A Comparison of Input Methods for Panning and Zooming: Touch-based and Facial Tracking

Vincent Chu
Dept. of Electrical Engineering and Computer Science
York University
Toronto, Ontario, Canada M3J 1P3
cse13261@cse.yorku.ca

ABSTRACT
A user study comparing the accuracy and speed of facial-tracking and touch-based input methods was performed. The study yielded results that show significantly worse effectiveness of facial-tracking than touch input. Facial-tracking was 6 times slower on average, upwards of more than 5 minutes to acquire a target, and had 64.8% accuracy rate compared to 97.2% for touch input. Limitations in hardware and software of facial-tracking also contributed to the slowness, averaging 58.1% of the time to complete the task was spent unable to detect the user's face or the user's intended controls.

Keywords
Android, lean, facial-tracking, zoom, panning

INTRODUCTION
Mobile devices such as smartphones and tablets commonly have smaller displays than traditional desktop computers. This constraint often results in content that is either too large to be viewed fully on-screen at any given time or has details that are too indistinguishable when scaled to fit. In order to overcome these challenges, implementers conventionally utilize two touch gestures to manipulate the content displayed. Users can drag their fingers across the screen to move the content up, down, left and right. By pinching two fingers together or moving them apart, users can increase or decrease the scaling of the content revealing more details in a particular region or more of the content as a whole while trading off finer details.

It is commonplace for mobile devices to also be equipped with cameras and a multitude of sensors. These cameras and sensors enable possibilities for novel, new solutions to these problems. This paper focuses on the front-facing camera of the devices.

Front-facing cameras are normal used for video-conferencing or taking selfies. They are rarely used as a means of input to control user interfaces. However, on occasion using the front-facing camera to map the position and orientation of the user's face to an input control may provide a useful hands-free alternative to more traditional input modes.

For instance, when a person has difficulty seeing an object, he or she naturally leans in towards the object to better inspect it [4]. Using the front-facing camera to track the user's face, a computation of the distance between the user's eyes can be used to infer the user's distance from the screen. As the user moves closer to the camera, the display can increase the magnification and conversely, as the user moves away the magnification is decreased.

This paper describes the methodology and findings of an experiment that will be conducted to compare the user performance of conventional touch-based pan and zoom implementation with the facial-tracking approach as presented earlier.

Related Work
Bulbul, Cipiloglu, and Capin [1] developed a face tracking algorithm that only uses color comparisons, rather than geometric properties of the facial features in order to avoid high computations. They found that lighting conditions and background affects that performance of their algorithm. They suggested extending their solution to take into consideration geometric features of the face in addition to color calculations.

Manresa-Yee, Varona and Perales [6] devised a method of tracking the position of the head using a webcam, in order to control a mouse pointer. Their system 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.

Joshi et al. [5] evaluated touch-free position and zooming control for viewing large imagery on mobile devices. Their experiment compared touch-based position control to various combinations of touch-free position, zoom and rate controls. They used face-tracking to compute the face's position and estimate the distance from the camera. Using these values they determined the viewing angle of the user. The *z*-axis was linearly mapped to zoom level. They found that zooming interfaces required approximately twice the amount of time compared to non-zooming interfaces. The experimenters attribute the increased time to the additional degree-of-freedom resulting in an increased precision required to acquire the target. Additionally, latency of the zoom and oscillation resulted in user frustration at times.

Harrison and Dey [4] similarly studied the lean and posture of users in correlation with the magnification of the displays. They implemented a system that calculated the user's amount of lean and increased the magnification proportionally. The calculation for the lean was computed based on the distance between the eyes when the user is in

a "nominal posture". When the user leans towards the screen, the measured distance between the eyes increases, and is used to determine the amount of magnification to apply. The participants of the study had generally positive impressions of the system. However, the study also found a degradation in performance when seeking items in the display.

