November 2021

Combined Head Gestures for Improved User Interaction

Lei SHI

Aobo ZHOU

Combined Head Gestures for Improved User Interaction

ABSTRACT

Fine-grained gestures such as eye gaze, facial expressions, etc., can be useful as input mechanisms for smart devices. However, single gesture inputs such as eye gaze are inefficient, and it is difficult to perform complex operations with such inputs. This disclosure describes techniques to fuse multiple head gestures, e.g., head, eye, mouth, or eyebrow movement, to provide superior user interaction. Head gestures are classified as analog (e.g., eye movements, which provide continuous input) or binary (e.g., frowns, which indicate one-time operation). An analog gesture is fused with multiple binary gestures to improve efficiency. For example, to move an object, the user selects the object by looking and frowning at it; moves the object using eye gaze; and frowns again to set it in position. Different facial expressions can be flexibly assigned to accommodate varied user abilities and preferences.

KEYWORDS

- Accessibility
- Gesture input
- Facial gesture
- Gaze detection
- Expression recognition
- Fine-grained gesture recognition
- Gesture fusion
- Multimodal user interaction
- Facial landmarks
- Digital maps
- Neural networks
- Head-pose detection

BACKGROUND

Fine-grained gestures detected using computer vision and sensing techniques can be useful as input mechanisms, especially for accessibility applications. For example, a webcam feed can be used to accurately sense a user's facial expression, eye gaze, head position, etc. Eye gaze can be used by people with motor impairment as an alternative input method for mouse control and/or text entry.

However, single gesture input mechanisms such as eye gaze aren’t as efficient as common user interface controllers such as computer mice. For example, on a mouse, users without disabilities can use a combination of actions such as hold-and-drag to move an object. However, it is difficult to synthesize the hold-and-drag effect using a single eye gesture. To move an object using eye-gaze control, the user dwells on the object (mimicking the right-click event of a mouse), selects move, moves the object to a target position, and dwells on the final position again to complete the task.

The limitations of a single gesture input mechanism can lead to user inconvenience in products with rich interaction and functionalities, such as map applications. Key for people with and without disabilities, map applications have rich functionalities that make their interactions complex: to find a place or plan a trip, a user enters search information, zooms in/out, pans the map, selects a point of interest to check detailed information, and browses street views. These interactions require non-trivial effort even for mouse or touchscreen users without disabilities.

The multiple user interaction modes, e.g., speech, gesture, touchscreen, keyboard, etc., typically used in computing devices become more powerful when users combine the modalities [1], e.g., combining speech commands with keyboard-presses. Yet there is more gain to be had by combining combinations *within* a modality rather than (or in addition to) combinations *across* modalities.

Maps being a popular and useful smartphone application, considerable design, research, and engineering effort has been expended to make them accessible. For example, it is possible to interact with map applications using speech input. However, speech input itself can have accessibility problems, for example, for users with speech disabilities.

DESCRIPTION

This disclosure describes techniques to fuse (or combine) multiple head gesture input mechanisms, e.g., eye movement, head movement, mouth movement, eyebrow movement, etc., to provide a superior interaction experience. Head gestures commonly detectable by smartphone cameras or computer webcams, are classified into two categories: analog gestures and binary gestures. Analog gestures are those that provide continuous input signals (e.g., eye movement, head movement), while binary gestures are ones that indicate one-time operation (e.g., a frown).

An analog gesture can be fused with multiple binary gestures to improve efficiency. For example, to move an object using the described fused gestures, the user can

- select the object by looking at the object and making a frowning gesture;
- move the object to a target position using eye gaze; and
- make the frowning gesture again to complete the task.

Similar to the action of moving a mouse and clicking a button, analog gestures and binary gestures can occur simultaneously and independently of each other. Moreover, different facial expressions can be flexibly assigned as binary gestures to accommodate different user disabilities, user preferences, and/or use cases. It is also possible to fuse multiple analog gestures to provide multi-dimensional navigation.

