Experimental Evaluation of User Interfaces for Visual Indoor Navigation

Andreas Möller [1], **Matthias Kranz** [2], **Stefan Diewald** [1], **Luis Roalter** [1], **Robert Huitl** [1], **Tobias Stockinger** [2], **Marion Koelle** [2], **Patrick Lindemann** [2]
[1] Technische Universität München, Arcisstraße 21, 80333 Munich, Germany
[2] Universität Passau, Innstraße 43, 94032 Passau, Germany
andreas.moeller@tum.de, matthias.kranz@uni-passau.de, {stefan.diewald, roalter, huitl}@tum.de, {tobias.stockinger, marion.koelle, patrick.lindemann}@uni-passau.de

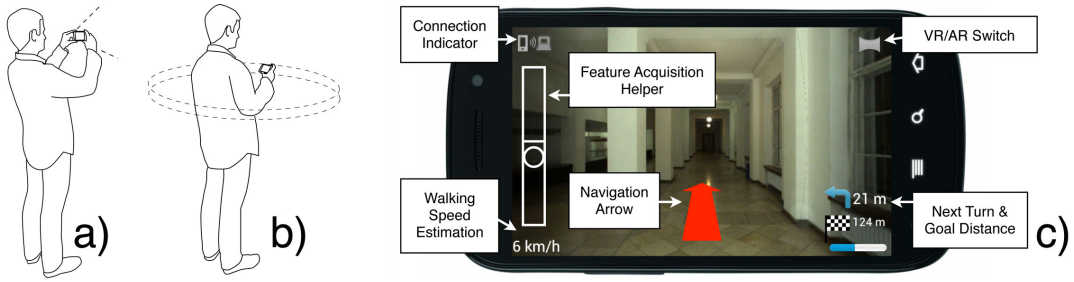


Figure 1. We present and evaluate a novel user interface for indoor navigation, incorporating two modes. In augmented reality (AR) mode, navigation instructions are shown as an overlay over the live camera image and the phone is held as depicted in Picture a). In virtual reality (VR) mode, a correctly oriented 360° panorama image is shown when holding the phone as in Picture b). The interface particularly addresses the *vision-based* localization method by including special UI elements that support the acquisition of "good" query images. Screenshot c) shows a prototype incorporating the presented VR user interface.

ABSTRACT

Mobile location recognition by capturing images of the environment (visual localization) is a promising technique for indoor navigation in arbitrary surroundings. However, it has barely been investigated so far how the user interface (UI) can cope with the challenges of the vision-based localization technique, such as varying quality of the query images. We implemented a novel UI for visual localization, consisting of Virtual Reality (VR) and Augmented Reality (AR) views that actively communicate and ensure localization accuracy. If necessary, the system encourages the user to point the smartphone at distinctive regions to improve localization quality. We evaluated the UI in an experimental navigation task with a prototype, informed by initial evaluation results using design mockups. We found that VR can contribute to efficient and effective indoor navigation even at unreliable location and orientation accuracy. We discuss identified challenges and share lessons learned as recommendations for future work.

Author Keywords

Virtual Reality; Augmented Reality; Indoor Navigation; Visual Localization; Mobile Interaction.

CHI 2014, April 26 - May 01 2014, Toronto, ON, Canada

ACM 978-1-4503-2473-1/14/04...$15.00.
http://dx.doi.org/10.1145/2556288.2557003

ACM Classification Keywords

H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous

General Terms

Human Factors; Design; Measurement.

INTRODUCTION AND MOTIVATION

Imagine you are at the airport, at a mall or in a museum and your smartphone gives you directions to your departure gate, that hot new fashion store, or the famous Dalí painting you want to visit. While mobile navigation is omnipresent outdoors, it is not inside buildings. Reliable *indoor* navigation is still a "hot topic". While researchers are still looking for the optimal localization method, appropriate novel user interfaces for these scenarios have to be investigated.

An analysis of existing indoor localization techniques (which we discuss in the Background section), shows *visual localization* to have multiple advantages to concurrent methods for indoor usage. Using computer vision, this technique captures and matches images of the environment with previously recorded reference images of known locations. However, we found that existing user interfaces (UIs) for pedestrian navigation are not appropriate for that (relatively new) technique, since they do not particularly address the characteristics of *visual* localization. As the device uses the camera to orientate and position itself, visual localization works similar to human orientation and wayfinding (e.g., based on landmarks and salient objects). The technical implications of this localization method should be reflected in the user interface to

the advantage of both the UI and the underlying localization mechanism. In that way, the UI can benefit from the strengths of visual localization, and the (perceived and actual) localization quality can be improved through UI elements and the user interactions with them.

In this paper, we implemented a novel UI concept for an indoor navigation system which is specially fitted to visual localization, and we provide a first evaluation of this UI, based on experimental simulation, compared against the conventional augmented reality (AR) technique. Moreover, our work represents an example for interweaving the UI and the underlying localization technique of an indoor navigation system to the advance of both, arguing that localization and UI should be treated jointly for being most effective.

The structure of this paper is as follows: We begin with presenting related work, where we focus on existing user interfaces for navigation systems and on the particularities of visual localization. Subsequently, we describe the implemented interface concept and UI elements. We introduce the conducted study and discuss our experimental findings in a comprehensive way. We finally share lessons learned in order to inform the design of future visual indoor navigation systems.

BACKGROUND AND RELATED WORK

Discussion of Visual Localization and Other Techniques

First, we discriminate visual localization against other techniques to localize a device inside buildings, and outline the advantages of visual localization. By visual localization, we understand the usage of computer vision to correlate query and reference images by characteristic properties (so-called feature matching). We explicitly do not comprise marker-based approaches (e.g. [17]) by this term.

Feature matching has the advantage that image can be captured with the device's camera at any location, which then serve as query images. No augmentation of the infrastructure with fiducial markers (i.e., points of reference) is necessary any more. Other infrastructure-based approaches, e.g. WLAN fingerprinting [9], require dense coverage of access points. This coverage is in many buildings not available, and it costs money and effort to establish.

Furthermore, a common camera-equipped smartphone is sufficient for visual localization. By contrast, approaches based on signal metrics, such as angle (AOA) or time of arrival (TOA), require special hardware, such as directional antennas or ultra-accurate timers [10]. Signal-*strength*-based measurements are feasible with common hardware, but the location often can only be determined within a radius of 1 m or more [9], even in laboratory tests. In the real world, where the typical density of access points is mostly lower, expected localization accuracies are likely to be inferior to those in controlled experiments. Fiducial markers provide exact localization only at locations where such markers are placed. Apart from these "key locations", the position needs to be estimated with relative positioning techniques, such as dead reckoning.