Francone and Nigay [3] studied the camera-based head tracking to calculate the position of the device relative to the user's head. This allowed them to display a 3D scene according to the vantage point of the user, enhancing the depth perception and creating a virtual 3D space within (or beyond) the screen. Participants in the study found the interaction with the system "natural and immersive". However, the authors noted that there were limitations in the camera's field of view, where at certain angles the face would be out of the range of the camera.

In a study comparing facial-tracking and accelerometer-based input, Cuaresma and MacKenzie [2] used a game to evaluate user performance when controlling a cursor-like object on the screen. The evaluation found that facial-tracking was inferior to tilt-based input. The authors proposed that facial-tracking was inferior due to its heavy-resource requirements that result in a noticeable system degradation and the lack of user experience with facial-tracking compared with the conventional tilt-input.

METHOD

A user study was performed to compare two input methods for zooming and panning: touch input and facial tracking. Each method was evaluated for speed and accuracy to determine its effectiveness and efficiency. Touch input is the more common input method and will be used as a reference point for the comparison.

Participants

This experiment involved eight voluntarily participants from the local university campus. Six participants were male, two female. Ages ranged from 21 to 23 years. All participants were smartphone users. The participants were given no incentives or compensation.

Apparatus

This experiment used a *Samsung Galaxy Tab 4* tablet running *Google Android 4.4.2 Kit-Kat* operating system. See Figure 1. The device has an 8″ (203.1 mm) display with a resolution of 1280 × 800 pixels and a pixel density of 188.68 pixels/inch.

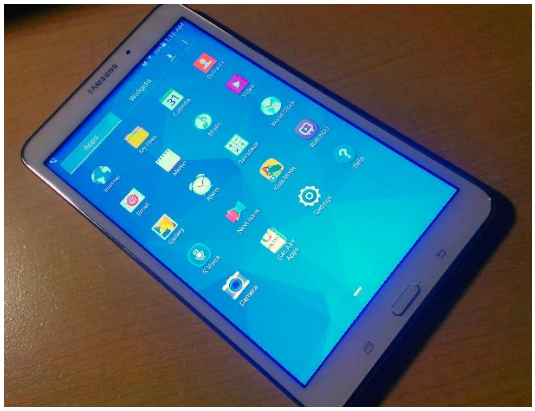


Figure 1 *Samsung Galaxy Tab 4*

The software was developed in Java using the Google Android SDK and *Qualcomm Snapdragon Facial Recognition API* [7], an API only available on certain devices running on a *Qualcomm Snapdragon* processor.

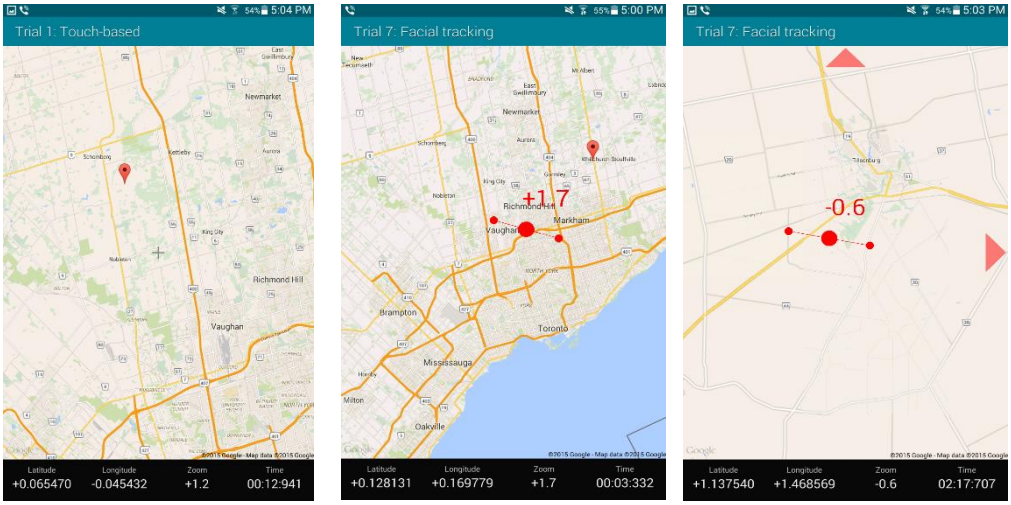
The experiment application was a map-like application developed specifically for this research. It implemented two modes of operation to accommodate for the two input methods being investigated.

