Attention in Mobile Interactions: Gaze Recovery for Large Scale Studies

Abstract
Understanding human attention in mobile interaction is a relevant part of human computer interaction, indicating focus of task, emotion and communication. Lack of large scale studies enabling statistically significant results is due to high costs of manual penetration in eye tracking analysis. With high quality wearable cameras for eye-tracking and Google glasses, video analysis for visual attention analysis will become ubiquitous for automated large scale annotation. We describe for the first time precise gaze estimation on mobile displays and surrounding, its performance and without markers. We demonstrate accurate POR (point of regard) recovery on the mobile device and enable heat mapping of visual tasks. In a benchmark test we achieve a mean accuracy in the POR localization on the display by \(\approx 1.5\) mm, and the method is very robust to illumination changes. We conclude from these results that this system may open new avenues in eye tracking research for behavior analysis in mobile applications.

Author Keywords
Human attention; gaze recovery; mobile interaction heat maps.

ACM Classification Keywords
H.1.2 User/Machine Systems - Human information processing.
**Introduction**

The evaluation of interaction designs in mobile applications requires in general the investigation of human attention by means of eye movements, in analogy to usability analysis for the iterative development of desktop technologies [9], such as for websites. Mobile computing technology allows having electronic devices available anywhere, anytime. Designers of mobile applications like palmtop computers and smartphones have to face unique challenges, because location and environment are usually less predictable than in desktop applications [2]. These devices have the common problem of small visual displays and limited input techniques, performance is often substantially worse than in the desktop context [3]. Multitasking and support for task interruption are of high relevance, since in a mobile context the frequency of distracting events is much higher than for a desktop application and tasks with interruptions take longer to complete on a mobile device than with a desktop application [4]. Special interest is therefore on increased competition with regard to attracting the users' attention and on interaction as a non-primary task in a certain context [5]. However, current technologies for the mapping of point-of-regards (PORs) to mobile displays do not enable natural interaction with the mobile device: users of the mobile device are either conditioned to act with tightly mounted displays and hence prevented from performing in the typical mobile user’s environment or they are distracted by markers in the view (Figure 3).

We propose a novel and successful approach that enables natural interaction with the mobile device and its application (Figure 1). We use eye tracking glasses (ETG) to capture the user’s fixations on its environment and apply computer vision for highly accurate recovery of human gaze on the mobile display. In a benchmark study we achieve a mean accuracy of POR localization on the display of $\approx 1.5 \pm 0.9$mm. Furthermore, we demonstrate that the approach enables the analysis of large user studies in contrast to using state-of-the-art annotation tools. In contrast to state-of-the-art methodologies, our approach enables natural interactions, from these one can draw more valid conclusions and better interaction designs, in the frame of usability research in both outdoors and indoors studies. We conclude that this will open new avenues for understanding mobile interaction.

**Related Work**

Eye tracking devices are frequently used to analyze mobile interaction. In indoor studies, mapping of PORs to mobile displays is usually performed by measuring with tightly mounted phones (Figure 3a,b). For example, [6] screw the mobile device under a table, and enable evaluation of mobile interaction only when sitting (Figure 3a). [7] investigated and proved the significance of eye tracking in mobile applications.
usability testing, using a head mounted eye tracker. [8] developed an eye tracking pilot test on validating a smartphone based pedestrian navigation system, describing the relationship between reality and navigation instructions, using a standard annotation toolbox. [10,16] presented static display localization in eye tracking tasks. However, their approach is specific, localization of mobile devices involves more degrees of freedom. [11] referred to the concept of mobile device detection but lacked quantitative information on experimental results. [12] mentioned Dikablis based marker tracking (Figure 3d) which principally distracts the user’s attention.

Alternative approaches use the smartphone camera directly with its rear view on the user to estimate the eye gaze on the mobile phone [9]. This method achieves accuracies only in centimeters range, is vulnerable to illumination conditions, and cannot provide information about the surrounding that contains relevant information [15].

Smartphone Eye Tracking Toolbox
We developed a software toolbox package, i.e., the ‘smartphone eye tracking toolbox’ (SMET) that will assist usability analysis towards fully automated analysis of human attention processes in user studies. Figure 2 depicts a schematic sketch of the information flow in the toolbox: data capture is applied with the ETG and the screencast recorder; synchronization and image analysis provide a data stream with synchronized POR and smartphone display events. Finally, geometric transformation and heat mapping provide the basis for attention analysis.

