, which may again create a feeling of social debt for blind users. Ultimately, we would face many of the scale/latency issues in human-powered systems if all auto-generated photo descriptions needed to be verified by photo owners or subjects.

The decision to design an AI system that acts *on behalf of* the photo owner to describe to blind people what the image is about leads to our discussion about *agency*. While algorithm-curated user-generated-content (UGC) on today's web is increasing (e.g. trending topics on Twitter and Facebook, relevant articles recommended by Medium), the application of machine intelligence to descriptive tasks is perhaps more sensitive and bolder. This issue is exacerbated by the fact that our users consistently asked for richer and more intelligent descriptions that cover "*all important elements*" (survey respondent). As the boundary between computer-generated descriptions and user-provided descriptions becomes harder to distinguish, we need to better understand the perspectives of photo owners and the implication on creative ownership.

Another design challenge is to set the boundary of algorithms. Universally, participants desired richer descriptions for photos. Expanding the concept space will require careful design considerations. For instance, visually impaired users were eager to know about people in photos, such as their identity, emotions and appearance. However, as identity, emotion, and appearance are personal, social, and fluid, it is extremely difficult to train computer algorithms to interpret these concepts in context. Meanwhile, knowing that to provide more accurate description on these dimensions, we need to train the algorithm with more detailed personal data, should we proceed? For example, while facial recognition turned out to be one of the most requested features from our participants and current technology is viable, we chose to wait before implementing this feature in our current system, as there are privacy implications that will need to be assessed and accounted for before we incorporating face recognition in AAT in a responsible way. We understand how much our blind users would appreciate this feature, since sighted people could recognize friends and public figures in photos even when their identities are not tagged or labeled. We are looking forward to conducting and seeing more research on this issue.

CONCLUSION AND FUTURE WORK

We built a system that uses computer vision technology to identify basic objects and themes in photos on Facebook, and constructs alt-text using the identified concepts. We evaluated our system in lab sessions with 4 visually impaired participants, and in a 2-week field study with 9K screen reader users. Our user studies provided strong evidence that our system will be useful to the vision loss community, and will make their experience with photos on social networking sites more enjoyable. We also explored two key dimensions of such a system: the trade-off between algorithmic accuracy and descriptive power, and the potential impact on people's interactions with photos and friends. While we tried to assess longer-term effects of our system on on interaction with photos on Facebook, we could only see the effect on photo engagement via self-reported survey responses but not in behavorial logs.

For future work, we are investigating constructing captions that not only list objects and themes but also reveal the relationship among them (e.g., "A person and a dog next to a table in a café"). We are also training and evaluating more concepts to expand our image description vocabulary, especially concepts related to people such as their actions. With the necessary privacy measures in place, we also plan to provide an open API for our computer vision system to share its value and invite human input to improve our algorithm.

ACKNOWLEDGEMENTS

We thank Matt King for being a dedicated test user and for the timely, helpful feedback. We thank Brett Lavalla and Hermes Pique for implementing our system on Android and iOS. We thank Facebook AI Research, especially, Dario Garcia Garcia, Manohar Paluri and their colleagues for developing state-of-the-art computer vision technology. We thank our participants for their valuable insights, incredible patience, and encouraging words. Lastly, we thank our reviewers for feedback and suggestions.

REFERENCES

1. Jeffrey P. Bigham, Samuel White, Tom Yeh, *et al*. VizWiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd annual ACM symposium on User Interface Software and Technology* (UIST '10), 333-342.

