A Preliminary Evaluation of Head and Facial Feature Tracking for Input on Mobile Devices

Kathryn Carnegie, Stuart Fleming, Esfandiar Ammirahimi, Andreas Komninos
Glasgow Caledonian University
70 Cowcaddens Rd.
Glasgow G4 0BA, UK
+44 141 3313095

{kcarne10, sflemi10}@caledonian.ac.uk, esfandiar@esfandiar.info,
andreas.komninos@gcal.ac.uk

ABSTRACT
This paper discusses the concept of using head and facial feature tracking as an input mechanism for mobile devices. We present our concept and ideas, along with preliminary findings from two prototype implementations. Suggestions for further work and implications are presented in the final sections of the paper.

Categories and Subject Descriptors
H.5.2 [**Information Interfaces and Presentation**]: User Interfaces

General Terms
Human Factors, Design, Theory

Keywords
Mobile device interfaces, text input, computer vision

1. INTRODUCTION & BACKGROUND
People with mobility or dexterity problems, such as people with movement disabilities or degenerative diseases, typically face difficulty controlling computing devices. Computer control through body motion tracking (e.g. using web-cams for facial feature detection) has been available to disabled users and computer game players for many years. With the advent of 3G mobile phones that incorporate user-facing cameras, we began to wonder whether the tracking of user head movements or facial feature movements might open up new possibilities in creating more natural human computer interfaces for mobile devices in the future. Our initial idea was to afford simple control of mobile devices by using facial feature (eye/nose) or head movement detection in two axes (vertical, horizontal). However we felt that it might be possible to extend our idea to allow text input.

The development of Tanaka-Ishii's TouchMeKey4 keypad [1] proved that text input speeds comparable to those achieved with traditional 12-key keypads could be achieved using only four buttons. An evaluation of the TouchMeKey4 system with eight subjects showed that, after becoming familiar with the system over the course of 10 sessions, users could achieve speeds of between 12 and 23 words per minute. Dunlop [2] also presented findings that support Tanaka's work. The results of Tanaka-Ishii and Dunlop are promising as they suggest that a mobile phone could be controlled using movement in four directions, i.e. up, down, left and right. As such, detecting movement using the secondary (user-facing) camera of a mobile device could possibly be used as an alternative input modality to replace "joystick" controls on mobile devices and also be used to enter text.

Significant research has been conducted into feature and gesture tracking using cameras for the purposes of controlling computers. One method is HeadDev [3] which tracks the position of the head using a webcam, in order to control a mouse pointer. HeadDev tracks the user's nose in order to determine the position of the head, which in turn controls the mouse pointer. The mouse is clicked by the system detecting eye blinks. Another proposed method [4] involves feature tracking for the purpose of menu selection, using an eye-camera which is attached to the top of a workstation monitor. The pose of the head is determined by tracking the position of the eyes, eyebrows and nose. The coordinates of both of the eyes and the nose are then used to determine the user's gaze direction. Selections are made by opening the mouth while the head is stationary. Kawato and Ohya [5] conducted research into the real-time detection of nodding and head-shaking. The research proposed a method for detecting these gestures by tracking the area between the eyes ("between-eyes") and using it as a template. For every frame, the system tracks the movement of the head by selecting candidates using the template, and then updating the template. The method has a high success rate provided the user does not move their head too quickly or rotate it outwith the range of the camera.

Unfortunately, most past research seems to concentrate on the use of standard desktop computers with high processing power and stable cameras that can afford fairly good resolution video for analysis. We haven't been able to find much work relating to real-time video analysis from mobile device cameras, apart from Tierno & Campo [6] who implemented a real-time surveillance system using a mobile phone and J2ME. Severe limitations were encountered due to the inability of the Java platform and slow processor speeds to process images with adequate performance. Closer to our own work, Wang et al [7] looked at the performance of camera phone-based motion sensing as a method of control. Their real-time system uses a mobile device's built-in forward-facing camera to detect the user's hand movements, which can then be used to interact with applications. The evaluation, of the

MobileHCI09, September 15 - 18, 2009, Bonn, Germany.

completed system, showed results that often exceeded those of traditional input methods.

2. MOBILE FACE FEATURE TRACKING

We began our investigation by attempting to utilize the Java (J2ME) environment on a mobile platform (SonyEricsson K800). We implemented a simple algorithm to detect the iris in a pre-defined search area (darkest pixel set). From a set of 10 subjects, we took a set of low-res still images taken using the phone's secondary camera under relatively favourable lighting conditions (lab). Five images of each subject were taken (looking straight, up, left, right, down) and were processed on the device. We found that overall our algorithm was 96% successful in detecting relative movement from centre (up, down, left, right) with an average processing speed of 18ms. Encouraged by these findings, we attempted to implement a real-time eye tracking system but due to the multimedia APIs provided by J2ME, we could not achieve a usable framerate for our system (4fps was the best we could achieve using a SonyErcisson K800 and Nokia N82).

