Human Factors Evaluation of a Vision-Based Facial Gesture Interface

Gamhewage C. De Silva, Michael J. Lyons, Shinjiro Kawato & Nobuji Tetsutani

ATR Media Information Science Labs
2-2-2 Hikaridai, Keihanna Science City
Kyoto, 619-0288, Japan
mlyons@atr.co.jp

Abstract

We adapted a vision-based face tracking system for cursor control by head movement. An additional vision-based algorithm allowed the user to enter a click by opening the mouth. The Fitts law information throughput rate of cursor movements was measured to be 2.0 bits/sec with the ISO 9241-9 international standard method for testing input devices. A usability assessment was also conducted and we report and discuss the results. A practical application of this facial gesture interface was studied: text input using the Dasher system, which allows a user to type by moving the cursor. The measured typing speed was 7-12 words/minute, depending on level of user expertise. Performance of the system is compared to a conventional mouse interface.

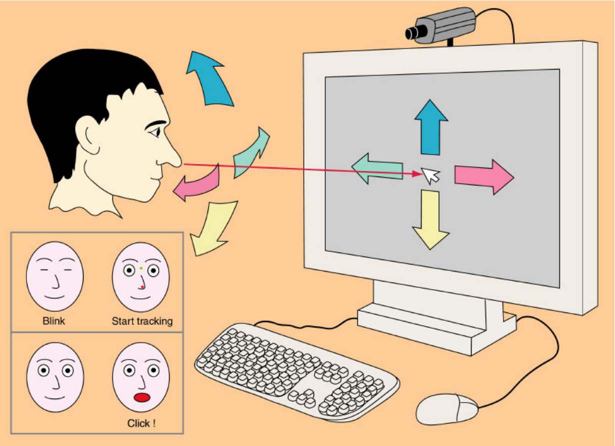


Figure 1: Schematic of the face tracking interface.

1. Introduction

There is a growing interest in computer interfaces that go beyond the standard keyboard/mouse human-computer interaction (HCI) paradigm. Input methods which integrate information from several sensory modalities, including sight and sound, are being considered for use in perceptual user interfaces, or PUIs [21]. Steady advances in CPU performance, the widespread availability of inexpensive plug-and-play cameras which capture video at full frame rate, and significant advances in computer vision research make vision-based methods one of the most promising areas for progress in PUI research. The importance of the face in human communication suggests that user interfaces (UIs) which process facial information will be popular and, not surprisingly, numerous works have studied vision-based methods for tracking the face [1, 3, 5, 6, 15, 19, 20] - to list only a few.

A review of the face-tracking literature is not the aim of the present communication, however suffice it to say that several groups have developed algorithms are approaching the level of robustness needed for general use in real-world applications. Most of these works end with a successful demonstration of face-tracking functionality, and sometimes include a simple application of their method. An important exception is a study [4] of the HCI aspects of using a face-tracking UI for dual-pointing tasks, which found that use of the head-based input stream was appropriate for asymmetric control tasks. Other than this recent study,there seems to have been relatively little effort to characterize the HCI aspects of vision-based UIs. Before vision-based UIs can be taken seriously as input methods, however, the human side of the interaction needs to be considered.

A significant fraction of the HCI literature is concerned with systematic and reliable evaluation of input devices such as mouse, joystick, and trackball, beginning with the influential work of Card and colleagues [2]. In the current paper we apply standard human factors evaluation methodology that has evolved in the HCI field, to characterize a facial gesture UI developed in our group, which allows the user to point and click using movements of the head and mouth. To demonstrate the value of the standard evaluation methodology we show how it may be used to accurately predict performance with a real-world application, hands-free text entry.

An outline of the paper is as follows: section 2 describes the algorithms used in this paper to (a) initialize the face-tracking by blink detection (b) track the tip of the nose for pointing (c) detect mouth opening to allow the user to input click events; section 3 briefly introduces Fitts law and standard methodology for measuring the information throughput of a pointing device, then describes our experimental measurement of the throughput of our UI, comparing it to that of a standard computer mouse; section 4 describes a usability assessment using a user questionnaire; section 5 describes how we harnessed our perceptive UI to drive the recently developed Dasher text entry software so that head movements could be used to type, and reports typing speeds, comparing them with predictions from the throughput measurements of the previous section; finally section 6 concludes the paper with some suggestions for improvements and further study.