2. Erin Brady. 2015. Getting fast, free, and anonymous answers to questions asked by people with visual impairments. *SIGACCESS Access. Comput.* 112 (July 2015), 16-25.
3. Erin Brady and Jeffrey P. Bigham. 2015. Crowdsourcing Accessibility: Human-Powered Access Technologies. *Found. Trends Hum.-Comput. Interact.* 8, 4 (November 2015), 273-372.
4. 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* (CHI '15).
5. Erin Brady, Meredith Ringel Morris, Yu Zhong, Samuel White, and Jeffrey P Bigham. Visual Challenges in the Everyday Lives of Blind People. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '13). 2117-2126.
6. Erin L. Brady, Yu Zhong, Meredith Ringel Morris, and Jeffrey P. Bigham. 2013. Investigating the appropriateness of social network question asking as a resource for blind users. In *Proceedings of the 2013 conference on Computer Supported Cooperative Work* (CSCW '13). ACM, New York, NY, USA, 1225-1236.
7. Maria Claudia Buzzi, Marina Buzzi, Barbara Leporini, and Fahim Akhter. 2010. Is Facebook really "open" to all? *International Symposium on Technology and Society, Proceedings, IEEE*, 327–336.
8. Cameron Cundiff, Alt Text Bot. Retrived Aug 06, 2016 from http://connectability.devpost.com/submissions/37785-alt-text-bot
9. Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. 2010. Every picture tells a story: generating sentences from images. In *Proceedings of the 11th European conference on Computer vision: Part IV* (ECCV'10), Kostas Daniilidis, Petros Maragos, and Nikos Paragios (Eds.). Springer-Verlag, Berlin, Heidelberg, 15-29.
10. CBSNews. Google Apologize for Mis-tagging Photos of African Americans. Retrieved Oct 08, 2015 from http://www.cbsnews.com/news/google-photos-labeled-pics-of-african-americans-as-gorillas/
11. Google. Search for pictures with Google Goggles – Search Help. Retrieved Oct 08, 2015 from https://support.google.com/websearch/answer/166331?hl=en
12. Chandrika Jayant, Hanjie Ji, Samuel White, and Jeffrey P. Bigham. 2011. Supporting blind photography. In *Proceedings of the 13th International ACM SIGACCESS Conference on Computers and Accessibility* (ASSETS '11). ACM, New York, NY, USA, 203-210. DOI=http://dx.doi.org/10.1145/2049536.2049573
13. Shaun K. Kane, Jessie a. Shulman, Timothy J. Shockley, and Richard E. Ladner. 2007. A web accessibility report card for top international university web sites. In *Proceedings of the 2007 international cross-disciplinary conference on Web accessibility* (W4A '07), 148-156. http://doi.acm.org/10.1145/1243441.1243472
14. Richard E. Ladner, Beverly Slabosky, Andrew Martin, *et al*. Automating Tactile Graphics Translation. In *Proceedings of the 7th International ACM SIGACCESS Conference on Computers and Accessibility* (ASSETS '05), 150-157. http://dx.doi.org/10.1145/1090785.1090814
15. Jonathan Lazar, Aaron Allen, Jason Kleinman, and Chris Malarkey. 2007. What Frustrates Screen Reader Users on the Web: A Study of 100 Blind Users. *International Journal of Human-Computer Interaction* 22, 3, 247–269.
16. Jonathan Lazar, Alfreda Dudley-Sponaugle, and Kisha Dawn Greenidge. 2004. Improving web accessibility: A study of webmaster perceptions. *Computers in Human Behavior* 20, 2, 269–288.
17. Jonathan Lazar, Brian Wentz, C. Akeley, et al. 2012. Equal access to information? Evaluating the accessibility of public library web sites in the state of Maryland. *Designing inclusive systems: Designing inclusion for real-world applications*, 185–194.
18. Jennifer Mankoff, Holly Fait, and Tu Tran. 2005. Is your web page accessible? In *Proceedings of the SIGCHI conference on Human factors in computing systems* (CHI'05), 41-50.
19. Mary Meeker. Internet Trends 2014 – Code Conference. Retrieved Oct 08, 2015 from http://recode.net/2014/05/30/mary-meekers-annual-rapid-fire-internet-trends-talk-video/
20. 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* (CHI '16). ACM, New York, NY, USA, 5506-5516.
21. Christopher Power, André Freire, Helen Petrie, and David Swallow. 2012. Guidelines are only half of the story: accessibility problems encountered by blind users on the web. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems* (CHI '12), 433–442. http://doi.acm.org/10.1145/2207676.2207736

22. Krishnan Ramnath, Simon Baker, Lucy Vanderwende, Motaz El-Saban, Sudipta Sinha, Anitha Kannan, Noran Hassan, Michel Galley, Yi Yang, Deva Ramanan, Alessandro Bergamo, and Lorrenzo Torresani. AutoCaption: Automatic Caption Generation for Personal Photos. *IEEE Winter Conference on Applications of Computer Vision* , March 2014
23. Ravic Ringlaben, Marty Bray, and Abbot Packard. 2014. Accessibility of American University Special Education Departments' Web Sites. *Universal Access in the Information Society* 13, 2, 249–254.
24. Pierre Sermanet, David Eigen, Xiang Zhang, Michael Mathieu, Rob Fergus, and Yann LeCun. OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks. *International Conference on Learning Representations (ICLR2014),* April 2014.
25. Ben Shneiderman. 2000. Universal usability. *Communications of the ACM* 43, 5, 84–91.
26. Kevin Tang, Manohar Paluri, Li Fei-Fei, Rob Fergus, and Lubomir Bourdev. Improving Image Classification with Location Context. *International Conference on Computer Vision (ICCV),* 2015.
27. TapTapSee, http://www.taptapseeapp.com/
28. Shari Trewin, Diogo Marques, and Tiago Guerreiro. 2015. Usage of Subjective Scales in Accessibility Research. In *Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility* (ASSETS '15). ACM, New York, NY, USA, 59-67. DOI=http://dx.doi.org/10.1145/2700648.2809867
29. Using Artificial Intelligence to Help Blind People 'See' Facebook, Retrieved May 25, 2016 from https://newsroom.fb.com/news/2016/04/using-artificial-intelligence-to-help-blind-people-see-facebook/
30. Violeta Voykinska, Shiri Azenkot, Shaomei Wu, and Gilly Leshed, How Blind People Interact with Visual Content on Social Networking Services. In *Proceedings of the Computer-Supported Cooperative Work and Social Computing (CSCW '16), to appear, Feb 27, 2016, San Francisco, USA.*
31. Web Accessibility Initiative. 2012. WCAG Overview. Retrieved May 25, 2016 from http://www.w3.org/WAI/intro/wcag.php
32. Shaomei Wu and Lada Adamic, *Visually Impaired Users on an Online Social Network*. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '14), 3133-3142.
33. Ning Zhang, Manohar Paluri, Marc'Aurelio Ranzato, Trevor Darrell, Lubomir Bourdev. PANDA: Pose Aligned Networks for Deep Attribute Modeling. In *Proceedings of the Ieee Computer Society Conference On Computer Vision and Pattern Recognition* (CVPR '14*), 1637-1644*.
34. 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* (CHI '15), 2353–2362.