Detection of analog gestures

Eye movements can be tracked accurately in real-time with RGB cameras in smartphones, personal computers, and other devices using various computer vision and neural network techniques, for example, as described in [2]. Head movement can be tracked with various computer vision techniques, e.g., as described in [3]. An RGB image dataset of human head images can be used to train a neural network to recognize head poses. Head movement can be detected as a time sequence of detected head poses.

Detection of binary gestures

Facial expressions, e.g., smiling, frowning, etc., can be detected through mouth and eyebrow movements, in turn recognized by the movement of facial landmarks, as described, e.g., in [4], which describes techniques to detect eleven facial landmarks. Binary classifiers to detect various facial gestures like smiling, frowning, mouth-opening, mouth-closing, etc. can be built based on facial landmarks.

Example application: Digital Maps

While the described techniques of combining gestures apply generally to mobile device (or other smart device or computer) applications, they are described herein using a digital map application as an example. User interactions with digital maps can be sophisticated. A user can use a map application to find a restaurant, to plan driving routes, to preview the street views of destinations, etc. Such map application tasks require analog and binary interactions, and many of these tasks conventionally require multi-dimensional navigation, such as pinching to zoom the map, dragging to move the map, etc. Tables 1-3 illustrate some use cases where combined head gestures described herein can be used to interact with a map in various ways, e.g., to pan the map, to zoom in and out, to explore images of streets, etc.

Gesture 1→ **Gesture 2↓**	**Turn head left**	**Turn head right**	**Turn head up**	**Turn head down**	**Head centered (relative)**
State 0: Mouth closed	Move pointer left	Move pointer right	Move pointer up	Move pointer down	NA
State 1: Mouth open	Pan map left	Pan map right	Pan map right	Pan map down	NA