2. Face Tracking System

The face-tracking system used here modified a previous face-tracking system developed in our group, to explicitly track the location of the tip of the nose and use it for cursor control (a schematic is shown in Figure 1). First the eyes are detected. Next, the location of the nose tip is estimated within a small region below the eyes. The nose tip is convex and usually only one point in the estimated region has this feature [8]. This reduces the difficulty of detecting the nose tip, which may then be tracked using template matching, with an adaptive template that is updated each video frame.

2.1. Detecting and tracking the eyes

Blink detection is used to detect the eyes [11]. In detection mode, differences the previous and current frames are calculated, and pixels where the luminance change exceeds a threshold are extracted. If blinking occurs, a pair of regions corresponding to eyelid movement areas are expected, having certain geometrical properties (size, distance, alignment). This is complicated by the fact that the head may move, hence thresholded pixels can include not only the eyelid regions but other parts of the face. To cope with this problem, head movement is estimated and used to cancel pixels where predicted and observed luminance change match. Subsequently, we apply labeling and connected component analysis and geometrical relation tests for candidate eye pair locations. Head movement estimation is not based on area patterns but only on pixels extracted from the frame difference image, which means the computation is not intensive and may be done in real-time [11].

Subsequent to eye detection and location, the eyes are tracked. However, blinking causes rapid and drastic changes and even adaptive templates cannot follow these changes. Instead, a “Between-the-Eyes” template [12] is used to track the location of the face. Compared to other areas of the face its pattern is relatively stable for changes in facial expression. It has a relatively bright part at the nose-bridge and relatively dark parts at the eyes like wedges on both sides. This is a very good feature for accurate location by template matching. After the detection of “Between-the-Eyes,” the eyes are searched again in very small areas, because their positions relative to the “Between-the-Eyes” template are known in the previous frame. In turn, the “Between-the-Eyes” template is updated each frame, based on the current eye positions [11].

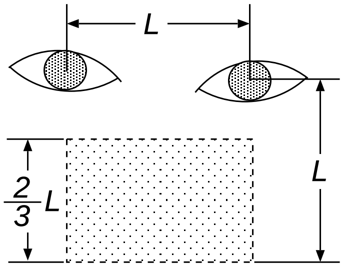


Figure 2: Nose tip search area relative to the eyes.

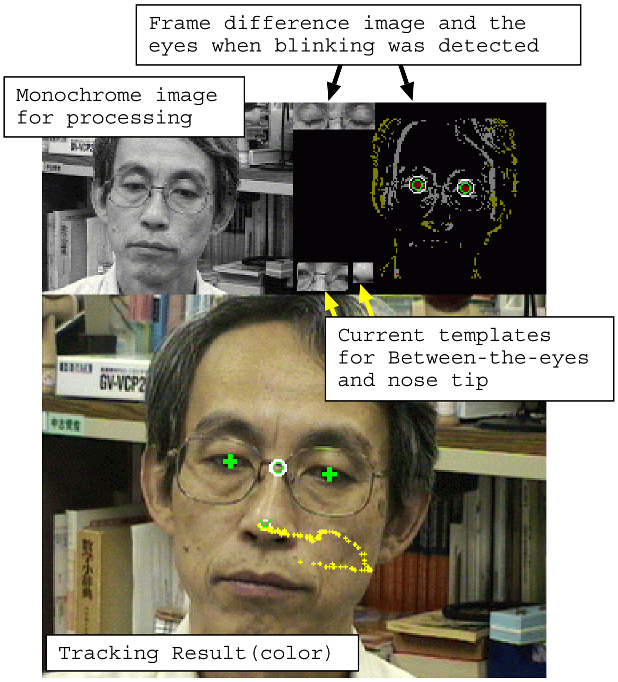

Frame difference image and the eyes when blinking was detected
Monochrome image for processing
Current templates for Between-the-eyes and nose tip
Tracking Result (color)

Figure 3: Detecting and tracking points corresponding to between-the-eyes, eyes, and nose tip.

