October 2020

Contextual Action Recommendations For A Virtual Assistant Based On Gaze Information

Anonymous

Contextual Action Recommendations For A Virtual Assistant Based On Gaze Information

ABSTRACT

Wearable devices such as augmented reality/mixed reality glasses can detect a user's gaze and objects that are within a user's field of view. The devices also include virtual assistant functionality that enables a user to provide commands to a virtual assistant, e.g., via voice, gestures, or other input. This disclosure leverages both these aspects to provide recommended actions to a user based on objects that the user gazes at. Upon detection that the user is gazing at a particular object, recommended actions for the virtual assistant are identified and ranked based on the object and the user's context. The ranked actions are provided to the user via a suitable modality, e.g., displayed via their glasses, provided as audio, etc. based on the user's context, along with indication of input the user can provide to trigger the virtual assistant to perform the action(s). The input modality can be selected based on the user's context. The recommended actions can be across various domains that the virtual assistant supports, e.g., communications, shopping, scheduling, providing information, etc.

KEYWORDS

- Gaze detection
- Object recognition
- Wearable device
- Head-mounted device (HMD)
- Augmented reality
- Mixed reality
- Virtual assistant
- Contextual recommendation
- Input modality

BACKGROUND

Wearable devices such as augmented reality/mixed reality headsets are popular among users. Such devices are equipped with hardware, e.g., cameras and other sensors, that can be

used to detect the direction in which a user that wears such a headset is looking. Some devices are also configured with software, e.g., a virtual assistant, that can perform multiple tasks or actions as commanded by the user. The types of tasks or actions that the virtual assistant can perform can span a large number of domains.

DESCRIPTION

This disclosure describes techniques to determine that the user has gazed at particular objects, where the gaze time meets a threshold for a likelihood that the user has an intent with respect to objects in their field of view. Based upon determining that the user has gazed at the particular objects, personalized and contextually appropriate recommended actions that a virtual assistant can perform are provided to the user as suggestions. For example, the techniques may be utilized in a head-mounted device (HMD). Context can include the user's location (e.g., workplace, residence, social setting, etc.); date/time; user's past activity; and other factors. The type and/or identity of the objects in the field of view is utilized to rank possible actions and the top-ranking actions are provided as a suggestion. Further, the context is also a factor that determines the activation modality for the recommended actions, e.g., gesture input, touch input, voice input, etc.

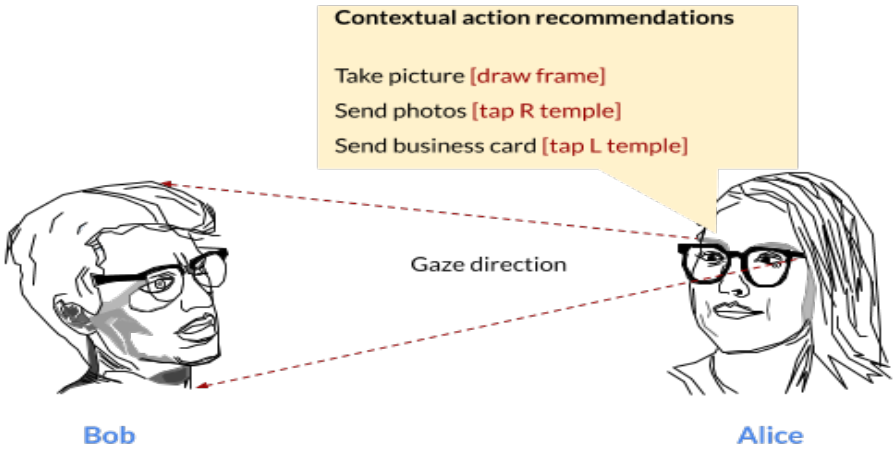

Fig. 1: Gaze-based action recommendations

Fig. 1 illustrates an example of recommending actions that can be performed, e.g., by a virtual assistant provided via a head-mounted device (e.g., AR/VR/MR glasses) based on detecting an object in the user's gaze. As seen in Fig. 1, a user Alice wearing an HMD is looking towards another person Bob. The gaze direction can be detected by on-board hardware of the HMD. It is determined whether the user Alice has looked in this direction for at least a threshold time, and that during this time, the object (the person Bob) has been present in the user's field of view.