With a database of sufficiently densely recorded reference images, visual localization can be performed at almost any location and on centimeter level [19]. Based on the position of feature points, even the pose (i.e., the viewing angle) can be detected, which is usually not the case with other approaches. However, the image database must be built up once (by mapping the environment) and updated regularly when buildings and objects therein significantly change.

There are several concrete implementations of camera-based location recognition systems [4, 17, 19]. Hile and Borriello correlated a floor plan or a previously captured reference image to estimate the device's pose and to calculate an information overlay [4]. However, the system only works for static images. Mulloni et al. [17] relocalized a phone by recognizing visual markers and displayed the new location on a map. Schroth et al. [19] presented localization approaches through feature-based image matching, but without focusing on a specific user interface.

UIs for Pedestrian Navigation Systems

After having motivated the visual localization technique, we provide an overview of pedestrian navigation user interfaces. Kray et al. [7] use sketches, maps or pre-rendered 3D views according to the quality of the location estimate or device capabilities. Butz et al. [3] propose a simple directional arrow when localization accuracy is high, and suggest to use a 2D map and more additional cues in case of decreasing accuracy.

Besides rendered graphics, augmented reality (AR) is considered an intuitive way to visualize a location and has been used in manifold ways [1]. In AR, virtual elements are superimposed over a live camera view, so that users do not need to translate between the virtual representation and the real world [18]. Liu et al. [11] presented a smartphone-based indoor navigation system with superimposed directional arrows and textual navigation instructions. They found that adapting the interface to the users' preferences is particularly important. AR can also convey information beyond navigation instructions. Narzt et al. visualized elements in a car navigation system that are invisible in the real world, such as highway exits that are hidden behind a truck [18]. Similar ideas could be adapted for pedestrian navigation. Miyashita et al. [12] used AR for a museum guidance system. Augmentations enhanced exhibits with additional information. Visitors were guided along a predefined route through the museum when they searched with their phone for the next AR object. An AR system which employed floor-projected arrows as way directions was evaluated better in terms of usability than a map-based system [20].

Researchers also recognized the value of landmarks for orientation (particularly for outdoor pedestrian navigation). Hile et al. [5] created route descriptions that include geo-tagged images as additional cues besides textual instructions. A similar approach is presented by Beeharee and Steed [2]. Miyazaki et al. [13] use panoramic images to provide additional information on surrounding buildings in an AR-like manner, but the location must have been determined before with GPS or manually on a map. Mulloni et al. suggested different perspectives for displaying panoramas [16]. They found that by top-down and bird's eye views of a panorama, users were quicker to locate objects in the environment than using a frontal view.

UI Challenges with Visual Localization

When going beyond key-point localization, as used in many prior systems [4, 17], towards continuous guidance, as known from outdoor navigation, new challenges emerge. In that case, the visual system must capture query images on a regular basis. The challenge is here that quality and distinctiveness of the query images impact the location estimate. Ideal query images are crisp and show characteristic areas with lots of visual information. However, the camera-visible scene can happen to be blurred due to motion of the device, or can be not sufficiently unique (e.g., plain corridors often look very similar). The pose of the device plays a role as well – the typical orientation when holding a phone (about 45° downwards) entails that rather the floor is visible to the camera, but not corridors and rooms and the objects therein (which would be good candidates for reference images).

USER INTERFACE CONCEPT

Our implementation is based on the UI concept we have presented in earlier work [14]. It includes a panorama-based view as a complement to Augmented Reality and proposes different visualizations for motivating users to record "good" query images. The concept is dedicated to visual localization and conceived as "live interface" during the entire navigation process, i.e., it is used not only for (re-)localization at a certain point on the route, but allows continuous guidance. Additionally, it is prepared for the use of context-based services by interacting with objects in the environment.

Augmented and Virtual Reality

The interface consists of two modes for continuous guidance: *Augmented Reality* (AR) and *Virtual Reality* (VR). *Augmented Reality* enhances the video seen by the smartphone's camera by superimposing navigation information, such as a directional arrow. Since users need to hold the phone upright for visual localization (so that the camera can see the environment), this seems a reasonable interface for a visual localization system. Users hold the phone as illustrated in Fig. 1a) and "look through" the phone in order to see the augmentation directly on their way. However, this pose might be inconvenient for long-term or frequent use (e.g. in unknown environments).

The alternative mode is *Virtual Reality*, which can be employed also when the phone is carried in a lower position. It displays pre-recorded images of the environment (downloaded from a server) that are arranged to a 360° panorama on the mobile device. Navigation arrows are directly rendered into the panorama, so that their orientation is fixed in relation to the virtual 360° view. This is expected to have several advantages. First, the device can be held in a more natural and comfortable way, as illustrated in Fig. 1b), since no alignment of the overlays with live video is required. Second, we expect that the "hard-embedded" navigation arrows provide a more reliable navigation, as they also show the correct way in the panorama if the orientation estimate is not perfectly accurate. Furthermore, in case no reliable localization estimate is possible, the frequency in which panoramas are updated can be lowered. Hence, we expect VR to be more robust than the more conventional AR view.

Specially Designed UI Elements For Visual Localization

Dedicated UI elements for the visual localization method shall help to improve localization accuracy. We assume that a visual localization system can determine its location better when the device is held upright, as if taking a photo. In that pose, the camera points at regions in eye height, such as exhibits, posters or signs, which are potentially more discriminative motives for feature matching than if the camera were pointed downwards. Consequently, if localization certainty has reached a lower bound (this value could e.g. be determined by the localization system or by user preferences), an indicator prompts the user to actively point at regions containing more visual features. The user is thereby asked to bring the phone from a pose as in Fig. 1a) to one as in Fig. 1b). Four indicator types fulfilling that purpose are proposed:

- *Text Hint*: A notification to raise up the phone appears until the pose is such that sufficient features are visible.
- *Blur*: The live video view turns blurry; the closer the device is moved to a feature-rich position, the sharper the image becomes. This metaphor is inspired by an autofocus camera, motivating the user to find the "best" shot.
- *Color Scale*: A colored scale, ranging from red to green, indicates the quality of the current scene for relocalization. The user should steer the indicator into the green area.
- *Spirit Level*: The user must align the bubble of a spirit level in the middle of the scale to find the ideal inclination, so that the camera points at a feature-rich region.

Involving the user to help the system improve its position accuracy has already been used in other contexts for self-localization. For example, Kray et al. [8] asked users whether they can see certain landmarks from their point of view in order to perform semantic reasoning about their position.