2.2. Detecting and tracking the nose tip

After the eyes are located, it is relatively easy to detect the nose tip, which is convex shaped and somewhat specular, hence possesses a highlight. Although the precise location of the highlight depends on face orientation and lighting direction, it is located on the nose tip. Moreover, for fixed lighting, to extent that the nose tip approximates a sphere, the precise location of the highlight is not strongly affected by orientation of the face.

Figure 2 shows the search area for the nose tip relative to locations of the eyes. The brightest point in this area is taken as a candidate for the nose tip. If the distances from this point to the two eyes are nearly equal, we assume it is the nose tip, and begin tracking.

In tracking the nose tip, we also use a continuously updated template. A small rectangular pattern centered at the nose tip is saved as a template for the next frame. In the current frame, the best matching point with the template is searched around the previous position. Then the nose tip is registered again to the brightest point in a very small region around the matching point. Then, the nose tip template is updated. If it goes out of region shown in Fig. 2, we assume the nose tip is lost, and start again from detection. Figure 3 shows the results of the facial tracking algorithm, indicating positions of the detected eyes and nose tip.

2.3. Displacement/Angle Gain

Selection of a control movement to display output gain setting is important for pointing devices and head pointers are no exception [17]. Two possible measures of gain are (a) D/A: the ratio of D, the displacement on the display device (in centimeters) to the rotation angle A of the head (in degrees), and (b) A'/A, where the displacement on the display device is measured as a visual angle, A'. With 3D face tracking it is a simple matter to have a linear angular display displacement for angular head rotations, but it has been suggested that a constant D/A ratio is appropriate for head-controlled pointing [17].

We track the nose in the 2D image, and map its displacement from the center of the image to displacement of the cursor from the center of the display. As a check, head rotation angle A was measured by setting gaze landmarks on a wall behind the computer display. For the range of angles needed to cover the display area the relation between D and A was found to be approximately linear. We adjusted the gain so that the entire horizontal display could be covered without uncomfortable movements resulting in a gain of 0.47 cm/deg, equivalent to an angular gain of about 4.35. This lies close to the gain found by earlier study [17] to minimize cursor movement times.

Rotation of the head (looking left and right) requires less effort than extension/flexion (looking up and down), so choose the gain to be anisotropic with a fixed ratio of vertical to horizontal D/A gain of 1.4. This allowed the vertical range of the display to be covered without uncomfortable movements.

2.4. Using the mouth to click

Intentional motion of the mouth is facile, hence we decided to use a mouth opening gesture for a “mouse click.” We implemented a simplified version of the previously studied interface for using mouth movements for input to a computer interface [13]. First, the approximate location of the mouth is estimated using the face tracking results. This is refined based on the location of a local intensity minimum corresponding to the region just beneath the upper lip. Finally, an open mouth is identified by the size of a region of pixels having intensity less than a certain threshold and red component larger than another threshold. This procedure makes use of the fact that under a large range of lighting conditions the open mouth robustly exhibits a shadow area which is darker than surrounding teeth and skin areas. This is more robust to changes in lighting conditions than are skin-color thresholding algorithms, for example, because the region is not a surface, but a cavity. In this work, we mapped the opening of the mouth to a single click event. Closing the mouth, or keeping it open has no effect. In future work, we plan to study more elaborate mappings of mouth gestures, to allow, for example, double clicks, or clicking and dragging.

3. Performance Evaluation

3.1. Fitts’ Law

Based on studies of the tradeoff between speed and accuracy in aimed movement tasks, Fitts [7] proposed a linear relationship between the *difficulty*, $ID = log_2(2D/W)$ (for a target of width W, at a distance D), of a task and the movement time, MT, taken for its completion. Numerous studies have used Fitts law to compare input devices [2]. Fitts’ law has evolved and recent work [14, 9] uses the following form:

$$MT = ID_e/Throughput \qquad (1)$$

The throughput, in bits/second, is an index of performance of the human-machine interface for an aimed target acquisition task, and an international standard method for measuring the information throughput of pointing devices, the ISO 9241-9 [9] has developed. In the ISO standard (also see [14]) the effective index of difficulty, ID_e is an explicit function of the accuracy with which targets are selected:

$$ID_e = log_2(D/W_e + 1) \text{ with } W_e = 4.133 SD_x \qquad (2)$$

where SD_x is the standard deviation of the target selection coordinate measured along the axis of approach to a target. The effective index of difficulty depends on the effective target width, which is based on the accuracy with which the task is performed.

