June 18, 2019

Utilizing Gaze Detection to Enhance Voice-Based Accessibility Services

Matthew Crain

Pingmei Xu

Alex Salo

Tomer Shekel

Utilizing Gaze Detection to Enhance Voice-Based Accessibility Services

Abstract:

For users that use voice as their primary mode of input, operating a computing device can be difficult due to potential false positives (*e.g.*, unintended voice commands by the user, background noise such as a radio). Voice commands can also be difficult to decipher, resulting in the voice-based accessibility service needing additional, clarifying user input to disambiguate the auditory commands.

This publication describes techniques and procedures for utilizing gaze detection to enhance voice-based accessibility services on a computing device, such as a smartphone or computer. The computing device utilizes camera image input and a machine-learned model to produce an estimated x-y coordinate of where the user is gazing on a display of the computing device. Utilizing the machine-learned model, if the computing device determines that the user is looking at the computing device's display (*i.e.*, giving the device attention), then auditory commands are accepted; otherwise, if the user is not giving the device attention, then auditory commands can be ignored. Additionally, the techniques and procedures can assist in disambiguation (*e.g.*, similar sounding commands, identically titled functions). Finally, the techniques and procedures can be used as an alternative means for controlling the scrolling of the display of the device.

Keywords:

Gaze tracking, gaze detection, eye tracking, eye movement, voice-based accessibility services, voice input, voice command, machine-learning, computing device, smartphone, gaze location, x-y coordinate, disambiguation of auditory commands, assistive technology, disability, gaze scrolling

Background:

For users who use voice commands as a means of input, it can be difficult to operate their computing device in environments with external noise sources (*e.g.*, a TV, a radio). Many false positives (*e.g.*, unintended voice commands by the user, background noise sources accidentally commanding the device) can be picked up by the computing device and perform undesired operations. Users frequently need to repeat auditory commands to assist the computing device in determining desired operations. These complications can make utilizing voice-based accessibility services quite frustrating for the user.

Therefore, a comprehensive solution that utilizes analyzing gaze tracking in cooperation with voice-based accessibility services on a computing device is desirable. With such a solution, users do not need additional hardware or services beyond the primary voice-based accessibility service they already use. This solution is inexpensive, simple to configure, and reduces the possibility of failure with future updates.

Description:

This publication describes techniques and procedures for utilizing gaze detection to enhance voice-based accessibility services on a computing device, such as a smartphone or

computer. These techniques enable the computing device to perform multiple operations that manage the accessibility of the device. Such operations include capturing an image of the user, determining an estimated x-y coordinate of user gaze location, and determining user attention. These operations work in cooperation with voice-based accessibility services to assist the device distinguish between actual user voice commands and false positives (*e.g.*, unintended voice commands by the user, background noise sources accidently commanding the device).

Figure 1 illustrates an example computing device and elements of the computing device that support utilizing gaze detection to enhance voice-based accessibility services.

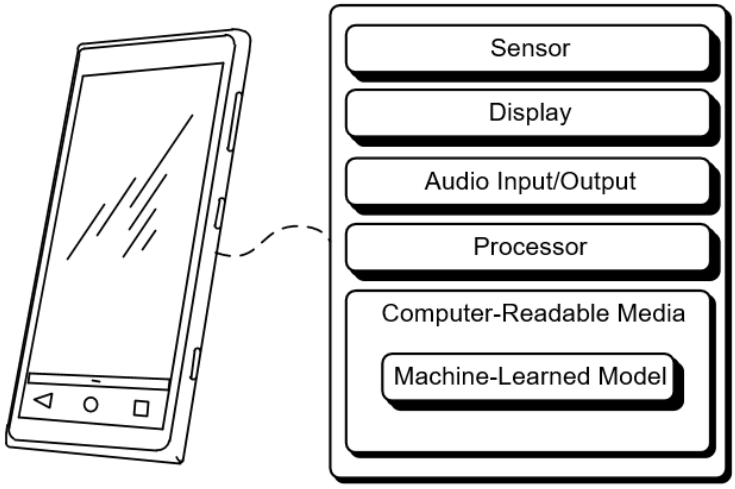

Figure 1

As illustrated in Figure 1, the computing device is a smartphone. However, other computing devices (*e.g.*, a tablet, a computer, a wireless-communication device) can also support the techniques and procedures described in this publication. The computing device includes a sensor (*e.g.*, a camera), a display (*e.g.*, a light emitting diode (LED) display, a liquid crystal display (LCD)), and an audio input/output mechanism (*e.g.*, a microphone, a speaker). The computing device also includes a processor and a computer-readable medium (CRM) that contains executable instructions for implementing the techniques and procedures described in this publication. The

CRM may include the operating system (OS) of the computing device and applications installed on the computing device. The CRM may include any suitable memory or storage device such as random-access memory (RAM).

