Design and Implementation of a Multimodal Input Device Using a Web Camera

Jongwhoa Na, Wonsuk Choi, and Dongwoo Lee

ABSTRACT—We propose a novel input pointing device called the multimodal mouse (MM) which uses two modalities: face recognition and speech recognition. From an analysis of Microsoft Office workloads, we find that 80% of Microsoft Office Specialist test tasks are compound tasks using both the keyboard and the mouse together. When we use the optical mouse (OM), operation is quick, but it requires a hand exchange delay between the keyboard and the mouse. This takes up a significant amount of the total execution time. The MM operates more slowly than the OM, but it does not consume any hand exchange time. As a result, the MM shows better performance than the OM in many cases.

Keywords—Multimodal mouse, workload-based evaluation.

I. Introduction

Recently, various types of camera mice using the inexpensive COTS web camera have been introduced [1], [2]. The camera mice operate as follows: (1) recognize a facial feature (such as a nose) in the input image, (2) find the location of the nose, and (3) map the location to a pixel on the monitor.

The problem with the camera mouse (CM) is that there is a size mismatch between the imaging sensor of the web camera (about 0.25″) and the monitor (15″ to 24″). If the CM uses one-to-one matching between the pixel of the imaging sensor and that of the monitor, the stability of the monitor is decent. However, the area the user can conveniently reach is limited, which means that the user must extend his/her body to compensate. On the other hand, if the system has one-to-many matching, the area the user can reach becomes larger, but the mouse pointer becomes less controllable. Moreover, the human body cannot be completely immobile due to breathing. Thus, unwanted movement in the imaging sensor occurs so that the cursor of the CM may exhibit some vibration. An optimized matching algorithm is still an open research issue. A detailed discussion about the CM can be found in [1] and [2].

Multimodal systems process two or more combined user input modes in a coordinated manner to improve usability, accessibility, and execution time. Multimodal systems utilizing speech and pen input, speech and lip movement, and gaze tracking and manual input have been proposed [3], [4] for various applications, such as the human-centric word processor, portable voice assistant, VR aircraft maintenance training, and voice activated control (VAC). Until recently, multimodal systems have suffered from slow signal processing for speech or image inputs. However, due to endless development in IT technologies, some commercial multimodal products have been launched. For example, using an advanced automated speech recognition engine, the VAC of Lockheed Martin's F-35 Joint Strike Fighter provides the pilot with immediate access to flight management system functions, even as the pilot maintains hands-on control of the aircraft [4]. Multimodal systems have been an active research area due to their potential. (see [5]).

We propose the multimodal mouse (MM) which uses two modalities: face recognition (FR) and speech recognition (SR). The proposed MM is unique in two aspects. First, it is designed as a novel input device for the PC. It uses FR, SR, and a keyboard concurrently so that the user's hands do not have to leave the keyboard. From an analysis of Microsoft Office Specialist (MOS) test problems, we found that approximately 70% to 80% of the problems require both keyboard and mouse, so they require hand exchanges. The MM can eliminate the time consuming hand exchange between keyboard and mouse. Using Microsoft Office

Manuscript received Apr. 18, 2008; revised May 19, 2008; accepted June 17, 2008.
This paper was supported by the Korea Aerospace University Research Grant.
Jongwhoa Na (phone: + 82 2 300 0410, email: jwna@kau.ac.kr), Wonsuk Choi (email: choiws@kau.ac.kr), and Dongwoo Lee (email: dongwoo81@kau.ac.kr) are with the Department of Electrical Engineering, Korea Aerospace University, Goyang, Rep. of Korea.

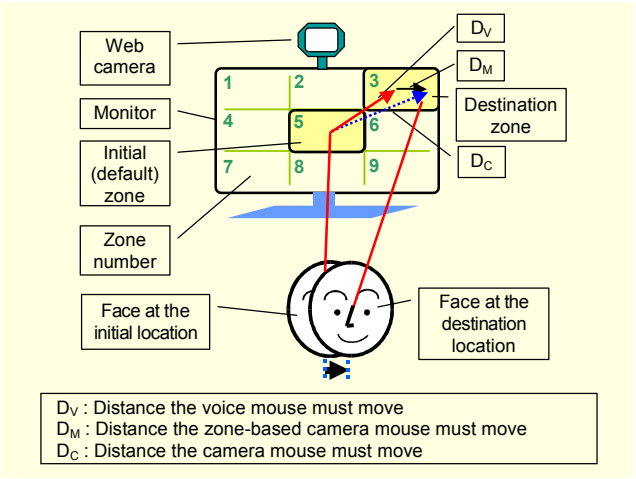

Fig. 1. Overview of the multimodal mouse.