Another way to draw the users' attention to feature-rich objects is to explicitly highlight them in the viewport. Object highlighting is motivated by an additional benefit for the user: context-based services. Like this, stores in a mall, individual shop windows, or even doors and doorplates can become points of interaction. However, a convenient side effect is that typical "interaction areas" like posters or signs often have a very characteristic appearance and therefore also serve well as reference images for localization (we though have to note that they are also subject to frequent change, see Discussion section). If they attract the user's attention and are focused with the smartphone's camera, they implicitly help improve the system's certainty of the location estimate.

ANALYSIS OF CONCEPT EVALUATION

A non-functional mock-up of the proposed UI concept has been evaluated in an online survey in prior work [14]. We summarize and analyze the results of this evaluation as a starting point for our investigation of the concept's effectiveness in practice. Extending on this prior work, we developed a working system which was evaluated in a laboratory study.

Research Questions and Results Summary

Perceived Accuracy and User Preference for AR/VR

In order to have subjects estimate how they perceive accuracy in the AR and VR modes, videos of a pre-recorded

sample navigation task were played back alongside with the simulated output of the system. The video demonstrations contained the simulated field of vision (i.e., the "reality") in the upper part, and the simulated visualization on the smartphone in the lower part. In four videos for each mode, different types of errors (position, orientation, both error types together) were induced to the system's location estimate, so that the simulated output changed accordingly. Subjects rated the perceived accuracy and quality of the guidance instructions they saw in the videos. In the individual ratings of each video, AR was preferred in case of reliable localization, but VR was perceived as more accurate when errors were introduced. The panoramas in VR helped subjects to orient themselves even if the location estimate of the system was incorrect. However, when asked which method subjects would generally prefer, 58% chose AR. This inconsistency motivated us to gain a deeper understanding of users' preferences.

Understandability and Level of Distraction

Subjects rated four visualizations (text hint, color scale, blur, spirit level) with respect to how likely it would make them raise the phone. The most effective visualizations were the text instructions and spirit level metaphor, followed by color scale and blur. Furthermore, subjects compared two object highlighting visualizations: *Frame* showed a rectangle around the object of interest, while *Soft Border* showed a semi-transparent overlay, smoothly fading out at the borders. We hypothesized that *Soft Border* better hides the inherent inaccuracy and jitter effects of object tracking due to the lack of a sharp border, adding to a more stable, calm visualization. As a consequence, distraction from the navigation task would be reduced with *Soft Border* compared to *Frame*. In fact, subjects rated the *Soft Border* visualization equally attention-raising as *Frame*, but at the same time less distracting.

Discussion and Motivation for Experimental Evaluation

We draw the following conclusions and lessons learned from this initial evaluation, which motivate us to a further iteration of the presented concept, and to an experimental evaluation.

1. A questionnaire-based survey with mockup videos might not reveal the true strengths and weaknesses of AR and VR modes. Users did not actually navigate in a building and thus could not evaluate certain aspects in situ (e.g., the experience on a small screen, or the additional effort to carry the phone). Moreover, using the interfaces while walking (secondary task) might have produced different results than evaluating them in a video (primary task).
2. Subjects perceived the VR mode to be more reliable in case of inaccurate localization. However, they widely preferred AR in a direct ranking, which seems contradictory. We hypothesize that in situ, preference for AR would be lower, since the phone must be carried in an uncomfortable pose for AR to work. Such physical usage factors cannot be determined in an online study. AR probably appeared in the mockup as the more elegant solution, compared to a "flip book" impression of VR.
3. No combined evaluation of AR and VR has been performed to see which mode subjects actually use more frequently in a navigation task.
4. The additional UI elements (indicators to raise the phone up) were only evaluated in terms of understandability, but not in terms of effectivity. Results do not tell if these elements really lead to more detected features and thus to improved localization. It was only examined which of the *Frame* and *Soft Border* visualization is believed to be less distracting (based on mockup videos), but not what was their actual effect based on actual object tracking.

PROTOTYPE

In order to evaluate the previously presented UI in an experiment, we built a prototype in Android 2.3[1] following the tool requirements in [15]. We implemented the described VR and AR modes as shown in Fig. 1c). Users can either manually switch between VR and AR with a button on the top right of the screen, or the system can switch modes automatically based on the gravity sensor readings. In an upright pose as in Fig. 1a), the system switches to AR; in a pose as in Fig. 1b), the VR visualization is selected. Based on empirical trials, we set the threshold angles to an inclination of 35° for switching to AR, and to an inclination of 30° for switching back to VR.

Simulation of Self-Localization and Navigation

We implemented the navigation mechanism with a Wizard-of-Oz (WOz) approach [6]. This allow us to modify the accuracy of position and orientation estimates throughout the different study conditions. Further, WOz enables comparable conditions for all participants. A live localization system would not guarantee reproducible behavior in all trials.

We built a WOz app (see Fig. 4) to control the navigation instructions that subjects see on a predefined path in a reproducible way. With this app, the experimenter sends location information to the subject's device at the desired position of the route, and can deliberately trigger localization and orientation errors. The subject's device uses this information to render the VR or AR visualization accordingly (see Fig. 1c).

The navigation interface on the subject's device is implemented with OpenGL ES 2.0. For VR, it displays 360° panorama images of key locations and draws the navigation arrow on top. For AR, the directional arrow is anchored to virtual "key point" locations similar to VR, except that it is overlaid on live video from the rear camera. The panorama photos of the route used in the experiment and the associated walking arrow directions for each key point have been prepared and stored in the WOz app. For both AR and VR, the compass was used to auto-rotate the visualization, accounting for device orientation. In VR, users could also drag and hold panoramas to rotate them manually; lifting up the finger re-enabled auto-rotation.

Elements Specific To Visual Localization

Out of the four suggested indicators to motivate users to raise the phone up (*Text, Blur, Color, Spirit Level*), we chose a combination of the spirit level metaphor and a text hint, as these two were evaluated best in prior work [14]. The indicator can

[1]As of July 2013, still >33% of devices run Android 2.3 or lower (http://developer.android.com/about/dashboards/index.html, last visited: September 2013)