A machine-learned (ML) model (*e.g.*, neural network) can be trained on labeled images of eyes and the respective gaze direction. After sufficient training, the machine-learned model can be deployed to the CRM. Figure 2, below, illustrates an example of how the OS of the computing device and/or applications installed on the computing device may use the machine-learned model to produce an estimated x-y coordinate of user gaze.

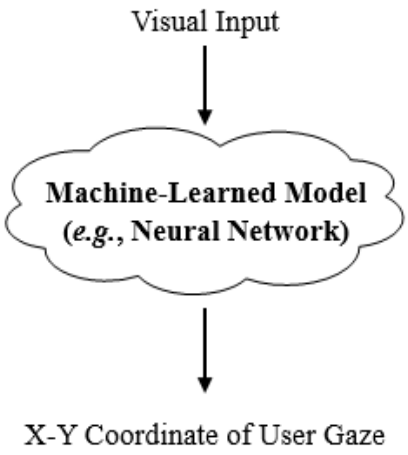

Figure 2

In Figure 2, the sensor (*e.g.*, camera) of the computing device measures visual input (*e.g.*, the size of the pupil of the eye, the location of the pupil of the eye, the pitch of the head, the tilt of the head, the yaw of the head, the motion of any eye relative to the head) relating to the user. The visual input may include images of the user's eyes taken by a camera. This visual input is then inputted to the machine-learned model to produce an estimated x-y coordinate on the device of where the user is looking. With this estimated x-y coordinate, the computing device (*e.g.*, an application on the CRM) can determine user attention (*i.e.*, whether the user is looking at the

device, whether the user is not looking at the device), gaze location of the user with respect to the device's screen coordinates with vertical and horizontal buffer, and relative location of the user (*i.e.*, the relative point from which the user is looking).

If the computing device determines that the user's attention is currently given to the display of the computing device, then voice-based commands measured by the computing device's audio input can be used to control the device. Otherwise, if the computing device determines that the user's attention is not focused on the display of the computing device, attention, then voice-based commands can be ignored to avoid accidental input from the user or other external sources.

The beneficial cooperation of the machine-learned model with voice-based accessibility services can be appreciated in the following example. While engaged in conversation with a friend, a quadriplegic computing device user desires to send a digital photograph from a recent trip to a friend. The user activates the voice input service on his computing device to navigate to the photograph. While the user is speaking commands to the computing device, his friend asks him a question. Naturally, the user removes his attention from the display of his computing device to look at his friend and answer the question. Since attention is no longer given to the display of the computing device, the computing device can control the voice-based accessibility services to ignore any potential voice commands received from the user or the friend. The user, upon finishing his response to the friend, can look back at the display of the computing device and continue providing voice commands to the device. An illustration of the computing device determining the user's attention and receiving his commands is illustrated in Figure 3.

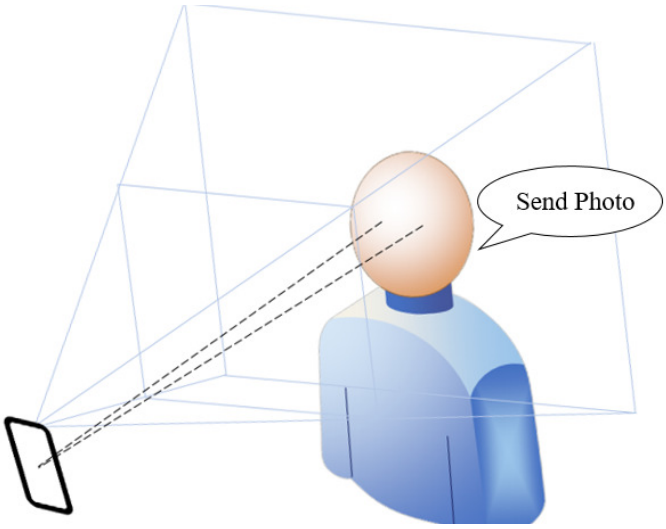

Figure 3

As illustrated in Figure 3, the computing device is a smartphone. The smartphone's camera measures visual input relating to the user (*e.g.*, the point of gaze of the user, the motion of any eye of the user relative to the user's head). The computing device utilizes the machine-learned model to determine when the user is giving the display of the device attention and receives the auditory command. Once the user has finished the task, he can then pause voice-based accessibility services by looking away or he can disable the gaze tracking feature entirely.