Each participant was given five trials for each of the two input modes, ten trials in total. For each trial, a marker was placed randomly on the map. The marker was ensured to be within the field of view at the start of each trial. A status area at the bottom of the screen indicated the differences in latitude, longitude and zoom from the user's present view to the desired zoom. A positive difference in zoom indicates that the user must zoom in, and a negative difference indicates that the user must zoom out. Similarly a positive difference in latitude or longitude indicates that the user must move north or east respectively, while a negative difference in latitude or longitude indicates the need for a south or west movement.

The goal for the participants was to center the marker and zoom until the differences were approximately or equal to zero given some threshold of tolerance, discussed later. A measure of the time duration required to complete each trial and the difference in zoom level and coordinate distance from the marker to the focal point for every 100 milliseconds was recorded.

Each trial was allowed to run for a maximum of 5 minutes. When a trial was completed successful or timed out, an audio tone indicated the completion of the trial and users pressed a *Finish* button to continue to the next trial. If participants failed to focus the marker at the desired zoom and position within the allotted time, the trial was considered a failure attempt. No additional trials were given for failed attempts.

In touch mode, the participants used swiping to pan and two-fingered pinching to zoom in and out. A crosshair was provided as a visual indicator for where to center the marker. See Figure 2 (a). Each touch trial was initiated with the press of a *Start* button at the bottom of the screen



(a) (b) (c)

Figure 2 Experiment map-like app (a) in touch mode, (b) in facial tracking mode with eye indicator and (c) in facial tracking when target outside of the focus

followed by an audio tone. The threshold for a touch trial to be considered successful was within 10 pixel for positioning and 0.1 for the zoom level.

In facial-tracking mode, the participants held the device in a fixed position and moved their head from side to side to center the focal point and lean their head towards and away from the screen to zoom in and out. In facial tracking modes, each trial started after the device has been held still for five seconds at which time an audio tone indicated the start of the trial. An indicator for the perceived focus of the left- and right-eyes as well as the central focus between the two eyes was displayed on the screen when the camera was able to successfully detect the face. See Figure 2 (b). The zoom level difference was also shown above this indicator as a visual aid. When the marker was outside of the field of view indicators were given to shown the direction the user should move. See Figure 2 (c). The threshold tolerance for success in facial-tracking mode was within 50 pixels and 0.5 zoom levels.

Procedure

Participants will be briefed on the purpose and procedures of the experiment, and the experimenter will briefly demonstrate each of the two modes of the application. The test will occur in a quiet room with favourable lighting conditions where the participants will be seated at a table. For the facial tracking mode, participants will be instructed to keep their hands and forearms on the table to ensure that the device does not move. The device will be tilted back slightly such that the participant's face is centered in the camera's field of view. For the touch modes, participants will free to pick up the device. Participants will not be given practice trials. Following the completion of all ten trials, participants will complete a questionnaire soliciting demographic information and impressions about each of the input methods.

Design

A 2 × 5 within-subjects design will be used. There will be two independent variables: input methods (touch input and facial tracking) and the trials (1, 2, 3, 4 and 5). The dependent variables are the time of completion for each trial, and the difference of the zoom level and coordinate distance between the marker and focal point recorded over time. Participants are grouped into two groups, where each group is assigned a different ordering of the input methods in the trials. In total, there will be 80 trials for eight participants each with five trials for each of the two input methods.

RESULTS AND DISCUSSION

Of the 40 face-tracking trials, 4 trials timed out before the target was reached. All touch trials were completed successfully.

Speed

The grand mean time to complete all 80 trials was 46.7 seconds. The mean time to complete the touch-based trials was 13.4 seconds, whereas the mean time for facial-tracking was 80.0 seconds or 6 times longer.

