Image Explorer: Multi-Layered Touch Exploration to Make Images Accessible

Jaewook Lee
University of Illinois at Urbana-Champaign
Champaign, IL, USA
jaewook4@illinois.edu

Yi-Hao Peng
Carnegie Mellon University
Pittsburgh, PA, USA
yihaop@cs.cmu.edu

Jaylin Herskovitz
University of Michigan
Ann Arbor, MI, USA
jayhersk@umich.edu

Anhong Guo
University of Michigan
Ann Arbor, MI, USA
anhong@umich.edu

ABSTRACT

Blind or visually impaired (BVI) individuals often rely on alternative text (alt-text) in order to understand an image; however, alt-text is often missing or incomplete. Automatically-generated captions are a more scalable alternative, but they are also often missing crucial details, and, sometimes, are completely incorrect, which may still be falsely trusted by BVI users. We hypothesize that additional information could help BVI users better judge the correctness of an auto-generated caption. To achieve this, we present *Image Explorer*, a touch-based multi-layered image exploration system that enables users to explore the spatial layout and information hierarchies in an image. *Image Explorer* leverages several off-the-shelf deep learning models to generate segmentation and labeling results for an image, combines and filters the generated information, and presents the resulted information in hierarchical layers. In a pilot study with three BVI users, participants used *Image Explorer*, Seeing AI, and Facebook to explore images with auto-generated captions of diverging quality, and judge the correctness of the captions. Preliminary results show that participants made more accurate judgements about the correctness of the captions when using *Image Explorer*, although they were highly confident about their judgement regardless of the tool used. Overall, *Image Explorer* is a novel touch exploration system that makes images more accessible for BVI users by potentially encouraging skepticism and enabling users to independently validate auto-generated captions.

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; **Accessibility technologies**.

ACM Reference Format:
Jaewook Lee, Yi-Hao Peng, Jaylin Herskovitz, and Anhong Guo. 2021. Image Explorer: Multi-Layered Touch Exploration to Make Images Accessible. In *The 23rd International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '21), October 18–22, 2021, Virtual Event, USA*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3441852.3476548

ASSETS '21, October 18–22, 2021, Virtual Event, USA

ACM ISBN 978-1-4503-8306-6/21/10.
https://doi.org/10.1145/3441852.3476548

1 INTRODUCTION

Understanding images on the web can be a challenging task for blind or visually impaired (BVI) individuals. These users often depend on alternative text (also known as alt-text) [3] in order to understand the content of an image. However, in part due to the rapid increase in the quantity of user-uploaded content online, a growing number of images are missing alt-text, leaving a large fraction of images inaccessible [13]. While some platforms have provided users with the option to add alt-text as they upload a photo, these options are rarely used [5]. Additionally, even when some users provide their own image alt-text, they may be of little use as user-written alt-text is often low quality or missing important details [14]. Recent work has instead turned to auto-generating image captions [15, 16] with the goal of providing high quality alt-text at scale. Automated systems have shown to greatly improve the coverage of alt-text [6, 7], but the quality and accuracy of these captions still remain questionable.

Even if some auto-generated captions are incorrect or misleading, without the means to verify correctness, BVI users place a high degree of trust in them, especially if they do not have access to additional information [10]. This prior work has also observed that BVI users often attempt to resolve issues in captions by filling in details and developing their own descriptions that could fit the scenario. One way to address this is to provide richer ways to interact with an image beyond a single caption, so that users can investigate the auto-generated captions for themselves after gaining a better understanding of the image's content and layout. For instance, Morris et al. found that the following two approaches are helpful in improving BVI people's understanding of images: 1) providing alt-text through multiple "layers", where deeper layers contain additional detail, and 2) supporting touch-based interaction with images [12]. In this paper, we aim to combine these two approaches in a single system, and compare it against state-of-the-art image exploration systems to understand if doing so will improve BVI users' abilities to identify errors in auto-generated captions and give rise to scepticism towards these descriptions.

In order to improve image understanding and to further investigate how BVI users leverage additional information when interpreting auto-generated image captions, we present *Image Explorer*, a touch-based image exploration system that divides image information into multiple layers to allow users to explore the spatial layout and information hierarchies in an image. *Image Explorer* relies on various deep learning models to gather a large quantity of information about an image, and presents the collected information in two layers. First, the essential, high-level objects in the

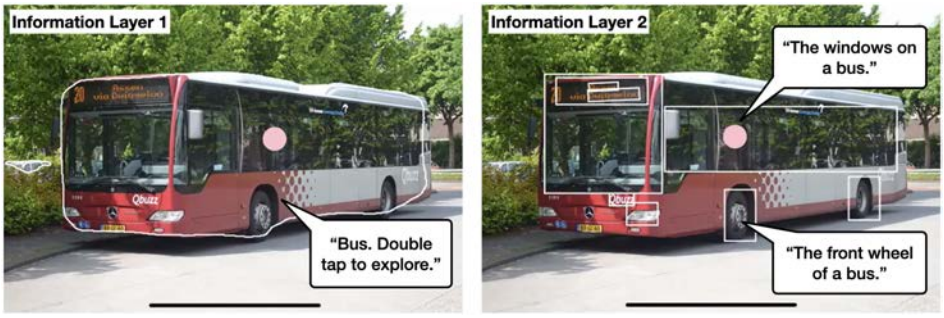

Figure 1: *Image Explorer* user interface. The first information layer shows primary objects in the image outlined with polygonal boundaries. After double tapping on an object, users enter the second information layer, which shows rectangular bounding boxes around various detailed sub-objects.