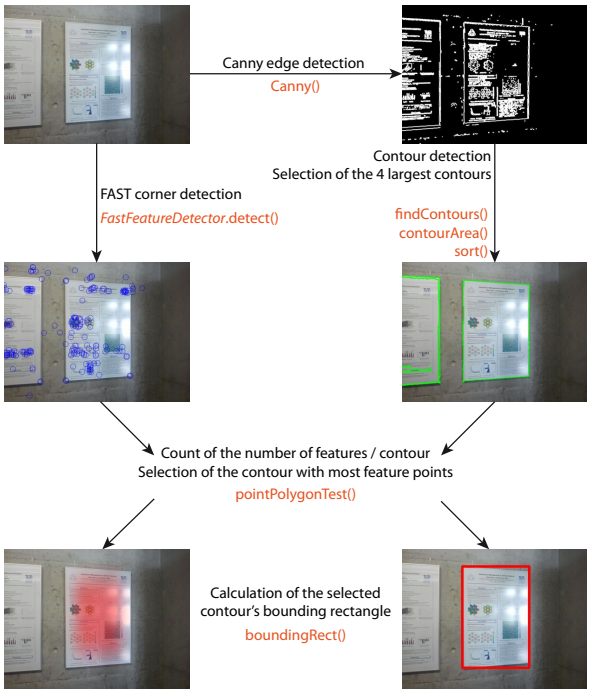


Figure 2. General proceeding for detecting and highlighting objects with two different visualizations: a soft border overlay, supposed to be less distracting (left), and a rectangular frame (right). Best viewed in color.

either pop up automatically when the number of visible features falls below a definable threshold, or it can be triggered through the WOz app. For the automatic trigger, we used a FAST feature detector from the OpenCV framework for Android to detect the number of features in the camera's live image. The anticipated position of the device (90° angle) is determined by the phone's gravity sensor.

We also implemented an object highlighting function which we trimmed to detect posters on uniform backgrounds using the image processing pipeline depicted in Fig. 2. For each frame, a contour detection is applied after edges have been enhanced by a Canny edge detector. The contour containing the most FAST features is regarded as the most interesting object in the scene, and is highlighted. We created two visualizations: for the *Frame* highlight, a red rectangle is drawn; for *Soft Border*, a semi-transparent texture with gradient borders is drawn at the position of the chosen contour.

EXPERIMENTAL EVALUATION

We evaluated the described user interface concept regarding its ability to deal with the previously exposed challenges. By these experiments, we aim at verifying the results of the initial mockup's evaluation. We conducted three experiments, covering the following aspects of the navigation interface: (1) efficiency, perception and convenience of AR and VR under different accuracy conditions, (2) effectivity of UI elements specific to vision-based localization, and (3) convenience and distraction of object highlighting.

In all experiments, subjects used a Samsung Galaxy S II (4.3-inch screen, 8 megapixel camera); the WOz app ran on a

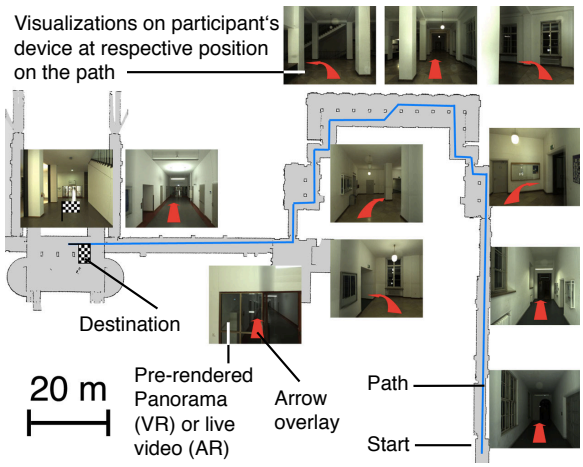


Figure 3. The indoor path used for the navigation task in the study (220 meters), alongside with some sample images and route instructions as they were displayed on the subjects' phone. Best viewed in color.

Samsung Nexus S (4-inch screen). Both devices had a screen resolution of 480×800 pixels.

Participants and Design

12 people (11 males, 1 female) between 23 and 27 years (average age: 24, standard deviation = 1.3) participated in the study. Most subjects were students; none were involved in our research project. No compensation was paid. The experimental design of all three experiments was within-subjects.

Experiment 1: Navigation using VR and AR

Hypotheses

We hypothesize that users reach their navigation destination faster with VR than with AR, i.e., that VR is more efficient **(H1)**. Further, similar to the online study, we suppose that VR will be perceived to be more accurate in case of errors **(H2)**. Although subjects preferred AR over VR in the online evaluation [14] (despite the higher perceived accuracy of VR), we hypothesize that VR would be generally favored in a hands-on study **(H3)**.

Task and Measurements

Subjects performed a navigation task in a university building on a path of 220 meters length (see Fig. 3), using both the AR and the VR mode. The accuracy of the system's location estimate was varied in four conditions (*No Error*, *Position Error*, *Orientation Error*, *Combined Error*), for both AR and VR. Consequently, each user traversed the path eight times. We decided to use the same path in all conditions for better comparability, but counterbalanced the order of conditions with a 4×4 Latin square to weigh out learning effects over all conditions. Subjects were asked to rely only on the given instructions, so that they could not be sure whether the path would not vary.

Navigation instructions were fed into the subject's phone by the experimenter (Wizard of Oz). The experimenter walked approx. one meter behind the subject and sent the appropriate panoramas in VR mode (and directional arrows in AR mode) to the subject's phone using the WOz interface (see Fig. 4,

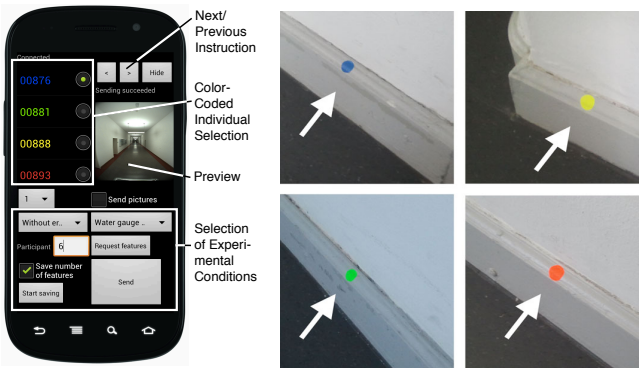


Figure 4. The WOz app for controlling visualizations on the subject's device and simulating localization errors (left). Markers in the corridors (right) helped the experimenter to trigger visualizations at identical locations for similar experimental conditions. Best viewed in color.

left). Colored labels in the app and on the skirting board (see Fig. 4, right) helped the experimenter to choose the correct image at the same locations.

In error conditions, the experimenter replaced correct images and instructions twice by short sequences of misplaced (*Position Error*) and misoriented panoramas (*Orientation Error*). Those errors were introduced at the same locations for all participants. Start and end time of each run (from receiving the first panorama until reaching the destination) were measured by the device. Users were asked to "think aloud" while using the system and answered a questionnaire after each run.

Results of Experiment 1

Efficiency

Subjects were in average 25 seconds faster to reach their destination with VR (averagely 2:39 minutes for the 220 m path) than with AR (averagely 3:04 minutes), which is a significant difference according to a paired sample t-test ($p = 0.002$, $\alpha < 0.05$), and confirms **H1**. With VR, no significant time differences between conditions were found. With AR, differences between conditions were partly significant. Subjects were slower in the *Orientation* and *Combined Error* condition than in the *No Error* or *Position Error* condition (see top right table in Fig. 5). This signifies that AR works worse in case of (particularly orientation) errors.