It was observed that the mean time to complete the facial-tracking trials decreased in the latter trials, peaking at an average of 63.8 seconds in the second trial and dropping down to a 46.8 seconds in the last trial, a 17-second decrease. This may be attributed to possible learning effects. See Figure 3.

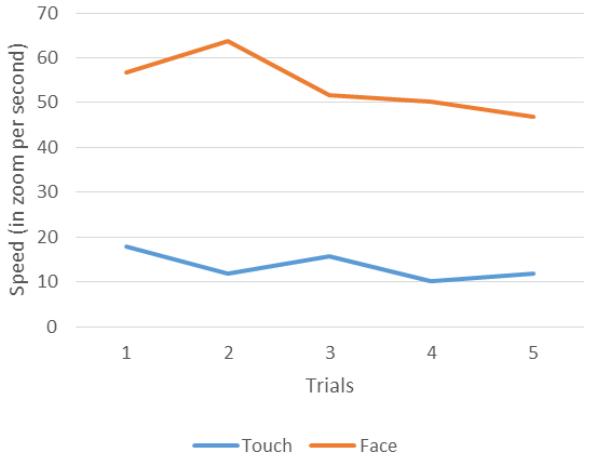


Figure 4 Average completion time per trial

Speed of navigation on the map using face-tracking generally remained unchanged throughout the experiment. No noticeably improvements in speed navigating in the latitudinal or longitudinal directions or in the zoom. Touch input was consistently faster in all three measures of speed. See Figure 4.

Accuracy

The accuracy of facial-tracking was also significantly poorer than touch based input. The average accuracy for touch-input was 97.2% whereas facial-tracking was only 64.8%. Accuracy was measured by the determining the amount of time in each trial where the user was navigating in the opposite direction of the desired target over the entire duration of the trial. Idle time or time in which the user was neither moving towards nor away from the desired target was also factored out when computing the accuracy. Accuracy did not noticeably improve over the duration of the experiment. See Figure 5.

The idle time in touch mode accounted for on average 31.9% of the time of a trial or 4.2 seconds. In facial-tracking mode, the idle time accounted 58.1% of the trial time or 46.48 seconds. The idle time can be attributed to the inability of the camera to track the face as participants leaned towards the device. The effective distance of the facial-tracking seemed to be at least 30 centimetres from the camera. Since the zoom control required leaning towards the display, participants often would get too close the device and the camera would lost tracking of their face. Limitations in the camera's field of view, where at certain angles the face would be out of the range of the camera also contributed to the idle time, and concurred with the findings of Francone and Nigay [3].

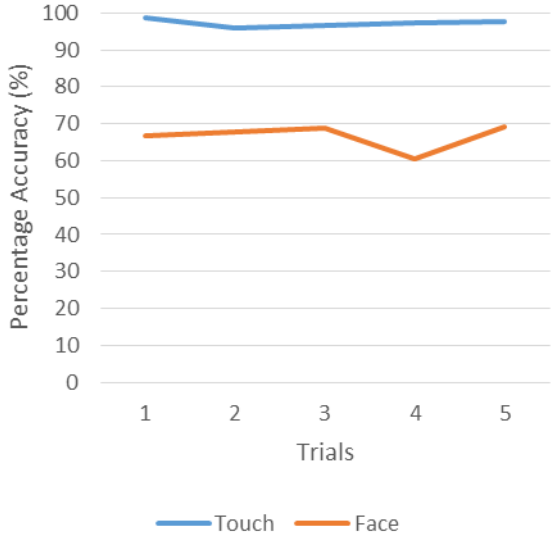


Figure 3 Percentage accuracy per trial

The observations gather for the differences in zoom and position over time for each trial reveal an explanation for the lack of accuracy. Figure 6 shows the difference in position and zoom for a touch mode trial on the left and a face-tracking trial on the right. Almost all the trials were