Table 1: Combined head gestures to enable (two-dimensional) panning of a map

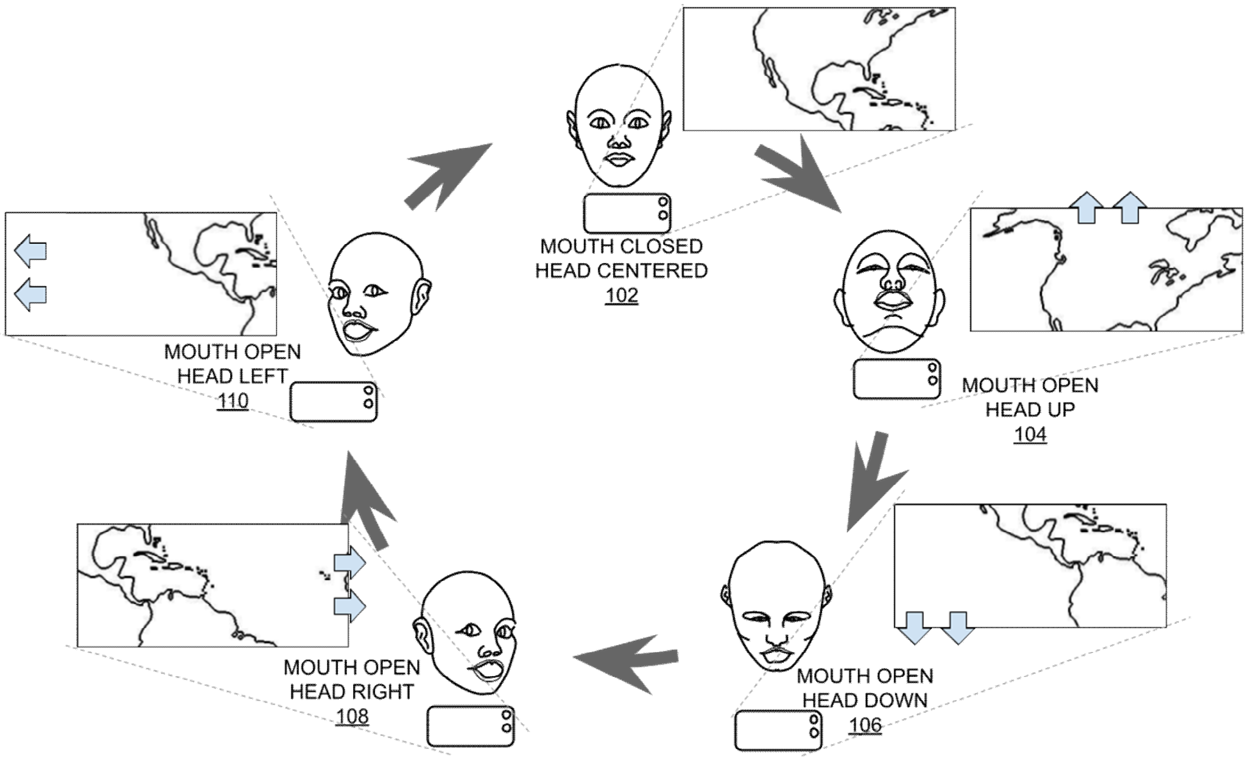

Fig. 1: Combining mouth and head gestures to effect panning

Fig. 1 illustrates combining mouth and head gestures to effect panning a digital map, e.g., as in Table 1. The user starts with mouth closed and head centered (102), which results in a stationary map. The user opens their mouth and moves their head up (104), resulting in northward panning of the map. The user keeps their mouth open and moves their head down (106), resulting in a southward panning of the map. The user keeps their mouth open and turns their head right (108), resulting in an eastward panning of the map. The user keeps their mouth open and turns their head left (110), resulting in a westward panning of the map. The user closes their mouth and re-centers their head to stop panning (102). The speed of panning can depend on the angle of the turn of the head.