In parallel, we carried out an experiment using 8 participants, in an attempt to judge the subjective quality of images in various lighting conditions, captured using the secondary (user-facing) camera of a Nokia N95. We took two sets of images, one with a light and one with a dark background. The reference image for each set was taken in natural daylight, and subsequent images for comparison were captured in fluorescent light, low natural light (dusk) and very low artificial light. Participants were shown the images in a pairwise fashion, with the reference image first, and then asked to rate the quality of the second image using a five-grade scale. The images were shown to the participants using a PowerPoint presentation and they were asked to rate the second image in each pair on a scale from 1 for poor to 5 for excellent in terms of being able to distinguish gaze direction. Due to the disparity in varying types of lighting conditions and backgrounds, it became apparent that detecting facial features would be problematic in realistic settings (mean ratings ranged from 5 for reference images to 2.4 for very low settings).

As such we began to wonder whether tracking the entire head (which contrasted fairly well even in dark backgrounds) would be a more feasible approach. We designed a simple user interface using a viewfinder (so the user can see their own movement) and a set of outlined direction arrows that "fill up" when movement in their direction is detected (fig. 1). To overcome the shortcomings of the J2ME platform we decided to implement using C++ on a Symbian platform (Nokia S60). We used Nokia's OpenCV library to set up a Camus algorithm to detect optical flow from a series of bitmaps, using a medium grid size and filter size, as well as a granularity of 4. These settings were determined empirically to be optimal in terms of performance.

3. PILOT RESULTS & FUTURE WORK

Five users were involved in pilot evaluations of our developed prototype. We found empirically that when the application was used within the optimal performance distance, of 22 and 38cm from the camera, the accuracy and performance was of a high standard. Use outwith the optimal distance had poorer results, but some accurate tracking did still occur. As such we asked participants to maintain a distance within the optimal performance thresholds as much as possible. Each participant's interaction with the system was observed and captured on video and later analysed. Participants were asked to move freely (we did not give them a pre-determined set of tasks, as we were recording them on video). We performed the tests in both normal and low lighting conditions. Overall the application performed satisfactorily during the usability testing (Normal lighting ~78%, low lighting ~65% correct detections), without any optimisation. Low lighting performance was somewhat surprising, compared to the results of the subjective still image evaluation. One of the participants had a darker skin tone than the other people involved with the usability testing. The application performed well; however successful operation with only one user is not enough to conclude that performance will be as good with darker skin tones as it is with lighter ones.

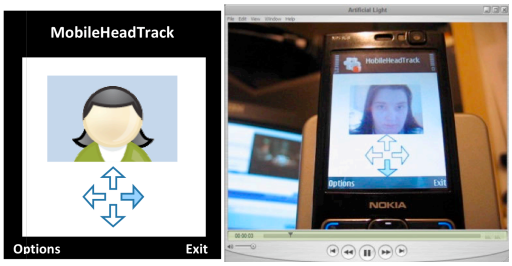

Figure 1. Prototype interface concept and implementation

With encouraging preliminary results, our hope is to extend our work in two ways: Firstly, by creating and evaluating an interface controlled through head movement and secondly, by combining it with a predictive text input system to allow the input of text for communication. Our aim is to test the usability of our prototypes with individuals with and without motor impairments.

4. REFERENCES

[1] Tanaka-Ishii, K., Inutsuka, Y. and Takeichi, M. (2002) Entering text with a four-button device. Nineteenth Intl. Conference on Computational Linguistics, pp. 1-7.

[2] Dunlop, M. D. (2004) Watch-Top Text-Entry: Can Phone-Style Predictive Text-Entry Work With Only 5 Buttons? Proceedings of Mobile HCI 04, pp. 342-346

[3] Manresa-Yee, C., Varona, J. and Perales, F.J. (2006) Towards hands-free interfaces based on real-time robust facial gesture recognition. Proceedings of the Fourth International Conference on Articulated Motion and Deformable Objects 2006, pp. 504-513.

[4] Bakić, V. and Stockman, G. (1999) Menu Selection by Facial Aspect. Proceedings of Visual Interface '99, pp. 203-209.

[5] Kawato, S. and Ohya, J. (2000) Real-time Detection of Nodding and Head-shaking by Directly Detecting and Tracking the "Between-Eyes". Proceedings of the Fourth IEEE Intl. Conference on Automatic Face and Gesture Recognition, pp. 40-45.

[6] Tierno J., Campo C. (2005) Smart Camera Phones: Limits and Applications, Pervasive Computing, vol. 4, 2, pp.84-87

[7] Wang, J., Zhai, S. and Canny, J. (2006) Camera phone based motion sensing: interaction techniques, applications and performance study. 19th Annual ACM Symposium on User Interface Software and Technology, pp. 101-1