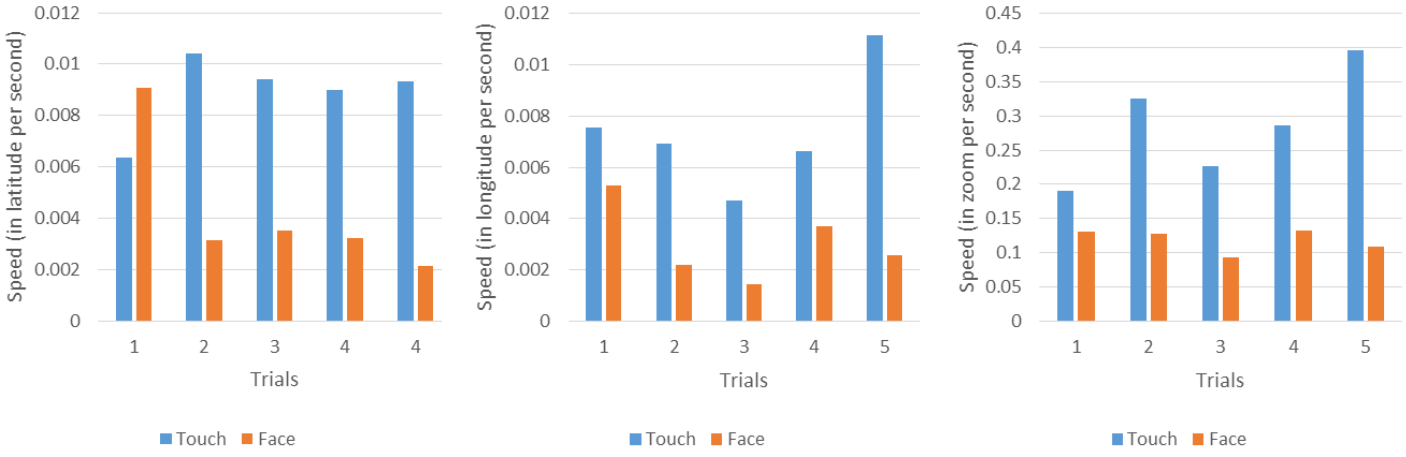


Figure 5 Speed of navigation in the (a) latitudinal direction, (b) longitudinal direction and (c) zoom per trial

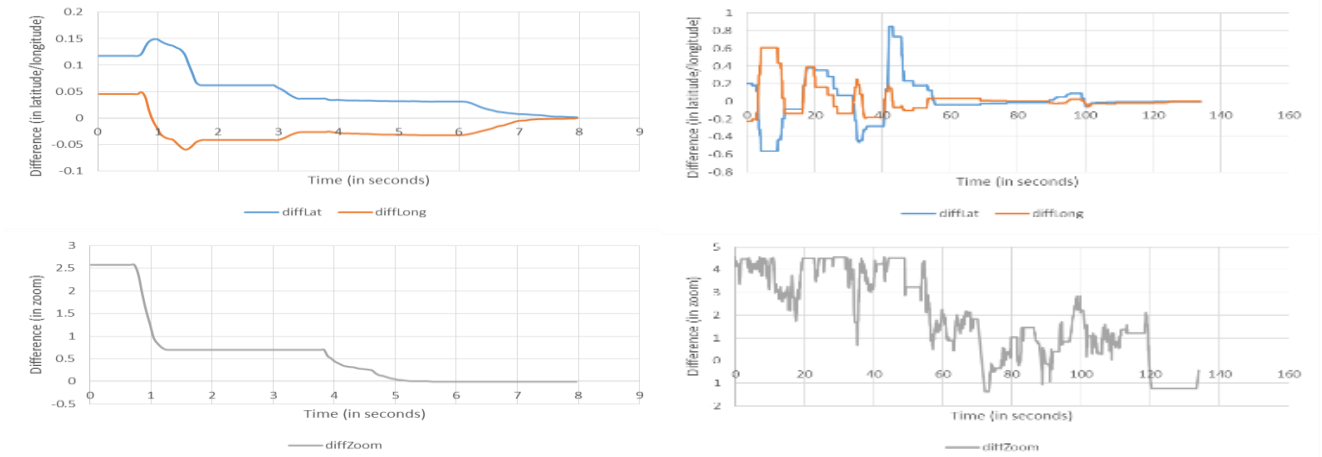


Figure 6 Differences in position and zoom over time for a trial. One trial with touch input on the left and one with face-tracking on the right.