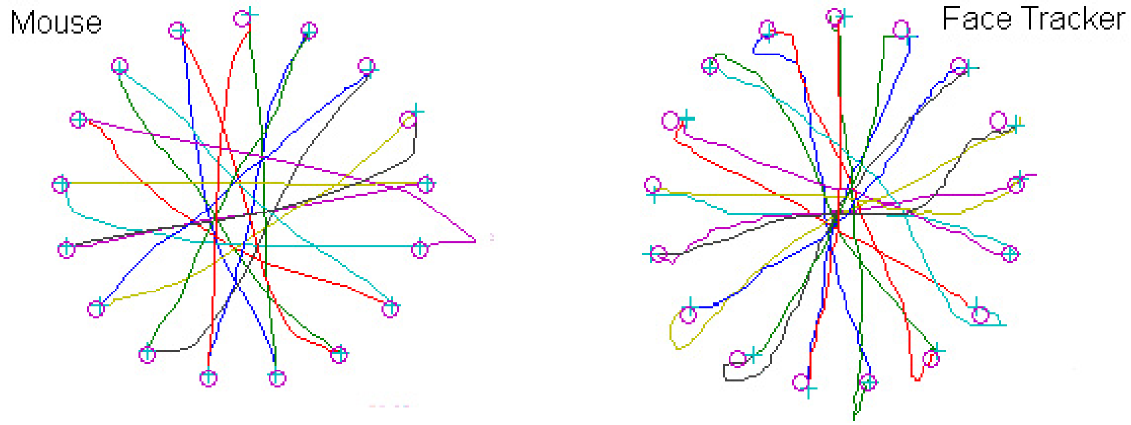


Figure 5: Sample trajectories for the ISO standard multi-directional tapping task.

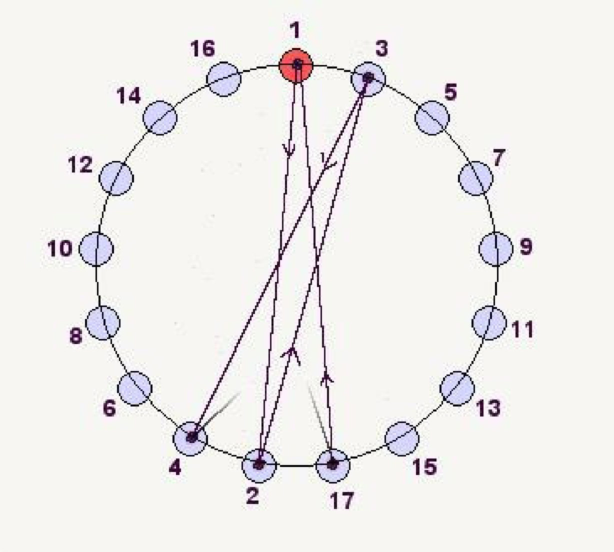


Figure 4: Multi-directional tapping task.

3.2. Multi-direction Tapping Task

Previous work using head-worn cursor control systems, has shown Fitts' law to hold for head movement tasks [10]. However, there seem to be no values in the published literature for the throughput of a vision-based face tracking pointer. Hence we decided to make a careful measurement of the information throughput of the pointing function of our UI, and at the same time a measurement of the performance of a standard mouse as a check of our methodology.

Figure 4 illustrates the ISO 9241-9 task [9] as implemented in our experiments. A 240 pixel diameter circle was displayed at the center of a 640x480 pixel resolution monitor. Seventeen circular targets (diameter: 21 pixels) were spaced equally around the perimeter of the circle. Subjects were required to move the cursor from one target to another, neighboring the diametrically opposite position, according to a pre-defined sequence indicated in the figure. Subjects pressed the space bar to indicate reaching a target, and the next target was highlighted by changing it's color to red. The index of difficulty of each trial of this task, calculated using the above dimensions and is 3.4, without the correction for effective target width.