Accuracy Perception

Subject rated the perceived accuracy in the conditions *Without Error*, *Position Error*, *Orientation Error* and *Combined Error*. Subjects were presented the following statements: *"The system seemed to know well where I am"* (relating to the position estimate), *"The system seemed to know well in which direction I am looking"* (relating to the orientation estimate), *"The navigation instructions were always correct"* (relating to the perceived correctness of individual instructions), and *"Overall, I found the guidance accurate"* (relating to the general guidance accuracy).

Agreements to each statement were indicated on a symmetric 7-step Likert scale where -3 corresponds to "strongly disagree" and +3 to "strongly agree". Fig. 5 summarizes the responses in box plots. As the response format approximates an interval-level measurement, the mean values are indicated in the diagram in addition to medians. However, in the following we only use medians (M) and non-parametric tests to report the results. α denotes the level of significance; W denotes the test statistic in Wilcoxon signed-rank tests.

Both in VR and AR mode, subjects clearly identified position and orientation accuracy differences between the *No Error* and the respective error condition. The Wilcoxon signed-rank test showed p-values below the significance level of $\alpha < 0.05$ for differences in position accuracy (AR mode: $W = 15$, $p = 0.037$, $\alpha < 0.05$; VR mode: $W = 28$, $p = 0.021$, $\alpha < 0.05$) and slightly higher p-values for orientation accuracy (AR mode: $W = 19.5$, $p = 0.073$, $\alpha > 0.05$; VR mode: $W = 55$, $p = 0.005$, $\alpha < 0.05$). This indicates that subjects were able to generally identify the induced position and orientation errors.

However, only with AR, p-values below 0.05 were observed for differences in perceived correctness between error and no error conditions ($p = 0.015$ for position and $p = 0.034$ for orientation). The perceived correctness of instructions was rated significantly higher for VR than for AR. With *Position Error*, rating medians were 3 for VR and 1 for AR ($W = 6$, $p = 0.030$, $\alpha < 0.05$). With *Both Errors*, medians were 2.5 for VR and 1.5 for AR ($W = 3.5$, $p = 0.023$, $\alpha < 0.05$). Only with *Orientation Error*, medians were slightly above significance (VR: $M = 2$; AR: $M = 1$; $W = 4.5$, $p = 0.065$, $\alpha > 0.05$). Those results indicate that VR is generally considered to be more accurate than AR (which supports **H2**).

Convenience and User Preference

Asked for the preferred system, 50% decided for VR, 33% for AR, and 17% were undecided (supporting **H3**). This strong tendency is presumably not only grounded in the quality of navigation instructions, which were perceived to be better in VR, but also in the convenience when using the system. Subjects found carrying the phone more convenient in VR ($M = 2$) than in AR ($M = 0$), which is a significant difference ($W = 0$, $p = 0.009$, $\alpha < 0.05$) The required upright position for carrying the phone in AR was physically constraining. One participant said that it could work *"well for 200 meters, but not more"*. Most subjects found it embarrassing to pass by other people in that pose, because others might fear being recorded. This problem was not given in VR, because the camera in that case pointed towards the floor.

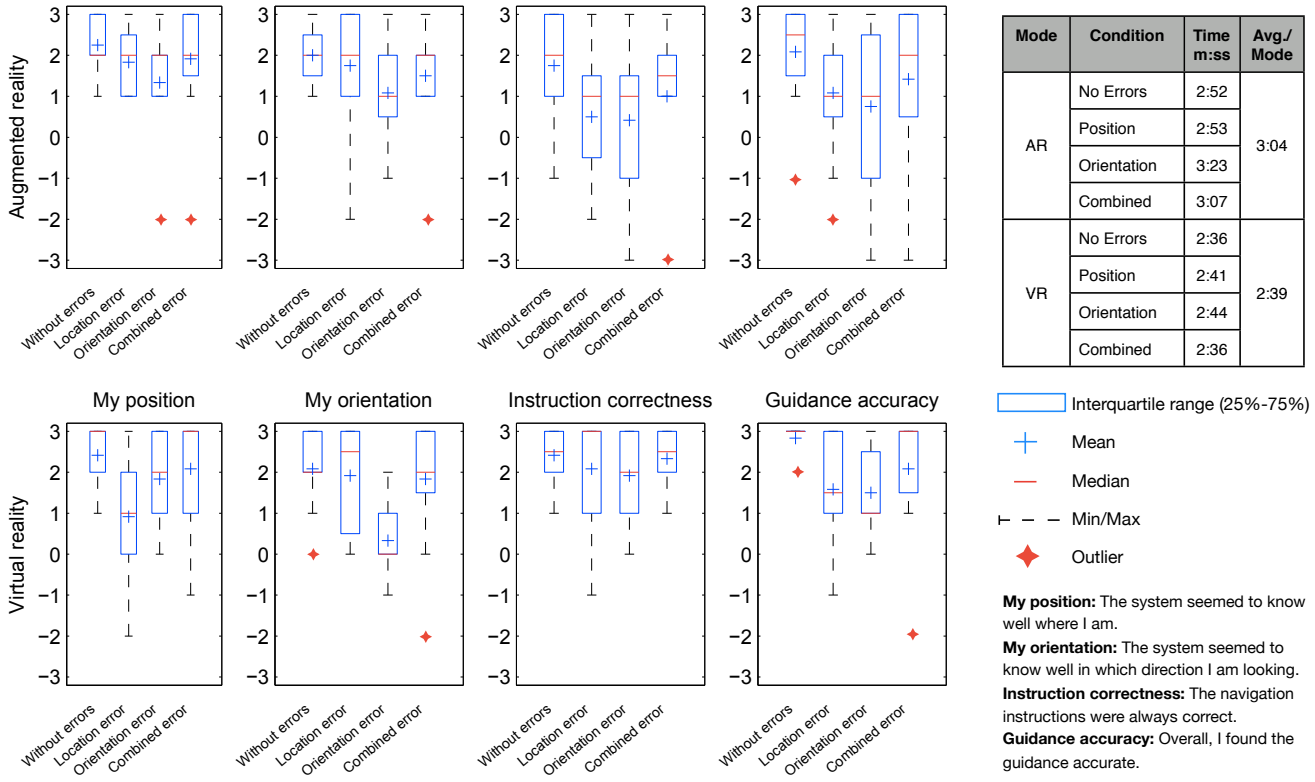
Experiment 2: Effect of Vision-Specific UI Elements with Combined Interface

Hypothesis

We hypothesize that the spirit level indicator actually makes subjects point at areas with more visual features and thereby increases localization accuracy. More precisely, we expect that the visibility of the indicator increases the average number of visual features in the captured images **(H4)**.