Based on detecting that Alice has been looking at Bob, contextual action recommendations are provided. The recommendations are ranked based on relevance. For example, as seen in Fig. 1, the recommendations can be to perform an action to take a picture of Bob, to send photos to Bob, or to send a business card to Bob. The library of recommended actions can automatically be expanded as the virtual assistant gains additional functionality. Corresponding gestures are also indicated so that the user Alice can provide the instruction to the virtual assistant to complete the action. In the example of Fig. 1, these are:

- ***Draw frame*** which corresponds to Alice drawing a frame in the air with her hands and acts as an input for the glasses to capture a picture of Bob;
- ***Tap R temple*** which corresponds to Alice tapping the right temple of the glasses to send photos to Bob; and
- ***Tap L temple*** which corresponds to Alice tapping the left temple of the glasses to send her business card to Bob, e.g., via a messaging application that Bob uses.

Action recommendations based on object and context

The recommended actions can be based on recognizing the object in the field of view and on the user's context. Object recognition can include semantic analysis, e.g., to determine

whether the user is looking at a person (e.g., known or unknown person), landmark (e.g., Eiffel tower), object (e.g., business card, foreign language signage, store hours, menus, etc.), and so on. Further, the recommended actions can be personalized to the user based on the user's prior actions and other user-permitted data. Only such actions that meet a threshold relevance are provided. Virtual assistants are capable of a very wide range of actions across a large number of domains, and selection of the recommended actions in this manner can ensure that the user is provided with suggestions of the most useful actions.

This is illustrated with some examples below:

- **User gazes at a person:** suggested actions can be to share media in which the person appears (e.g., photos, videos); schedule a meeting with the person (if the person is a co-worker or otherwise known in a work context); add as social media friend (if such a relationship doesn't currently exist); send business card (if new person); review related information (e.g., recent communications with the person); etc.
- **User gazes at an object:** suggested actions can include queries related to the object, e.g., nutritional information, price, etc. for a food item; prices at other stores when gazing at an item for sale at a store; dictionary lookup or translation when gazing at a printed publication; etc.
- **User gazes at a landmark:** suggested actions can include taking pictures of the landmark, viewing information regarding the landmark, viewing operational hours, etc.

Personalized action recommendations

The recommended actions are personalized to the user. For example, if the user is a resident of Paris and passes by the Eiffel tower, no recommendation to take a picture is provided

(e.g., since the user is likely familiar with the landmark and may already have pictures in their library) while for another user that is detected as being on a holiday in Paris, a recommendation is provided to take a picture.

In another example, if the user often gazes at printed materials in a foreign language and selects a "translate" action, such action can be shown with a high priority each time foreign language text is detected in the user's gaze. Personalization can be based on various contextual factors such as user's location (home or traveling), time/day (different recommendations for work hours on a work day and for other times), etc.

Different actions may be suggested for similar objects based on the user's context - e.g., if the user is in a business conference and gazes at an unknown person, actions to share a business card or to set up a meeting may be recommended, while if the same user is at a party, actions to add the unknown person as a social media friend may be recommended.

Gesture and modality selection based on user's context

The recommended actions can be associated with specific gestures. The gestures can be selected based on the user's preference, e.g., tap on right wrist for left-handed users, tap on watch shown based on the hand on which the watch is worn, etc. The user's context is taken into account when providing recommendations. For example, if the user is in a location or context where certain gestures are inappropriate (e.g., due to cultural context, if the user is driving or engaged in other activity, etc.), the corresponding recommended actions may be suppressed (or may be shown with a different suggested modality). For example, for a user that is driving, gestures that require the use of hands may be suppressed, and instead, the user may be encouraged to choose from the recommended actions via spoken commands.

Action recommendations are provided in a suitable modality

The recommended actions can be provided via a suitable modality, determined based on the user's context. For example, if the user is engaged in an activity that requires visual attention (e.g., driving, cooking, etc.), the recommendation may be provided via audio (e.g., turn L at the next junction, the foreign language road sign says "traffic ahead, slow down," etc.).

CONCLUSION

This disclosure describes techniques to provide recommended actions to a user based on objects that the user gazes at, detected by a device worn by the user. Upon detection that the user is gazing at a particular object, recommended actions for a virtual assistant are identified and ranked based on the object and the user's context. The ranked actions are provided to the user via a suitable modality, e.g., displayed via their glasses, provided as audio, etc. based on the user's context, along with indication of the input the user can provide to trigger the virtual assistant to perform the action. The input modality can be selected based on the user's context. The recommended actions can be across various domains that the virtual assistant supports, e.g., communications, shopping, scheduling, providing information, etc.