3.3. Experimental Procedure

Eight volunteers participated in the experiment. All were regular computer users and familiar with use of the mouse, but none were involved in development of the face-tracking system, or had previously used it. Each subject was briefed about the task at the beginning of each experiment. Warm-up trials were not permitted in order to allow observation of the effect of learning.

Each test subject performed a total of 20 repetitions of the multi-directional pointing task, alternating between using the mouse and the face-tracker to control the cursor, starting, in all cases with the mouse since the users are experienced mouse users. Breaks were allowed between blocks. Total time for completion of all 20 blocks was about 40 minutes. To calculate throughput values, pixel coordinates of the cursor at time of target selection were recorded, together with time taken to reach it. For further device characterization, coordinates of the cursor movement path were sampled at a rate of 32 samples/second and stored.

3.4. Results

Sample trajectories for one user completing one block of the experiment are shown in figure 5. The shakier movement of the cursor seen with the face tracker partially reflects jitter of the tracking system, but it is also a function of head movement behavior. As the mouse rests inertly on a flat surface its trajectories are not surprisingly smoother.

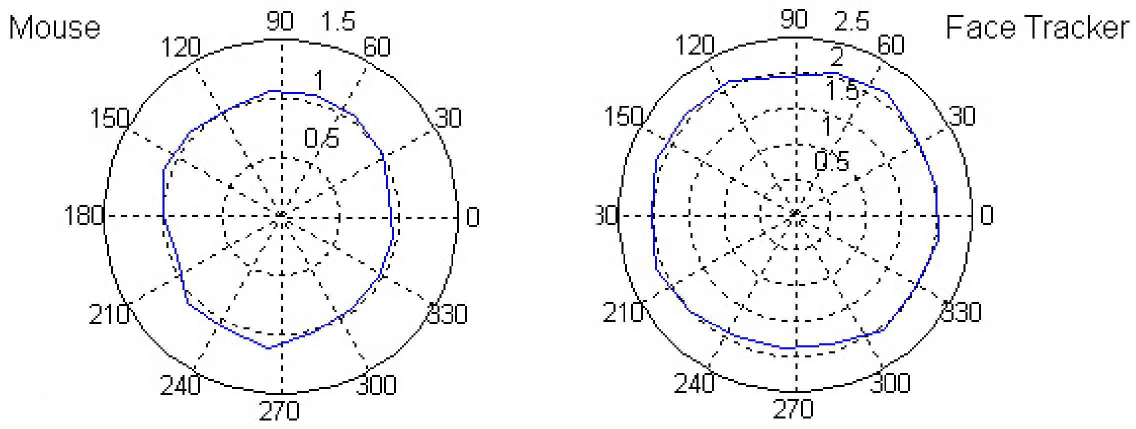


Figure 7: Average movement time (sec) versus orientation (deg) for the ISO standard multi-directional tapping task. Neither the mouse nor the vision-based face tracker shows significant orientation dependence of movement times.

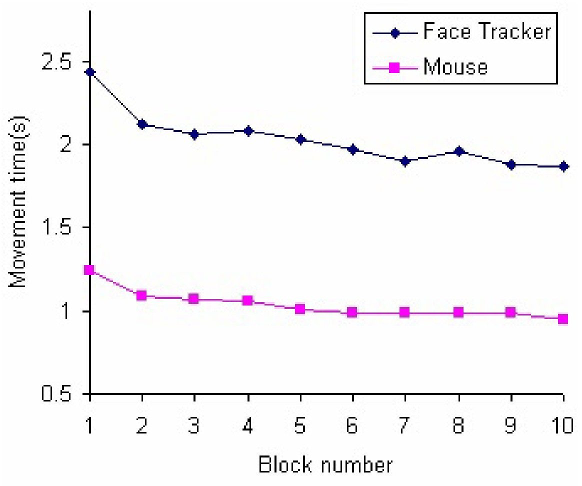


Figure 6: Learning curves.

Figure 6 plots movement times, averaged over the eight subjects, for successive blocks. Movement times decrease with successive blocks for both devices, reflecting motor learning. The mouse resulted in generally faster interaction than the face tracker. With the face tracker, we observed a more rapid decrease of movement times, especially for the first three blocks. Since the subjects had little experience using a face tracking system to point, this may be the result of learning two skills, the ISO task and use of the vision-based UI, at the same time.

