A Framework for Attention-Based Implicit Interaction on Mobile Screens

Peter Kiefer
ETH Zurich
Institute of Cartography and Geoinformation
CH-8093 Zurich, Switzerland
pekiefer@ethz.ch

Ioannis Giannopoulos
ETH Zurich
Institute of Cartography and Geoinformation
CH-8093 Zurich, Switzerland
igiannopoulos@ethz.ch

Abstract
We propose to keep track of the user’s attention during interaction with a small display screen, and use that attention history for later interface adaptation. A framework for attention-based implicit interaction is described that consists of attention measuring, mapping, logging, aggregation, interpretation, and interface adaptation. The framework is exemplified by an interaction method described in previous work: the GeoGazemarks approach.

Author Keywords
Gaze-Based Interaction; Visual Attention; Information Overload

ACM Classification Keywords
H.5.2 [Information interfaces and presentation]: User Interfaces

Motivation
Although wearable devices, such as smartwatches and head-mounted displays [24], are on the rise, the smartphone with its ‘one device fits all’ property is still the most prevalent mobile device these days. The high-resolution displays, high processing power, strong multimedia capabilities, good battery capacities, and the integration of numerous sensors turn smart phones into everyday companions.
assisting with a plethora of mobile tasks, such as messaging, navigation, or gaming.

The diversity of tasks smart phones support, however, makes the interaction and interface design challenging. Complex, diverse, and frequently changing content needs to be accommodated on the very limited screen. An inadequate interaction and interface design may easily lead to information overload, even more since mobile tasks are often performed highly parallel and during interaction with the real world. Not surprisingly, cognitive load has been recognized as an important criterion for usability of mobile interaction [13].

Different strategies have been proposed for coping with information overload during the interaction with mobile devices, the most obvious one probably being the development of screen design guidelines for mobile devices [21]. Another strategy consists in adapting to the user’s cognitive resource limitations based on context inference [2]. The Google Glass interface has explicitly featured short micro-interactions [24], while other wearable interfaces completely abandon the screen in favor for non-visual output, such as vibration [14, 11, 4]. Non-visual output modes, however, do not fit the general purpose requirement, and are hardly suited for content-rich interfaces, such as PDF reading, map exploration, or general web surfing.

In this paper, we propose to track the user’s attention during interaction and use that attention history for later interface adaptation in order to avoid information overload. We describe a general framework for attention-based implicit interaction that manages attention with the following steps: attention measuring, mapping, logging, aggregation, interpretation, and interface adaptation. Note that, although most of this paper focuses on visual attention, we see attention as a general concept in the line of attentive interfaces [25].

In the following section, we introduce the framework which is then exemplified by the GeoGazemarks approach in the Example section. We conclude the paper with an Outlook. Related Work is discussed throughout the paper whenever adequate.

Framework
Figure 1 displays the processing steps of our framework which are explained in the following.

Measuring Attention
As a first step, the user’s attention on the interface during normal interaction is measured. Different means can be used: attention data of high spatial and temporal resolution can be acquired by eye tracking, either through the front-facing smart phone camera [20], or through a head-mounted eye tracker [10, 22].

Eye detection is a more basic means of measuring visual attention, already implemented for current smart phones on the market\(^1\). Here, the presence of the eye in the front-facing camera is detected, but not the gaze position on the screen.

Next to measuring the visual attention directly, also non-visual interaction logs can serve as a proxy for visual attention. It is known that mouse and gaze positions show a high correlation in Desktop interaction [7]. Analogous, on a smart phone we can take the position of touch events as indication for visual attention. While eye tracking returns a stream of gaze points in screen coordinates (high resolution), the eye detection and touch events return (potentially

\(^1\)e.g., the Samsung Galaxy S4
fuzzy) regions on the screen (low resolution) as a basic unit for attention.

Note that the framework is flexible towards measuring attention on multiple screens [17]. The basic recorded unit of attention, such as a gaze position in pixel coordinates, would then be stored including information about the device it has been recorded on.

**Mapping Attention to Content**

Next, the screen coordinates must be mapped to the content that was visible on the screen at that moment. Depending on the resolution measured (point vs. fuzzy region), the density of the screen content, and the accuracy of the sensor, this mapping is not a trivial task. For cartographic maps, which contain specifically dense content, this problem has been investigated in [18] under the title Gaze Map Matching. It has been argued that the temporal sequence of attention, combined with transition probabilities between content elements, can help for attention mapping. The dotted arrow in Fig. 1 indicates that using the log can help for the attention mapping

**Logging and Aggregation**

Having identified the content elements the user has attended to, these are logged to an attention history. Because the attention data come at high temporal resolution (e.g., eye tracking at 30 Hz), and because the memory of mobile devices is limited, a forgetting and/or aggregation of the attention history may become necessary. Deciding what to forget and/or aggregate depends on what we would like to do with the attention data later, which is described in the following section.