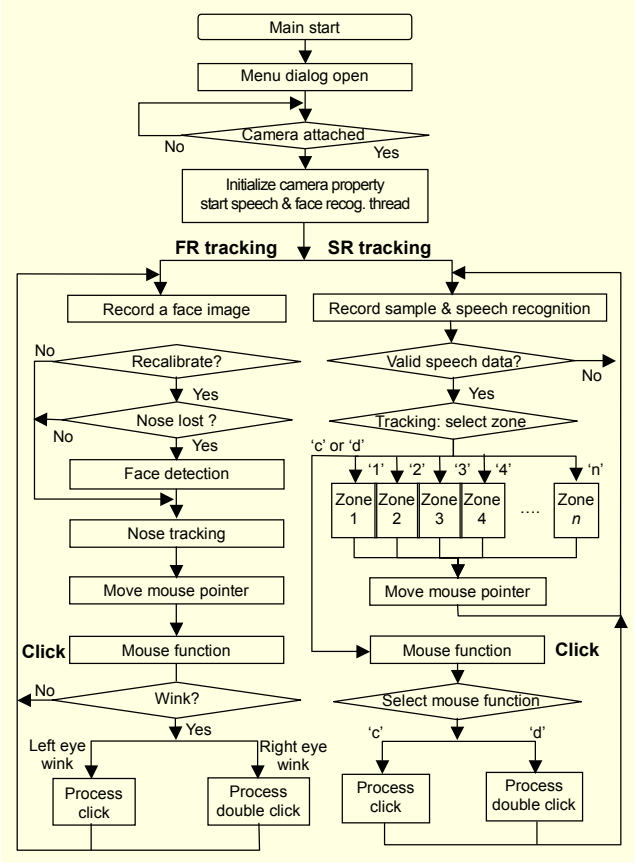

Fig. 2. Flowchart of the proposed multimodal mouse.

workloads, we have shown that the MM may outperform the optical mouse (OM). Second, the MM is designed and implemented to perform optimally even with an inexpensive COTS web camera.

The operation of the proposed MM is illustrated in Fig. 1. The MM is a CM augmented with SR for the functions of the mouse. The MM tracking process is the same as that of the CM, which uses face recognition. What distinguishes the MM from the CM is its triggering of zone selection by speech recognition. In the current implementation of the MM, the monitor area is partitioned into nine zones. Explained simply, the user can select any zone by speaking its number at the recording stage of the MM. After the recording stage, the MM sends the recorded speech data to the SR engine for interpretation. When the program receives a correctly-recognized zone number, it changes the coordinates of the mouse pointer from the current zone to the center of the newly-recognized zone. The user can then move to the target location from the center of the new zone by moving the mouse pointer along a reduced path.

For example, consider the user trying to move the mouse pointer from a point in the fifth zone to a destination point in the third zone. With a CM, the user must move to D_C by extending the upper body. However, with the MM, one can move to the destination by moving D_V using speech command (SR) and moving the reduced D_M using FR concurrently.

In the Microsoft Windows environment, many desktop buttons and menu boxes are near the edges of the monitor, while the usual location of the mouse pointer is at the center of the monitor. As a result, MM users can control the mouse pointer over the whole area of the monitor with less movement than a CM user.

II. Design of the Multimodal Mouse

The MM consists of two stages as shown in Fig. 2. In the first stage, tracking is performed using FR and SR concurrently. The FR-based tracking module finds the user's face using a Harr algorithm. Using a knowledge-based method, the tracking module finds the user's nose. Then, it relates the location of the nose to the location of the mouse pointer using an optical flow algorithm. For a detailed explanation of the FR-based tracking algorithm and its implementation, see [1] and [2]. In addition to FR, SR is used to track the mouse pointer. The SR-based mouse tracking module records voice commands from the user and performs SR. A recognized voice command is used to select the corresponding zone. We used Korea Telecom's KT-Huvois discrete SR engine.

In the second stage, after the tracking stage, the MM performs various mouse functions using the following three methods: FR, SR, and keyboard input. In the current implementation, the user performs the click function with a wink of the left eye and the double click function with a wink of the right eye. Also, the user can perform the same function by speaking "click" and "double click."