The mouse throughput, averaged over our last five trials, was 4.7 bits/sec, which is similar to the value of 4.9 bits/sec measured recently also using the ISO task [14]. For the face tracking pointer, the average throughput was 2.0 bits/sec. This value exceeds the 1.8 bits/sec reported for a joystick, but is lower than the 3.0 bit/sec for a trackball and the 2.9 bit/sec for a touchpad, measured previously using the ISO task [14].

Individual user mouse and nose pointer throughputs were not significantly correlated (Pearson r = -0.05; Spearman's ρ = 0.05), suggesting that there is no strong relationship between motor skills for using these pointing devices.

In contrast to findings with a head-worn head tracking system [16] the vision-based system we studied showed no significant dependence of throughput on movement direction orientation, as shown in the polar plot of throughput versus orientation, as may be seen in figure 7.

4 Usability Assessment

We conducted a preliminary assessment of usability and user comfort. Our greatest concern was for the possibility of neck fatigue through repetitive head movements. After completing the 20 blocks of the ISO task, the eight volunteers were asked to fill in a questionnaire adapted from the ISO 9241-9 document [9]. Subjects rated the vision-based user interface on eight criteria related to performance of the interface and comfort or fatigue of operation. A seven point response scale was used with 1 being the worst rating (very poor performance or greatest discomfort) and 7 the best (very good performance or greatest comfort). Below we list the criterion descriptor, response mean, mode (in parentheses with more than one response number if the distribution is flat), and the range of responses:

- Strength required 4.5 (5) 3-6
- Smoothness 3.9 (2,3,6) 2-6
- Effort required 4.0 (3,5) 3-5

- Accuracy 3.5 (2,3,4,5) 2-5
- Speed 4.1 (2) 2-7
- Comfort 4.2 (4) 3-5
- Fatigue 4.0 (4) 2-6
- Overall 4.9 (6) 4-6

Overall neck effort was rated on Borg's 11 point scale [9], with 0 indicating no neck effort and 10 indicating very, very strong neck effort. Mean response was 3.7 (3=weak, 4 = moderate), the range of responses was 0-6 with a mode of 4.

In the list above we note that average responses lie close to the middle of the usability scale. The weakest points were smoothness and accuracy of cursor control and the strongest points were the overall usability and low demand on neck strength.

In addition to rating using independent usability scales, the ISO standard also recommends comparative ratings with different devices. Since our volunteers had much greater familiarity with the conventional mouse than with the nose pointer, we decided that it would be difficult to conduct a meaningful comparison in this case and decided not to pursue this experiment.

4.1. Descriptive User Feedback

The questionnaire concluded with two qualitative questions. Answers to the first question, "What are your suggestions for improvements?" are listed below (number of subjects responding this way is indicated in parentheses):

- greater smoothness (3)
- greater accuracy (2)
- more displacement gain (2)
- less displacement gain (1)
- velocity rather than position control (1).

Most users suggested improvements in smoothness and accuracy of the pointing device, in agreement with the results from the first part of the questionnaire, listed in the previous section.

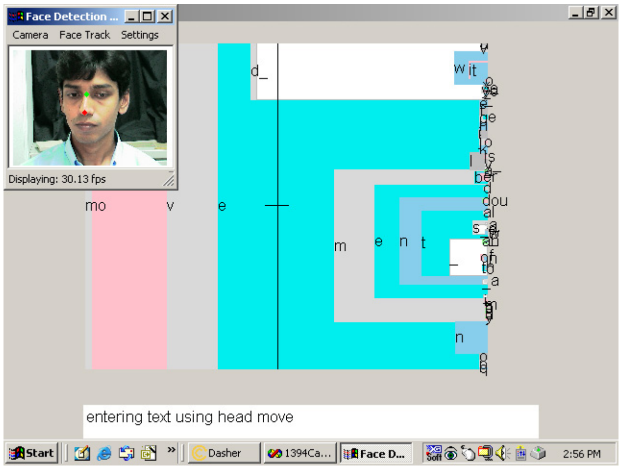
Answers to the second question,"How would you imagine the system being used?", included:

- interface for the disabled (5)
- for use as a dual pointer (2)
- interface for computer games (2).