The Graphical User Interface (GUI) of the SMET application (Figure 4) requires three kinds of input data: the scene camera based video, the smartphone video, and the captured POR data. Firstly, points that lie on image areas of the colored phone case are manually selected for initialization, but further then the color thresholds which are required by the tracking algorithm are automatically determined. The two videos are then synchronized in a semi-automated manner and the automated smartphone tracking process is applied. Once the smartphone has been detected in the image,
To enable robust tracking the smartphone is used with a color intensive case (Figure 5a) in order to identify the device well within the scene video. Through the detection of the colored frame lines and the corner points the image segment of the smartphone can be identified and visually tracked. The tracking algorithm makes adaptive use of color threshold values, which are manually selected in a first frame, and then automatically adjusted, so that color based segmentation is performed in appropriate color spaces (HSV, YCbCr) (Figure 5b). After noise removal by morphological filter operations, the remaining image regions are reduced to skeletons of the regions by a thinning algorithm. The resulting lines are then extracted based on Hough Transform [13]; noisy lines are removed and intersection points from the remaining lines are robustly found. Corner points of the smartphone are detected, are validated by previous tracking results and geometrical constrains of the smartphone area. The tracking algorithm is based on a linear prediction which performs well for a random movement of a single object [14]. Positions of missing corner points are estimated based on the movement of detected points and position predictions from tracking, resulting in localizations even if hands occlude parts of the area of interest or only small parts of the phone are observed within the image. Through the continuous adaptation of the color thresholds based on actual tracking results, the detection of the smartphone is robust in challenging illumination conditions (Figure 5c,d).

**Proof of Concept - Navigation Study**

The study targeted a proof of concept, investigating which level of accuracy in POR mapping can be achieved. Typical users of mobile applications were involved several times in an outdoor navigation task, including map based and augmented reality (AR) guidance, to find locations in a local hospital. To acquire video and eye-tracking data we used SMI Eye Tracking Glasses, with 30 Hz sampling rate of gaze and a 1280 x 960 pixel resolution scene camera. Figure 6a,b depict the camera view from the Eye Tracking Glasses, overlaid with the bounding box of the smartphone localization; Figure 6c presents a heat map from PORs mapped onto the display of the mobile device.

**Precision of Smartphone Localization** is crucial for the usability of automated annotation in ETG video frames. Table 1 presents most relevant evaluation results from 3 different test runs using the mobile service, under different weather conditions (2x sunny, 1x cloudy weather). A substantial subset of the number of video frames (‘detections’) represents the interaction of the user with the mobile phone, however, the detection method was applied to all video frames. The precision rate was sufficiently high to enable stable, cost-efficient
processing. The calibration error was estimated ≈3mm. The human error in the ground truth annotation was ≈2mm. The localization error fitted nicely to be ≈1.5±0.9 mm, in reference to index to icon buttons of, e.g., size 7x7 mm in standard Samsung displays.

**Semantic Annotation.** We are interested to measure the distribution of attention on regions of interest (ROIs) of the smartphone display and the surrounding in a navigation application (Figure 6a): “AR” depicts the augmented reality view, i.e., the orientation arrow overlaid on the live camera view; “MAP” refers to the map view on the environment and a red line representing the planned trajectory of a user; “BG” the information panel of the smartphone display; “SUR” the surrounding that is viewed beyond the smartphone. We analyzed the ETG video as described before and determined the PORs’ recovery on the device’s display. We then associated the POR to ROIs’ segments as depicted in Figure 6a.

**Performance Analysis.** From the statistics of 24,000 video frames from 2 users, among those 12,884 frames with automated smartphone detections, we deduce the fact that 62% of PORs were localized within “AR”, 9% on “MAP”, 9% on “BG” and 20% on “SUR”. The avg. dwell time on “AR” has been calculated to be 471 (±540) ms, for “MAP” 128 (±115) ms, for “BG” 88 (±96) ms and for “SUR” 176 (±198) ms. All quantities were extracted and calculated in a fully automated manner, and will be input for more profound analysis. Automated annotation of ROIs enables to perform large user studies without substantial efforts, which has not been applicable so far. In a user study with 100 persons, and a study with similar quantities as above (100 times 12884 frames to annotate), counting 3 sec./frame using standard (such as BeGaze) interactive annotation software, assuming 8 hours daily work –one operator would need 134 days (i.e., 26.8 weeks) to annotate - this in contrast to the presented automated annotation process with no intervention needs.
Discussion and Conclusion
We presented a novel and robust approach for marker free tracking of smartphones in the head video of eye tracking glasses. The localization precision is sufficient to enable analysis of button choices in the display. This enables natural interaction, especially in large user studies, using automated instead manual annotation, and carries a huge potential to be extended with computer vision methodology. Limitations are in extreme illumination, such as direct sunlight and dawn.

Potential applications are mobile interaction studies both outdoors (navigation, image/video based user interfaces, urban points-of-interest, mixed reality games, cell based interaction with ticket machines or urban displays) and indoors (multiple displays, mobile-TV interaction). Future work will focus on the consideration of attended semantics in the surrounding (detection of persons, objects, landmarks), and on the fusion with information from psychophysiological and activity sensors.

Acknowledgements
This work has been partly funded by FP7(2007-2013) (grant n°288587, MASELT0V) and by the Austrian Research Prom. Agency (grant n°832045) FACTS.

References