Task and Measurements

Subjects performed a navigation task on the path shown in Fig. 3, but in opposite direction as in Experiment 1, so that the path was not already too familiar. Three times during the walk, a relocalization procedure, as it would be required from



Mode	Condition	Time m:ss	Avg./ Mode
AR	No Errors	2:52	3:04
	Position	2:53	
	Orientation	3:23	
	Combined	3:07	
VR	No Errors	2:36	2:39
	Position	2:41	
	Orientation	2:44	
	Combined	2:36	

Figure 5. ***Left:*** **Perceived guidance accuracies in experimental conditions of AR and VR interfaces. The box plots visualize the level of agreement to the statements on the bottom right. (on 7-step Likert scales ranging from -3 to +3).** ***Top right:*** **Task completion time using VR and AR. In AR, Subjects on average took 25 seconds longer, and differences between conditions were higher. Best viewed in color.**

time to time in a self-contained system, was simulated. The experimenter triggered a spirit level visualization (cf. Fig. 1c) to appear on the subjects' device. The indicator told subjects to collect enough features for relocalization. As soon as subjects raised the phone until the bubble was centered on the scale, the indicator disappeared and a location update (i.e., the correct arrow/panorama) was displayed. To increase the degree of realism, the interface automatically switched between the AR and VR visualization based on the phone's inclination, as described in the *Prototype* section. Subjects were not given any instructions how they should carry the phone.

We logged the inclination of the phone (whether it was carried down or upright), whether the feature indicator was currently shown or not, as well as the number of detected FAST features (all in one-second intervals). After the experiment, users answered a questionnaire.

Results of Experiment 2

Reliable localization requires 100 to 150 features in the image (empirical values). While the indicator was visible, the average number of detected features per frame rose from 42 to 101. Given that the amount of frames in which more than 150 features were detected was 20.7% with active indicator, and 8.1% with inactive indicator, the indicator significantly increased the probability for successful re-localization, which confirms **H4**. While those ratios may in overall appear low, it has to be kept in mind that in practice, a certain amount

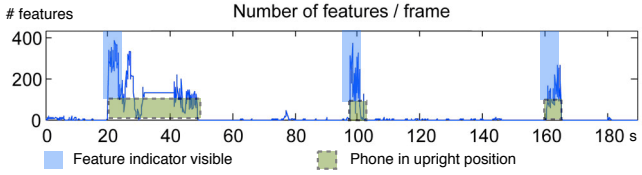


Figure 6. When the feature indicator is visible (light blue), users move the phone up (green) and more visual features are detected per frame. This diagram exemplarily shows one subject's data. Best viewed in color.

of frames will always be subject to motion blur, and 20% of frames with sufficient features still yields on average 5 frames per second (at 25 frames per second), which is sufficient for continuous visual localization. Fig. 6 illustrates, based on an exemplary excerpt of the experiment's data, how the number of features per frame was correlated with the phone inclination and the state of the indicator.

The experiment also showed that subjects preferred the lower carrying position for VR mode, compared to the upright pose for AR mode. They only raised the phone when told so by the visualization, but soon returned to the more comfortable carrying position. None of the subject deliberately chose to carry the phone upright which would have activated AR mode.

Subjects responded that they found the pose-dependent switch between AR and VR convenient (median of agreement M = 2.5). They also understood the meaning of the indica-

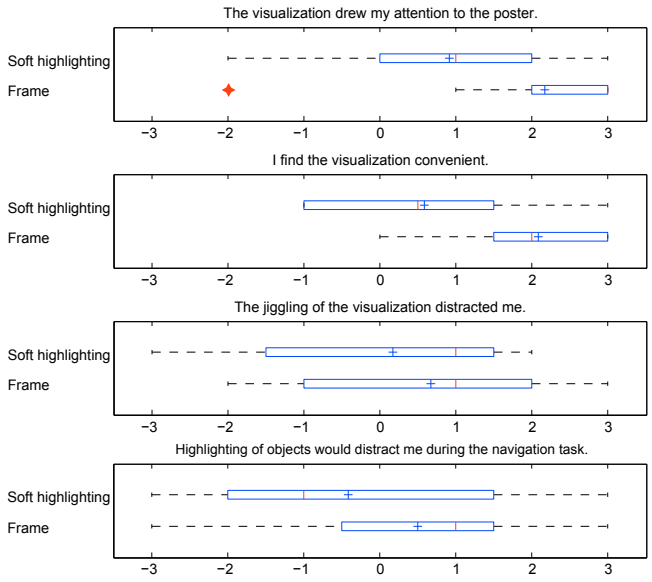


Figure 7. User feedback on *Frame* and *Soft Highlight* object visualization. Answers were given on a 7-step Likert scale, ranging from -3 (strongly disagree) to +3 (strongly agree). For symbol legend see Fig. 5.

tor: they agreed with M = 3 to the statement *"What I should do when the indicator appeared was clear to me"*, and with M = 3 to the statement *"I have been motivated by the indicator to raise the phone up"*.

Experiment 3: Object Highlighting Methods

Hypothesis

We hypothesize that highlighting objects might have a distracting effect, but that a soft border can reduce the effect size, compared to a simple rectangular highlighting **(H5)**.

Task and Measurements

We evaluated the two ways of highlighting objects, *Frame* and *Soft Highlight*, as described earlier and illustrated in Fig. 2. Our algorithm is currently optimized to detect square, feature-rich objects out of a uniform background. This applies to, e.g., a poster on a wall, which we chose as scenario for evaluating the object highlighting mechanism. It was tested beforehand that the posters could be robustly recognized. Subjects pointed at the posters using both highlighting visualizations. Feedback was afterwards collected by a questionnaire.

Results of Experiment 3

The results are summarized in Fig. 7. On a Likert scale from -3 to +3, subjects indicated that *Frame* drew more attention to the poster (M = 3) than *Soft Highlight* (M = 1). Given that the visualization signals a possibility to interact with the object, they found *Frame* more convenient (M = 2) than *Soft Highlight* (M = 0.5). The semi-transparency of *Soft Highlight* complicated readability of text on the poster. Regarding distraction, the visible contours of the *Frame* visualization were perceived as more unstable. During a navigation task, subjects would be more distracted by *Frame* (M = 1) than by *Soft Highlight* (M = −1). Although this is a tendency towards **H5**, this difference was not significant. However, we found significant differences between *Frame* and *Soft Border* for attention and convenience ratings (Student's t-test, $p < 0.05$).