The techniques and procedures described in this publication can also be utilized to decrease command disambiguation. If a user desires to interact with two similarly named items on a computing device, the user's gaze can be used to disambiguate between the two items without further user interaction. For instance, if voice-based accessibility services are processing and the user says "Settings" to click on an icon that looks like a gear in the top right of the screen, but another settings option is available on the bottom left of the screen, then the computing device can utilize gaze tracking analysis to distinguish between the two settings options based on the user's x-y gaze coordinate position on the display as predicted by the ML model. This way the computing device does not have to inquire further from the user as to which settings option is desired.

Another additional benefit of utilizing gaze detection to enhance voice-based accessibility services is device display scrolling. If the user, for instance, is exploring an online map by means of voice-based accessibility services, instead of commanding "Scroll up," "Scroll right," *etc*., the user can utilize gaze scrolling. As the user looks around the map conventional scrolling services can be imitated, such that the user can pan in the direction of their gaze to access more content. To accomplish gaze-based scrolling, the computing device by means of gaze detection marks the focal center of the user indicated view. The computing device then uses the focal center as the relative starting point and the gaze history to determine the intended direction of scrolling. A demonstration of the scrolling coordinates is illustrated in Figure 4 below.

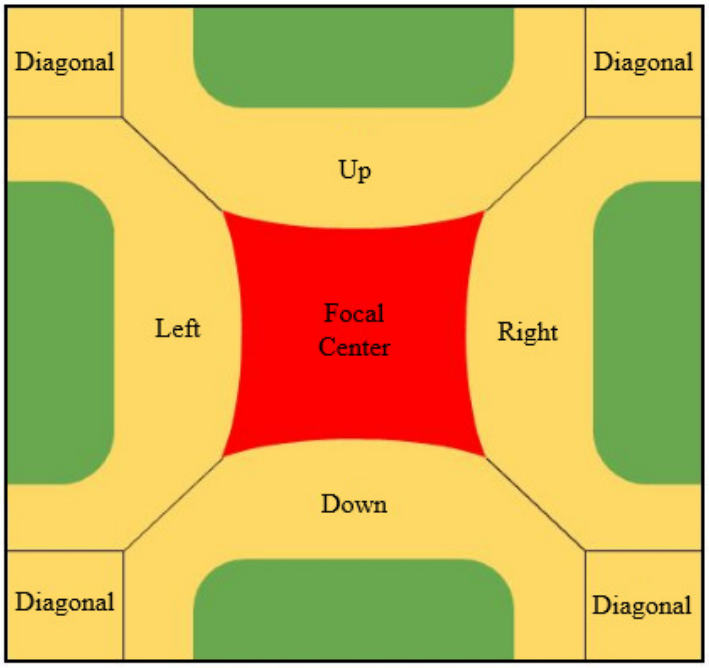

Figure 4

As illustrated in Figure 4, gaze detection on the computing device allows the user to scroll in different directions based on where the user's gaze focuses. The scrolling speed can accelerate depending on the magnitude of the distance from the current gaze position and the focal center. For example, in the red region (*e.g.*, the focal center), no scrolling is performed; in the yellow

region (*e.g.*, left, right, up, down, diagonal), scrolling is performed at a base speed in the direction the region indicates; and in the green region (*e.g.*, the extreme positions), scrolling is performed at an accelerated speed.

Finally, the accessibility service allows the user to enable or disable each feature (*e.g.*, attention tracking, gaze scrolling) and alters the frame rate that the images are processed to efficiently utilize the device's battery.

The images captured by the computing device can be immediately discarded after usage/processing. All computations necessary to detect and determine user gaze direction are performed on the device by the ML model. Further to the descriptions above, a user may be provided with controls allowing the user to make an election as to both if and when systems, programs or features described herein may enable collection of user information (*e.g.*, information about a user's social network, social actions or activities, profession, a user's preferences), and if the user is sent content or communications from a server. In addition, certain data may be treated in one or more ways before it is stored or used, so that personally identifiable information is removed. For example, a user's identity may be treated so that no personally identifiable information can be determined for the user, or a user's geographic location may be generalized where location information is obtained (such as to a city, ZIP code, or state level), so that a particular location of a user cannot be determined. Thus, the user may have control over what information is collected about the user, how that information is used, and what information is provided to the user.

The incorporation of gaze detection enhances voice-based accessibility services on computing devices. Other solutions rely on additional hardware that can be expensive and difficult

to configure. The addition of gaze detection to voice-based accessibility services can provide users convenient and affordable services.

Reference:

[1] Hennessey, Craig A., Jacob Fiset, and Simon St-Hilaire. System and Method for Using EyeGaze Information to Enhance Interactions. US Pub. 20140184550, filed March 7, 2014, published July 3, 2014.