**Interpretation**

Attention can, first and foremost, be seen as an indicator for interest in a part of the interface. This interest is driven by both, bottom-up and top-down processes [5]. The first are mainly determined by (visual) saliency while the second can be described as the user's activities, intentions or plans, depending on the level of abstraction. In order to provide an adequate adaptation it becomes necessary to determine which of these processes caused the visual attention, i.e., to interpret the user's attention log. This is an ongoing active research field in HCI [19, 6, 1, 3], and the algorithms used here are highly dependend on the interface type and tasks performed.

**Interface Adaptation**

By knowing which interface elements the user has paid attention to, and even more if we were further able to infer activities or intentions, intelligent interface adaptation can be designed. A number of papers in HCI investigate adaptive user interfaces (e.g., [16]). There is a clear agreement that adaptations must be designed carefully to avoid users feeling patronized, and to avoid causing information overload. Instead, the existing overload in the interface before the adaptation, such as, too many notification icons, should be reduced through the adaptation. The expected cost of the adaptation must be less than then the expected cost of not adapting [15]. Formal adaptation models are required which integrate a number of context parameters into the adaptation [23]. Moreover, adaptation will most likely change the user's attention, which will again be measured in the first level of the framework.

**Example**

GeoGazemarks is a gaze-based implicit interaction concept that provides the user's visual attention history on a mobile map as visual clue to facilitate orientation [10]. The proposed framework captures and divides the complexity among the components relevant for the GeoGazemarks concept.
Measuring Attention: The visual attention in the GeoGaze-
marks concepts is recorded using an eye tracking device,
returning gaze points in screen coordinates.

Mapping Attention to Content: These coordinates are
transformed to geospatial coordinates (latitude, longitude)
by mapping the visual attention to the geospatial content
displayed at the recorded screen coordinates.

Logging and Aggregation: Next, the geospatial coordi-
nates are aggregated to fixations and logged to the atten-
tion history. Fixations occur when the eyes remain relative
still over a period of time.

Interpretation: A clustering algorithm was applied in or-
der to aggregate fixations in spatial proximity. The size of
the cluster is based on the amount and duration of the fix-
atations. The more and longer fixations in the spatial proxim-
ity, the higher the interest for that coordinate. In GeoGaze-
marks, attention for a spatial coordinate on the map is in-
terpreted as having interest for content at that coordinate.
That means, it is assumed that visual attention includes
perception, and that perception includes interest. Higher-
level cognitive states, such as activities or intentions, are
not the focus of GeoGazemarks.

Interface Adaptation: Finally, once the user of GeoGaze-
marks performs a zoom-out interaction, the map adapts,
displaying the visual attention clusters on the map (see fig-
ure 2). This should help the user to orient based on the
map areas she was previously spending her visual attention
to.

Outlook
As future work, the framework will be applied to interfaces
in different domains, and different methods will be tested
on the individual processing levels. Open questions to be
investigated include: how reliant are the upper processing
levels on a precise attention measuring sensor (or can we
achieve good results also with other means of measuring
attention)? What are good models and algorithms for inter-
preting attention in terms of activities/intentions? Here, we
can rely on existing models from research on visual atten-
tion and cognitive models for the respective domain (e.g.,
[9] for airline pilots). How to design a good adaptation that
is really perceived as helping, instead of annoying and pa-
tronizing?

While applying the framework to multiple devices is straight-
forward for the lower levels – attention measuring, mapping
and logging – there is some challenge in the upper levels of
the hierarchy: how can we aggregate and interpret attention
during the interaction with multiple devices?

Extending the framework to attention-based interaction
of multiple users is another interesting area for future re-
search. One user could share her visual attention with oth-
ers, e.g., for teaching purposes [8], or multiple users could
share their attention to solve a task collaboratively [12].

REFERENCES
1. Thomas Bader, Matthias Vogelgesang, and Edmund
Klaus. 2009. Multimodal integration of natural gaze
behavior for intention recognition during object
manipulation. In Proceedings of the 2009 international

Proceedings of the 7th international conference on
Intelligent user interfaces. ACM, 15–22.

3. Roman Bednarik, Hana Vrzakova, and Michal Hradis.
2012. What do you want to do next: a novel approach
for intent prediction in gaze-based interaction. In