GENERAL DISCUSSION AND LESSONS LEARNED

We now discuss the findings of Experiments 1–3, also in comparative view to the initial mock-up study [14], and formulate lessons learned. We also report on issues that have not been addressed explicitly in our presentation of results, but which have become evident in the course of our study or were explicitly mentioned by participants when "thinking aloud".

VR as Main Visualization

VR mode turned out to be advantageous in several ways. In Experiment 1, it brought subjects significantly faster to the destination, independently of the accuracy condition. Further, the perceived correctness of instructions was higher in VR than in AR, which made the system more reliable even when panoramas were incorrect with relation to position and orientation. Navigating using VR was also more convenient from a practical point of view, since this visualization did not require subjects to hold up the phone all the time (which was perceived to be physically uncomfortable). Experiment 2 confirmed this, where subjects almost "automatically" chose VR when they had the choice how to carry the phone. An additional argument in favor of VR manifested through the "think aloud" technique, where multiple subjects reported that they felt like unwantedly recording or "stalking" other passers-by when walking around with active camera in AR mode.

In the direct vote, subjects clearly preferred VR over AR, in contrast to the initial mock-up study, where subjects liked the AR visualization better. This contradiction could be explained due to the fact that the AR UI might have appeared more appealing in the simulation, and that subjects could not really compare both in practice. Moreover, the physical constraints of AR – the required pose of the phone – seem to be a "knock-out criterion". Hence, we see the hands-on results as more plausible and more in line with the results for efficiency and convenience, which were likewise in favor of VR.

We thus recommend, as a guideline, the VR mode as primary interface for a visual navigation system. Particularly when localization accuracy is not perfect, it allows still reliable and fast guidance, compared to AR.

AR and Feature Indicator to Improve Localization

The AR view, by contrast, can play out its strengths in two cases. First, it can help to improve feature collection using the feature indicator. In the study, the spirit level visualization contributed to a rise of visual features in query images, thus increasing the probability of reliable re-localization. Hence, a visual navigation system could switch to AR mode when the location estimate is too inaccurate even for the robust VR mode, and ask and motivate users to relocalize themselves by pointing at a feature-rich scene.

Second, AR can integrate object highlighting which likewise may contribute to feature-rich query images. It also provides an anchor for interaction with those objects (a detailed analysis of location-based services via an AR interface, however,

is beyond the scope of this paper). We have investigated two highlighting visualizations and found that *Frame*-based highlighting of interesting objects contributed to high attention of users, but at the same time distracted them stronger during navigation. The *Soft Highlight* visualization reduced visual jiggling, but aroused less visual attention and resulted in worse readability of text on posters and signs. As another guideline, a way to combine the advantages of both visualizations could be to use *Soft Highlight* for peripheral objects during a navigation task in order not to distract subjects too much, and to employ the *Frame* visualization once a user focuses an object with the phone.

Automatic AR/VR Switching

Since both VR and AR are useful components of a vision-based navigation interface, future work will have to investigate how both can be combined even better. We need to examine which events could serve as triggers to select them automatically, in addition to the quality of the location estimate (see Fig. 8) and the phone's inclination.

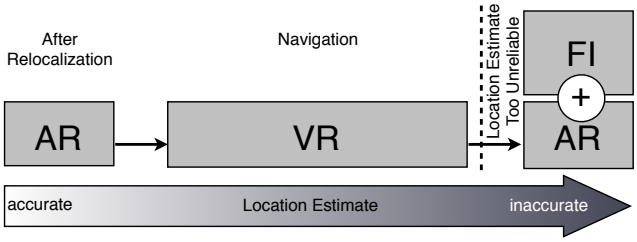


Figure 8. VR is used as main interface during navigation. AR is either used directly after re-localization (to highlight objects of interaction), or when the location estimate becomes too unreliable and a re-localization has to be enforced using an additional feature indicator (FI) element.

Discrepancies between Real and Virtual World

The photorealism of panorama images and the visible landmarks (e.g. posters, exhibits, fire extinguishers) contribute to a simpler identification of a location. However, the real environment often does not look exactly like the recorded panorama images. While color-invariant feature descriptors can minimize the matching problem for the localization algorithm, differences in lighting conditions and exposure changes between subsequent panoramas have been negatively noticed by subjects. However, it did not hinder them in finding their route. To some extent, image post-processing (e.g., exposure correction) could solve this issue.

An advanced solution could choose appropriately from multiple reference sets (e.g. recorded at day and at night) by the time of day. Mapping of (especially crowded) buildings, however, will often have to take place at night when they are closed for the public, and therefore exhibit significantly different lighting conditions than at day. In order to ease mapping of landmarks between panoramas and the real world, characteristic objects could be highlighted in the interface with a similar approach to what we presented in this paper.

Another challenge are permanent changes to the real environment. Posters or advertisements might be replaced from time to time (i.e., within several weeks). As an example, Fig. 9

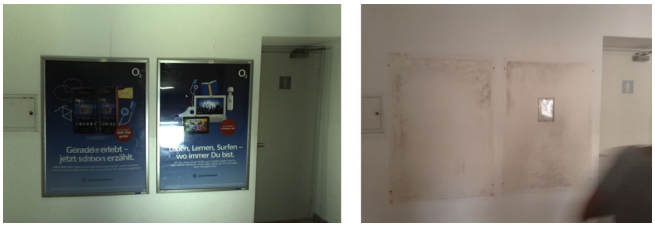


Figure 9. Advertisements that were present in the reference dataset (left) but have been removed at the time of the study (right) were irritating for users, as such salient points often serve as landmarks for orientation.

shows two advertisements in the reference dataset which were not present any more at time of the user study two months later. This is problematic in two ways: Such distinctive objects expose characteristic features and are thus important for visual localization. As a consequence, image matching could fail after a change in the real world. Second, also humans use landmarks for orientation. When they see, e.g., a poster in the VR panorama image, they might search for this poster in the real environment to orient themselves, which could be irritating if it is not present any more.

A possible solution for that problem could be crowd-based updates. Query images users take with their smartphone cameras can be included as new textures and continuously update the reference dataset. However, more profound changes in buildings (such as construction works) that entail detours require not only texture updates, but also adaptions of the underlying 3D model and a different navigation path, which might eventually require re-mapping (parts of) the building.

Frequency of Panorama Updates

Subjects reported that the frequent updates of the panorama images in VR mode (every few meters or less, independent of the walking speed) were partly irritating, especially when not permanently looking at the screen. Since each panorama was slightly different in perspective and lighting, they had to "re-check" their position in reference to the panorama each time they looked back at the display. Some stated to have used mostly the distance indicator (showing the distance to the next turn), and to have looked at the panorama only for double-checking when approaching the turn location.