image are presented as polygonal boundaries, which can be explored by dragging a finger across the screen. Upon coming into contact with an object's boundaries, users can double-tap to enter the second layer of the object, which consists of sub-objects, color, written text on an object, or other details (shown in Figure 1). We gathered early feedback regarding *Image Explorer* in a preliminary pilot study, and use the results to reveal key directions for improvements and future work. Using *Image Explorer*, participants made more accurate judgements about the correctness of auto-generated image captions. Overall, *Image Explorer* is an important step towards encouraging skepticism in AI-generated captions, allowing independent verification, and increasing image understanding.

2 IMAGE EXPLORER SYSTEM

Image Explorer is an image exploration system that uses touch and multiple information layers to allow users to explore the content of the image, their spatial relationships, and their hierarchy. It is intended to supplement auto-generated natural language captions, which often have errors; we instead focus on providing users with a variety of raw information from off-the-shelf models so that they can judge captions independently. *Image Explorer* was designed to allow users to identify common errors in auto-generated captions, including missing information, incorrect object labels, and incorrect layout descriptions (examples are shown in Figure 2). It first collects information from an image through a handful of deep learning algorithms, aiming to reduce the probability of missing information. It then separates this information into two presentation layers using a set of criteria, allowing users to review details about an object and correct mislabeling. Finally, it provides a touch interface supported by audio feedback for accessing object information, allowing users to explore the spatial relationships between objects.

2.1 Element Detection and Scene Hierarchy

Image Explorer leverages a variety of existing deep learning models to detect image content and create a scene hierarchy that can later be explored by users. It focuses on extracting common image elements: the location, boundaries, and descriptive labels of people and objects in an image, and transcriptions of printed text. To accomplish this, we relied on a combination of Mask R-CNN [8], DenseCap [9], and Google Cloud Vision's Labels, Face, Text, and Localized Object solutions [1]. Mask R-CNN was used to generate the first layer of information in the image. Compared to other object recognition models, Mask R-CNN is unique in that it generates segmentation masks, which are polygon-shaped borders that best fit elements of interest [8]. Because *Image Explorer* uses object boundaries to determine the scene hierarchy, tighter object borders resulted in a more accurate representation than traditional rectangular bounding boxes.

To create the second layer of information, *Image Explorer* relies on DenseCap and Google Cloud Vision API to supplement the information provided by Mask R-CNN. We leverage DenseCap to generate more specific object descriptions (e.g., "front wheel" and "back wheel" of a vehicle, identifying color of objects), setting the confidence threshold to 75% to filter inaccurate labels. This threshold was chosen empirically based on our observations of performance, such that it removed many misleading labels while still maintaining a sufficient amount of labels overall. Additionally, we use Google Cloud Vision to perform more general object (e.g., face) and text detection and labeling.

Finally, to organize the first- and second-layer information into a hierarchical structure, *Image Explorer* analyzes each second-layer element (i.e., rectangular bounding box with text layer) to determine which first-layer element, if any, it belongs to. We used the following criteria: (1) the area of the second-layer element must be smaller than that of the first-layer element, and (2) at least 75% of the second layer element must overlap with the first layer element. Second-layer elements were then paired with the first-layer element that they overlapped with most. Additionally, any areas in both layers less than 34 pixels by 34 pixels were omitted to remove elements that were too small to touch.

2.2 *Image Explorer* User Interface

Image Explorer provides a touch interface for exploring the content and hierarchy of an image that we extracted. An overview is shown in Figure 1. When the user opens an image, *Image Explorer* first vocalizes the number of elements available, and displays the first-layer elements as polygonal boundaries overlaid onto the image for users to explore. As users move their finger across an image, they receive audio feedback: when not touching any element, a background tone plays; when touching an element, its name is read verbally (e.g., "bus", "car", "person"). If an element contains sub-elements, users are then verbally prompted to double tap for more information (e.g., "bus, double tap to explore"). If the user chooses to double tap on a first layer element, the system will display its

Figure 2: Example images from *Image Explorer*'s pilot study highlighting common errors found in auto-generated image captions.

corresponding second layer elements, which the users can again explore using touch. Users can exit the second layer at any time by double tapping again anywhere on the screen. When the user returns to the first layer, the system will again repeat the number of elements that have yet to be explored, providing users an awareness of their exploration progress.