generally similar to that shown in the figure. It reveals that large amount of oscillation, fluctuation and jitter in the face-tracking that made it significantly more difficult to precisely control and direct the focus or zoom in a particular direction. These suggest that some form of smoothing or exponential weighted moving average needs to add to the implementation to dampen the amount of fluctuation and thus improve the control.

Based on the questionnaire, participant feedback was generally negative towards the facial-tracking input. Overwhelmingly, participants preferred touch-based input over facial tracking. Participants were frustrated by the difficulty of the control. One participants described his experience as:

> *Really hard to pan across the map with Facial Tracking. I wasn't sure whether the pan was based on the angle of my head or the positioning. It was also very noisy. My map would constantly zoom in and out at a very rapid rate with Facial Tracking. The zooming feature in Facial Tracking is pretty clever and was something I really liked. However, overall, the experience was frustrating, especially when trying to move the cursor to the target.*

Some found that they had to remove their glasses, because the camera had difficulty tracking their faces when wearing their glasses.

Others felt that the tracking may have favoured one direction slightly over the other, because the camera is slightly off-centre on the device. It was also noted that vertical panning, up and down, was particular difficult perhaps because the camera's location at the top of the device made it difficult for the camera to detect the upward or downward movement of the head.

However, one participant pointed out that face-tracking would be useful for hands-free applications if the control and tracking were better.

CONCLUSION

An experiment was conducted comparing touch and facial tracking input methods for an application controlling and navigating a map. The objective of this work was to better understand the effectiveness and efficiency of facial-tracking. The results of this study indicate that, limitations in facial-tracking at wide-angles and in close proximity to the device, noisy and imprecise data from the camera and face detection mechanism may significantly hampered the accuracy and usability of facial-tracking for tasks requiring high precise and control.

Improvements to algorithms and implementation for detecting the face gesture and movement may yield better more promising results. Additionally, improvements to hardware and software for facial-tracking will likely result in high accuracy and precise, thus better usability.

REFERENCES

1. Bulbul, A., Cipiloglu, Z., & Capin, T. (2009, September). A face tracking algorithm for user interaction in mobile devices. In *CyberWorlds, 2009. CW'09. International Conference* (385-390). New York: IEEE.
2. Cuaresma, J., & MacKenzie, I. S. (2014). A comparison between tilt-input and facial tracking as input methods for mobile games. *Proceedings of the 6th IEEE Consumer Electronics Society Games, Entertainment, Media Conference - IEEE-GEM 2014*, 70-76. New York: IEEE. doi: 10.1109/GEM.2014.7048080
3. Francone, J., & Nigay, L. (2011, October). Using the user's point of view for interaction on mobile devices. In *23rd French Speaking Conference on Human-Computer Interaction* (4:1–4:8). New York: ACM.
4. Harrison, C., & Dey, A. K. (2008, April). Lean and zoom: Proximity-aware user interface and content magnification. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (507-510). New York: ACM.
5. Joshi, N., Kar, A., & Cohen, M. (2012, May). Looking at you: Fused gyro and face tracking for viewing large imagery on mobile devices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2211-2220). New York: ACM.

6. Manresa-Yee, C., Varona, J., & Perales, F. J. (2006). Towards hands-free interfaces based on real-time robust facial gesture recognition. In *Articulated Motion and Deformable Objects* (504-513). Springer Berlin Heidelberg.
7. Qualcomm. Snapdragon SDK for Android, https://developer.qualcomm.com/mobile-development/add-advancedfeatures/snapdragon-sdk-android, 2014.