This leads to the idea of varying the frequency in which panoramas are updated during a path. Instead of showing always the closest view to the current location estimate, a reduced set of panoramas could be used along the route, illustrating particularly the turns and difficult parts. This could reduce the cognitive effort required for visually matching panoramas with the real world, at similar quality of guidance.

LIMITATIONS

Although the evaluation presented in this paper provides valuable insights, it also has limitations. First, this work evaluated interfaces with simulated localization data. This was necessary to test the ability of AR and VR interfaces to cope with varying levels of accuracy. Simulations can however not fully model a self-contained system. For example, although the usage of the spirit level indicator resulted in more visible features, this study cannot tell whether this increase actually

would lead to more reliable localization. It is subject to future work to evaluate our UI concept, which we have shown to be sound and useful, with an underlying live-working visual navigation system. Further, it was not part of this work to evaluate the accuracy of visual localization.

However, we have shown that VR mode provides reliable guidance even with low (simulated) accuracy, making the UI adequate to work on top of a variety of visual localization systems, including such with lesser accuracy. As we have tested responses to various error types and levels of accuracy, we believe that the results will be transferable to a broad range of real-world cases.

CONCLUSION AND FUTURE WORK

We have presented a user interface adapted to some unique challenges of visual indoor navigation, and evaluated a working prototype in a hands-on study. Our concept combines virtual and augmented reality elements, and proved in quantitative and qualitative experiments to provide reliable navigation instructions even with inaccurate localization. It actively contributes to feature acquisition which improves positioning certainty. We identified challenges of visual localization and outlined ways for solving them. We believe that vision-based approaches are a promising technique for indoor navigation. Future work will have to evaluate approaches addressing the mentioned challenges in real-world studies, with a larger user base, and with a live localization system.

REFERENCES

1. Azuma, R. A survey of augmented reality. *Presence-Teleoperators and Virtual Environments 6*, 4 (1997), 355–385.
2. Beeharee, A. K., and Steed, A. A natural wayfinding exploiting photos in pedestrian navigation systems. In *Proc. of the 8th Conf. on Human-Computer Interaction with Mobile Devices and Services (MobileHCI)*, ACM (2006), 81–88.
3. Butz, A., Baus, J., Krüger, A., and Lohse, M. A hybrid indoor navigation system. In *Proc. of the 6th Intl. Conf. on Intelligent User Interfaces (IUI)*, ACM (2001), 25–32.
4. Hile, H., and Borriello, G. Positioning and orientation in indoor environments using camera phones. *Computer Graphics and Applications, IEEE 28*, 4 (2008), 32–39.
5. Hile, H., Vedantham, R., Cuellar, G., Liu, A., Gelfand, N., Grzeszczuk, R., and Borriello, G. Landmark-based pedestrian navigation from collections of geotagged photos. In *Proc. of the 7th Intl. Conference on Mobile and Ubiquitous Multimedia (MUM)*, ACM (2008), 145–152.
6. Kelley, J. F. An empirical methodology for writing user-friendly natural language computer applications. In *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems (CHI)*, ACM (1983), 193–196.
7. Kray, C., Elting, C., Laakso, K., and Coors, V. Presenting route instructions on mobile devices. In *Proc. of the 8th Intl. Conf. on Intelligent User Interfaces (IUI)*, ACM (2003), 117–124.
8. Kray, C., and Kortuem, G. Interactive positioning based on object visibility. In *Mobile Human-Computer Interaction (MobileHCI)*, S. Brewster and M. Dunlop, Eds., vol. 3160 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2004, 276–287.
9. Li, B., Salter, J., Dempster, A., and Rizos, C. Indoor positioning techniques based on wireless LAN. In *1st IEEE Intl. Conf. on Wireless Broadband and Ultra Wideband Communications* (2006), 13–16.
10. Lim, H., Kung, L., Hou, J., and Luo, H. Zero-configuration, robust indoor localization: Theory and experimentation. Tech. rep., Univ. of Illinois, 2005.
11. Liu, A., Hile, H., Kautz, H., Borriello, G., Brown, P., Harniss, M., and Johnson, K. Indoor wayfinding: Developing a functional interface for individuals with cognitive impairments. *Disability & Rehabilitation: Assistive Technology 3*, 1-2 (2008), 69–81.
12. Miyashita, T., Meier, P., Tachikawa, T., Orlic, S., Eble, T., Scholz, V., Gapel, A., Gerl, O., Arnaudov, S., and Lieberknecht, S. An augmented reality museum guide. In *Proc. of the 7th IEEE/ACM Intl. Symposium on Mixed and Augmented Reality*, IEEE (2008), 103–106.
13. Miyazaki, Y., and Kamiya, T. Pedestrian navigation system for mobile phones using panoramic landscape images. In *Intl. Symposium on Applications and the Internet (SAINT)*, IEEE (2006).
14. Möller, A., Kranz, M., Huitl, R., Diewald, S., and Roalter, L. A mobile indoor navigation system interface adapted to vision-based localization. In *Proc. of the 11th Intl. Conf. on Mobile and Ubiquitous Multimedia (MUM)*, ACM (2012), 4:1–4:10.
15. Möller, A., Kray, C., Roalter, L., Diewald, S., and Kranz, M. Tool support for prototyping interfaces for vision-based indoor navigation. In *Workshop on Mobile Vision and HCI (MobiVis) on MobileHCI 2012* (2012).
16. Mulloni, A., Seichter, H., Dünser, A., Baudisch, P., and Schmalstieg, D. 360 ° panoramic overviews for location-based services. In *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems (CHI)*, ACM (2012), 2565–2568.
17. Mulloni, A., Wagner, D., Barakonyi, I., and Schmalstieg, D. Indoor positioning and navigation with camera phones. *Pervasive Computing, IEEE 8*, 2 (2009), 22–31.
18. Narzt, W., Pomberger, G., Ferscha, A., Kolb, D., Müller, R., Wieghardt, J., Hörtner, H., and Lindinger, C. Augmented reality navigation systems. *Universal Access in the Information Society 4*, 3 (2006), 177–187.
19. Schroth, G., Huitl, R., Chen, D., Abu-Alqumsan, M., Al-Nuaimi, A., and Steinbach, E. Mobile visual location recognition. *IEEE Signal Processing Magazine 28*, 4 (2011), 77–89.
20. Walther-Franks, B., and Malaka, R. Evaluation of an augmented photograph-based pedestrian navigation system. In *Smart Graphics*, Springer (2008), 94–105.