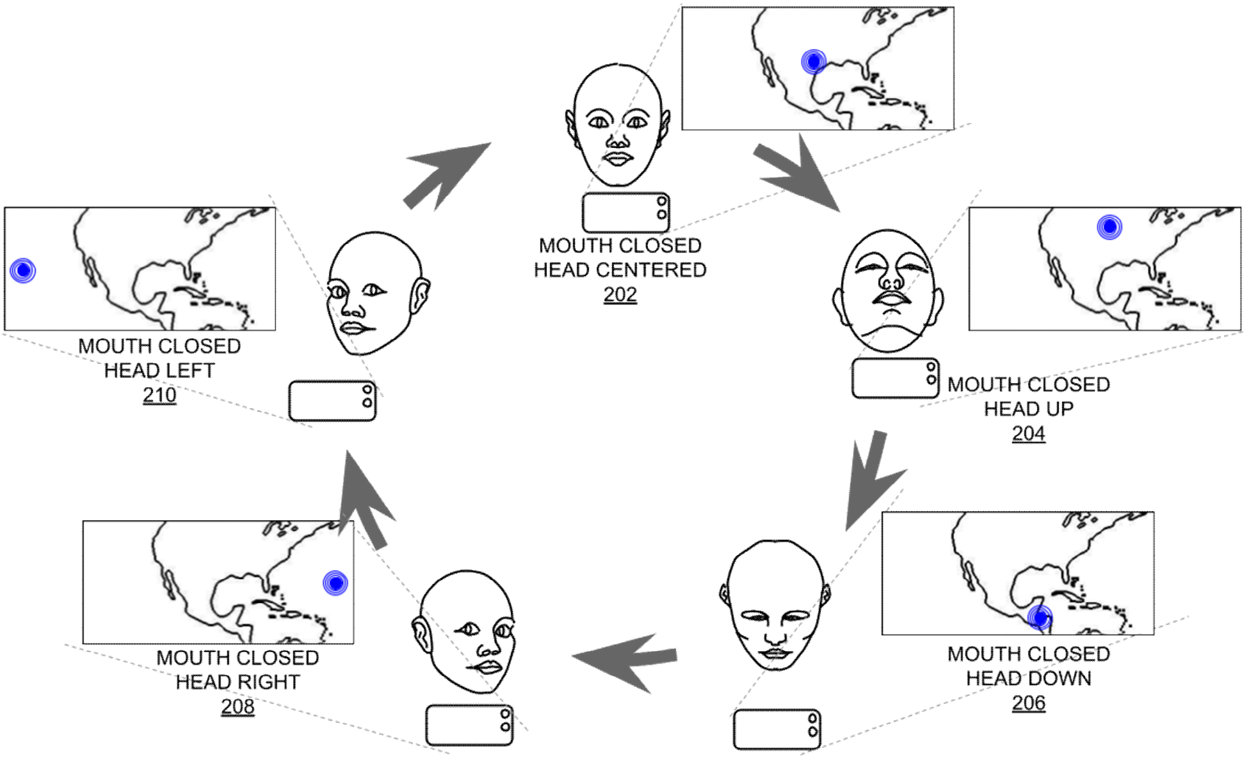

Fig. 2: Combining mouth and head gestures to effect pointing

Fig. 2 illustrates combining mouth and head gestures to effect pointing, e.g., as in Table 1. The user starts with mouth closed and head centered (202), which results in a centered pointer. The user moves their head up (204), resulting in the pointer moving northward. The user moves their head down (206), resulting in the pointer moving southward. The user turns their head right (208), resulting in the pointer moving eastward. The user turns their head left (210), resulting in the pointer moving westward. The user re-centers their head to bring the pointer back to the center (202).

Gesture 1→ **Gesture 2↓**	**Move head closer to the camera**	**Move head away from the camera**	**Head in stationary (relative)**
State 0: Normal	NA	NA	NA
State 1: Frown	Zoom in	Zoom out	NA