The absence of answers relating to mobile phone or palmtop computing may be a result of conducting the experiment with a desktop computer, rather than a strong indication that our users felt these were not desirable applications of the technology.

5. Hands-Free Text Entry

Conventional keyboards are inconvenient for text-entry with mobile phones and PDAs, as well as for disabled computer users. Hence text-entry, is a potentially important application for vision-based interfaces. The recently proposed Dasher system [22, 23] allows for efficient text-entry by two-dimensional gestures. Figure 8 shows a screenshot from the Dasher text-entry interface. Letters automatically drift from right to left at a speed controlled by the horizontal position of the cursor. Letters are selected by the vertical position of the cursor. The size of the selection zone for each letter depends the probability that it will occur next in the typed sequence. This probability is estimated using a model trained on a large corpus of English text. This greatly improves the efficiency of text entry: after some practice time users feel they can easily steer the cursor along paths which follow correctly spelled English text.



entering text using head move

Figure 8: Operating the Dasher text entry interface with head movements.

Dasher has been tested with an eyetracking system [23], however this requires special hardware, calibration, and restriction of the user's head position. The face tracking system reported here, by contrast, requires no special hardware and only minor or no calibration adjustments.

5.1. Typing Rates: Predicted and Measured

For the language model used by Dasher, the estimated Shannon information per character is about 2 bits [22]. So, ideally, with the measured throughput for the face tracker of 2 bits/sec, one should be able to type at a rate of 60 character/min. However, we can expect text entry with Dasher to be more difficult than the target selection task. Instead, we use the previously measured median (from a distribution of users) typing speed of 90 characters/min for Dasher operated with a mouse [22], and our measured ratio of mouse to face tracker throughput of 2.35, to predict a typing speed of 38 char/min with the face tracker.

To check this prediction, two subjects familiar with the use of the face tracker used Dasher to enter short excerpts from Hans Christian Andersen's story "The Little Match Girl", which is not contained in the corpus of texts used to train Dasher's language model. Eight excerpts of about 20 words in length were used, for a total of 160 words, or 819 characters. The average measured typing speed for the two subjects was 38 char/min (7.3 wpm), in excellent agreement with the above prediction. One of the authors, who has considerably more experience using Dasher than the two subjects above, averaged 61 char/min (12 wpm) on the same task, which is similar to the theoretical prediction based on the throughput. We note that faster rates should be possible with an improved language model or with adaptation to the user's writing style. As the information content of English text is estimated to be 1 bit/char [18], this could be as great as a factor of 2, for a possible typing speed of 24 wpm, for users with some practice.

6 Conclusion

In this article, we have implemented a video-based interface that allows the user to control cursor position by pointing with the nose and to enter single clicks by opening the mouth. The system is initialized in a hands-free manner by blinking while looking at the center of the display, though minor calibration is sometimes needed for novice users. Using the international standard method for evaluating pointing devices we measured the information throughput of the system and found it to be lower than a computer mouse but slightly higher than a joystick. The utility of measuring the throughput was shown by demonstrating reasonably accurate prediction of the typing rate of hands-free text entry with the Dasher software. With the ability to point with the nose and enter a single click with the mouth, Dasher may be operated in a completely hands-free manner and no special hardware beyond a USB or Firewire camera. This represents an application of the facial gesture interface with immediate usefulness.

Observation of movement trajectories as well as user responses from a questionnaire indicated that the system would benefit from improvements in smoothness and accuracy of the tracking algorithm, which we are currently investigating. In addition, this paper did not systematically evaluate the mouth click functionality of the system and we are still considering how best to do this. Future work with this system will also consider further applications of hands-free cursor and click control.

Acknowledgments

We thank the volunteers who took part in our HCI experiments. GC de Silva thanks the National University of Singapore and the University of Moratuwa, Sri Lanka, from which he took study leave in order to visit ATR. This work was supported by the Telecommunications Advancement Organization of Japan.

References

[1] G. R. Bradski, "Computer Vision Face Tracking for Use in a Perceptual User Interface," *Proc. WACV'98*, pp. 214-219, 1998.