The SR used is a discrete processing engine. The discrete SR engine wastes the synchronization time between the user and the MM. For synchronization, the MM used three methods: a keyboard-based interrupt, a periodic beeping sound, and the

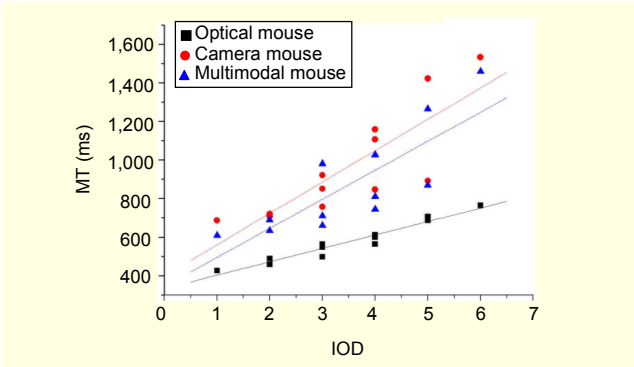

Fig. 3. Experimental results of the Fitts' test.

periodic display of an icon. First, the user presses the control key, which is hooked to the speech recording interrupt service routine. Then the SR processes the user's commands. Second, the user can respond at the sound of the beep to start recording the speech command. Third, for convenience, the MM displays a small red square in the upper-left corner of the monitor to indicate the recording stage.

III. Multimodal Mouse Implementation and Experiments

The MM was implemented using Visual Studio .NET 2003 on a 1.8 GHz Pentium 4 desktop PC equipped with an inexpensive COTS web camera of 300k pixels and an inexpensive microphone. In the current implementation of the MM, when the user wants to move the mouse pointer to the corner of the monitor, the user can control the mouse pointer easily by changing the zone in the MM. In Fig. 1, the nine zones are identified with the green lines on the screen. The MM displays a red square icon and sounds a beep to inform the user that the system is in the recording stage.

First, we performed Fitts' test to find the performance of the three input devices. In the Fitts' test, we set the movement distances D to 32, 64, and 128 pixels and the target widths W to 8, 16, 32, and 64 pixels. Three college students tested the three input devices 30 times over two days. The experimental results shown in Fig. 3 indicate that the OM showed the best moving time, followed by the MM and CM. The MM shows better performance than the CM due to improved controllability.

Next, we performed a workload-based evaluation to find the effects of the three input devices over the entire computer system. The experimental setup consisted of a test window and the workloads. The workloads were two representative test questions from the 2006 MOS Certification problems (see Table 1). The detailed workload-based evaluations of the OM, CM, and MM are presented in [6].

The experimental results for the OM, CM, and MM are summarized in Table 2. The PPT test requires four movements between a keyboard and a mouse; thus, the MM outperforms the OM by 6.7%. In the Outlook test, which requires three consecutive mouse movements, the OM is faster than the MM by 13.6%. Although the execution times of the no-hands mice are slower than the OM, when we consider the execution time of useful workloads, the difference in performance is negligible.

Table 1. Two representative problems in MOS test.

	Powerpoint (PPT) test	Outlook test
Step 1	Move "Title" box & click	Move to a mail & double click
Step 2	Type "My Show"	Move to Forward button & click
Step 3	Move to "Subtitle" box & click	Move to "text" box & click
Step 4	Type "My name"	Type "test"
Step 5	Move to STOP button & click	Move to STOP button & click

Table 2. Workload-based evaluation results (in seconds).

Test type	Optical mouse	Camera mouse	Multimodal mouse
PPT	9.6	9.1	9.0
Outlook	6.8	7.3	7.8

IV. Conclusion

In order to improve the performance and the usability of the no-hands mouse, a novel multimodal mouse (MM) is proposed. The proposed MM uses an inexpensive COTS web camera and microphone for tracking and clicking. In the MM, the monitor is divided into $n \times n$ zones such that face recognition is performed in a reduced monitor area while speech recognition is performed to change zones. The experimental results showed that the MM can be a viable alternative to the OM and CM.

References

[1] M. Betke et al., "The Camera Mouse: Visual Tracking of Body Features to Provide Computer Access for People with Severe Disabilities," *IEEE Trans. Neural Sys. and Rehab. Eng.*, vol. 10, no. 1, Mar. 2002, pp. 1-10.

[2] J.W. Na et al., "A Novel Camera Based Computer Input Device," *ICMOCA*, 2006, pp. 61-64.

[3] S.L. Oviatt and P. Cohen, "Multimodal Systems that Process What Comes Naturally," *Comm. of the ACM*, vol. 43, no. 3, pp. 45-53.

[4] Voice Activated Cockpits, http://www.airport-int.com

[5] R. Sharma et al., "Toward Multimodal Human-Computer Interface," *Proc. IEEE*, vol. 86, no. 5, May 1998, pp. 853-869.

[6] J.W. Na et al., "Applicability of No-Hands Computer Input Devices for the Certificates for Microsoft Office SW," *ICCHP*, 2008, pp. 1169-1176.