3 PILOT STUDY

In order to gather early feedback about *Image Explorer* and to understand how BVI users judge caption accuracy, we ran a pilot study with three blind iPhone users. We asked participants to use three systems (*Image Explorer*, Facebook's text-based 'Detailed Image Descriptions' feature [4], and Seeing AI's touch exploration feature [11]) to obtain more information about an image and judge if there was an error in a natural language caption generated by Microsoft Cognitive Services [2]. For the purpose of the study, we generated six pairs of images, each with one good caption (e.g. "a police officer on a motorcycle") and one bad caption (e.g. "probably calendar," but the image is a bag of cheese). The systems and image pairs were counter-balanced, and each participant used all three systems for at least one pair of images each. Participants were then asked to rate (1) the accuracy of the caption and (2) their confidence in their rating both before and after exploring the image with one of the three system conditions. We also asked participants to rate each system for its helpfulness and ease of use, and to rank the three systems overall.

We observed that *Image Explorer* allowed users to make more accurate judgements about the correctness of the auto-generated captions than with the two other systems, both identifying errors in captions and verifying correct captions more frequently. This supports the prior work by Morris et al. [12]. All three participants agreed that touch was useful in evaluating positions of objects relative to each other. For instance, P2 explored an image of a cat sitting on a table using *Image Explorer* (see Figure 2). P2 was able to determine that the caption, "a cat sitting on a chair," was incorrect because the cat was positioned above the table, and the chairs were located to the right of the cat. Additionally, all three participants expressed satisfaction towards the information layers in *Image Explorer* as they provided much more information than the other two systems and gave the participants an option to choose whether to view the additional information or not. However, all three participants found Facebook's text-based system to be the easiest to use, and ultimately, they preferred to use Facebook over the two touch-based systems. When using Seeing AI or *Image Explorer*, participants often struggled to locate all of the elements dispersed across the image. While touch shows great promise in providing BVI users with more information about an image, which helps them making more informed judgements about AI-generated captions, better interaction and audio feedback design could increase its usability, an important direction for future work.

4 CONCLUSION

In this work, we have presented *Image Explorer*, a system that supports BVI users to gain multi-layer image information through touch. Our pilot results indicate the usefulness of touch and layers in increasing image understanding, and demonstrate that this additional information could be useful in enabling BVI users to assess errors in captions. In the future, we hope to continue to improve the design of *Image Explorer*, and perform a larger study of caption evaluation by BVI users. Future work could also investigate how to better convey additional types of information, such as the shape and size of objects. Overall, *Image Explorer* is a step towards understanding the role that additional information plays in image interpretation, and can be used to improve the design of future image understanding systems.

REFERENCES

[1] Google Cloud. 2021. Cloud Vision API. https://cloud.google.com/vision.
[2] Microsoft Azure Cloud. 2021. Azure Computer Vision API. https://azure.microsoft.com/en-us/services/cognitive-services/computer-vision/.
[3] W3 Consortium. 2018. Web Content Accessibility Guidelines (WCAG) 2.1. https://www.w3.org/TR/WCAG21/
[4] Facebook. 2021. How Facebook is using AI to improve photo descriptions for people who are blind or visually impaired. https://ai.facebook.com/blog/how-facebook-is-using-ai-to-improve-photo-descriptions-for-people-who-are-blind-or-visually-impaired/
[5] Cole Gleason, Patrick Carrington, Cameron Cassidy, Meredith Ringel Morris, Kris M. Kitani, and Jeffrey P. Bigham. 2019. "It's Almost like They're Trying to Hide It": How User-Provided Image Descriptions Have Failed to Make Twitter Accessible. In *The World Wide Web Conference* (San Francisco, CA, USA) *(WWW '19)*. Association for Computing Machinery, New York, NY, USA, 549–559. https://doi.org/10.1145/3308558.3313605
[6] Cole Gleason, Amy Pavel, Emma McCamey, Christina Low, Patrick Carrington, Kris M Kitani, and Jeffrey P Bigham. 2020. Twitter A11y: A browser extension to make Twitter images accessible. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–12.
[7] 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* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3173574.3174092

[8] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*. IEEE, New York, NY, USA, 2961–2969.

[9] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. 2016. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. IEEE, New York, NY, USA, 4565–4574.

[10] 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* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 5988–5999. https://doi.org/10.1145/3025453.3025814

[11] Microsoft. 2021. SeeingAI. https://www.microsoft.com/en-us/ai/seeing-ai

[12] 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* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3173574.3173633

[13] 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* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 5506–5516. https://doi.org/10.1145/2858036.2858116

[14] Abigale Stangl, Meredith Ringel Morris, and Danna Gurari. 2020. "Person, Shoes, Tree. Is the Person Naked?" What People with Vision Impairments Want in Image Descriptions. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–13.

[15] 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*. IEEE, New York, NY, USA, 49–56.

[16] 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* (Portland, Oregon, USA) *(CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 1180–1192. https://doi.org/10.1145/2998181.2998364