[2] S.K. Card, W.K. English, and B.J. Burr, "Evaluation of Mouse, Rate-Controlled Isometric Joystick, Step Keys, and Text Keys for Text Selection on a CRT," *Ergonomics*, 21(8), pp. 601-613, 1978.

[3] J. L. Crowley and F. Berard, "Multi-Modal Tracking of Faces for Video Communications," *Proc. CVPR'98*, pp. 640-645, 1997.

[4] T. Darrell, N. Checka, A. Oh, and L.P. Morency, "Exploring Vision-Based Interfaces: How to Use Your Head in Dual Pointing Tasks", *MIT AI Memo 2002-001*, 2002.

[5] J.W. Davis and S. Vaks, "A Perceptual User Interface for Recognizing Head Gesture Acknowledgements" *Proc. PUI'01*, 2001

[6] G. J. Edwards, C. J. Taylor, and T. Cootes, "Learning to Identify and Track Faces in Images Sequences," *Proc. ICCV'98*, pp. 317-322, 1998

[7] P.M. Fitts, "The Information Capacity of the Human Motor System in Controlling the Amplitude of Movement," *Journal of Experimental Psychology*, 47, pp. 381-391, 1954.

[8] D. O. Gorodnichy, "On Importance of Nose for Face Tracking," *Proc. FG'02*, pp. 188-193, 2002.

[9] ISO 9241-9:2000(E), "Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs), Part 9: Requirements for Non-Keyboard Input Devices," International Standards Organization, 2000-02-15.

[10] R. J. Jagacinski and D. L. Monk, "Fitts' law in two dimensions with hand and head movements," *Journal of Motor Behavior*, 17, pp. 77-95, 1985.

[11] S. Kawato and N. Tetsutani, "Detection and Tracking of Eyes for Gaze-camera Control," *Proc. VI 2002*, pp. 348-353, 2002.

[12] S. Kawato and N. Tetsutani, "Real-time Detection of Between-the-Eyes with a Circle Frequency Filter," *Proc. ACCV 2002*, vol. II, pp. 442-447, 2002.

[13] M. J. Lyons and N. Tetsutani, “Facing the Music: A Facial Action Controlled Musical Interface” *Proc. CHI’01*, pp. 309-310, 2001.

[14] I. S. MacKenzie, T. Kauppinen, and, M. Silfverberg, “Accuracy Measures for Evaluating Computer Pointing Devices” *Proc. CHI’01*, pp. 9-16, 2001.

[15] N. Oliver, A. P. Pentland, and F. Berard, “LAFTER: Lips and Face Real Time Tracker”, *Proc. CVPR’97*, pp. 123-129, 1997.

[16] R. G. Radwin, G. C. Vanderheiden, and M. L. Lin, “A method for evaluating head-controlled computer input devices using Fitts’ law,” *Human Factors*, 32, 423-438, 1990.

[17] J. A. Schaab, R. G. Radwin, G. C. Vanderheiden, and P. K. Hansen, “A Comparison of Two Control-Display Gain Measures for Head-Controlled Computer Input Devices”, *Human Factors*, 38(3), pp. 390-403, 1996.

[18] C.E. Shannon, “Collected Papers,” ed. N.J.A. Sloane and A. Wyner, IEEE Press, New York, 1993.

[19] L. C. De Silva, M. Tahara, K. Aizawa and M. Hatori, ”Detection and Tracking of Facial Features by using a Facial Feature Model and Deformable Circular Templates”, *IEICE Trans.*, E78-D(9), pp.1195-1207, 1995.

[20] K. Toyama, “Look, Ma - No Hands! Hands-Free Cursor Control with Real-Time 3D Face Tracking” *Proc. PUI’98*, pp. 49-54, 1998.

[21] M. Turk and G. Robertson, “Perceptual User Interfaces”, *Communications of the ACM*, 43(3), pp. 33-34, 2000.

[22] D. J. Ward, A. F. Blackwell, and D.J.C. MacKay, “Dasher - a Data Entry Interface Using Continuous Gestures and Language Models”, *Proc., UIST’00*, pp. 129 - 137, 2000.

[23] D. J. Ward and D. J. C. MacKay, “Fast hands-free writing by gaze direction”, *Nature*, 418, p. 838, 2002.