Table 2: Combined head gestures for zooming in and out

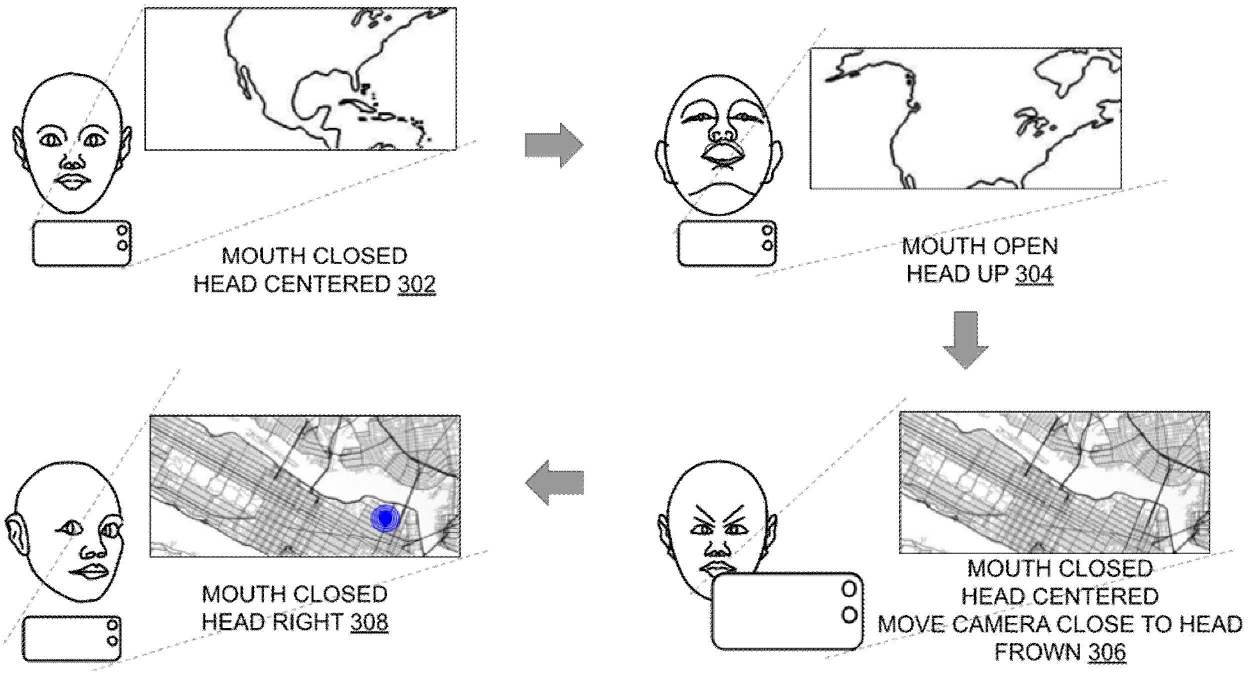

Fig. 3: Combined head gestures to pan, zoom, and point to a particular spot on the map

Fig. 3 illustrates combined head gestures to pan, zoom, and point to a particular spot on the map, e.g., as per tables 1-2. The user starts with mouth closed and head centered (302), resulting in a stationary map. The user opens their mouth and moves their head up, resulting in the map panning northward (304). The user centers their head, closes their mouth, moves the smart device close to their head, and frowns (306), resulting in a zoom-in to a particular point on the map. The user turns their head right (308), resulting in a particular spot in the zoomed-in map being pointed to.

Users can customize their own gesture combinations based on their abilities and preferences. For example, while some people may control zoom by frowning, others may do so by smiling. Furthermore, since binary gestures are defined by facial landmarks, users can define their own gestures.

Gesture 1→ **Gesture 2↓**	**Turn head left**	**Turn head right**	**Turn head up**	**Turn head down**	**Move head closer to the camera**	**Move head away from the camera**	**Head centered (relative)**
State 0: Mouth closed (same as Table 1)	Move pointer left	Move pointer right	Move pointer up	Move pointer down	NA	NA	NA
State 1: Mouth open	Look left	Look right	Look up	Look down	Move forward	Move backward	NA

Table 3: Combined head gestures to explore street images

In this manner, the techniques of this disclosure combine different types of head gestures to achieve superior user interaction with smart device applications such as digital maps. The techniques extend gesture combinations, which have traditionally been operative *between* modalities (e.g., between mouse and keyboard), to become operative *within* modalities (e.g., eye movements and head movements, both being types of head gestures). The combined head gesture user interface described herein augment existing accessibility options, e.g., speech commands of apps. The techniques improve application accessibility for people with and without disabilities.

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 the collection of user information (e.g., information about a user’s gestures, facial or other actions detected by a camera or sensor, profession, a user’s preferences, or a user’s current location), 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.

CONCLUSION

This disclosure describes techniques to fuse multiple head gestures, e.g., head, eye, mouth, or eyebrow movement, to provide superior user interaction. Head gestures are classified as analog (e.g., eye movements, which provide continuous input) or binary (e.g., frowns, which indicate one-time operation). An analog gesture is fused with multiple binary gestures to improve efficiency. For example, to move an object, the user selects the object by looking and frowning at it; moves the object using eye gaze; and frowns again to set it in position. Different facial expressions can be flexibly assigned to accommodate varied user abilities and preferences.

REFERENCES

[1] Turk, Matthew. "Multimodal interaction: A review." *Pattern recognition letters* 36 (2014): 189-195.

[2] Krafka, Kyle, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra Bhandarkar, Wojciech Matusik, and Antonio Torralba. "Eye tracking for everyone." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2176-2184. 2016.

[3] Sun, Wei, Yezhao Fan, Xiongkuo Min, Shihao Peng, Siwei Ma, and Guangtao Zhai. “Lphd: A large-scale head pose dataset for RGB images.” In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1084-1089. IEEE, 2019.

[4] Vezzetti, Enrico, Federica Marcolin, Stefano Tornincasa, and Pietro Maroso. “Application of geometry to RGB images for facial landmark localization: A preliminary approach.” *International Journal of Biometrics* 8, no. 3-4 (2016): 216-236.