Gaze-Guided Narratives: Adapting Audio Guide Content to Gaze in Virtual and Real Environments

Tiffany C.K. Kwok
Institute of Cartography and Geoinformation, ETH Zürich
Zürich, Switzerland
ckwok@ethz.ch

Peter Kiefer
Institute of Cartography and Geoinformation, ETH Zürich
Zürich, Switzerland
pekiefer@ethz.ch

Victor R. Schinazi
Chair of Cognitive Science, ETH Zürich
Zürich, Switzerland
scvictor@ethz.ch

Benjamin Adams
Department of Geography, University of Canterbury
Christchurch, New Zealand
benjamin.adams@canterbury.ac.nz

Martin Raubal
Institute of Cartography and Geoinformation, ETH Zürich
Zürich, Switzerland
mraubal@ethz.ch

Figure 1: A tourist using the Gaze-Guided Narrative system from a vantage point (red: example fixation sequence with fixation locations (fixations 1 - 9)). The system provides gaze guidance when the user does not find the next object (fixation 3), and adapts the content to what has previously been looked at (fixation 6).

ABSTRACT
Exploring a city panorama from a vantage point is a popular tourist activity. Typical audio guides that support this activity are limited by their lack of responsiveness to user behavior and by the difficulty of matching audio descriptions to the panorama. These limitations can inhibit the acquisition of information and negatively affect user experience. This paper proposes Gaze-Guided Narratives as a novel interaction concept that helps tourists find specific features in the panorama (gaze guidance) while adapting the audio content to what has been previously looked at (content adaptation). Results from a controlled study in a virtual environment (n=60) revealed that a system featuring both gaze guidance and content adaptation obtained better user experience, lower cognitive load, and led to better performance in a mapping task compared to a classic audio guide. A second study with tourists situated at a vantage point (n=16) further demonstrated the feasibility of this approach in the real world.

CCS CONCEPTS
• Human-centered computing → HCI theory, concepts and models; Usability testing; Field studies;

KEYWORDS
Gaze-Guided Narratives; Outdoor Eye Tracking; Tourist Guide

ACM Reference Format:

This is a pre-print of an article published in ACM DL. The final authenticated version is available online at: https://doi.org/10.1145/3290605.3300721
1 INTRODUCTION

Tourist destinations around the world have fascinating stories to tell and interesting facts to be revealed. Not surprisingly, tourism has always been among the early-adopting domains for novel HCI technologies [1, 5, 12, 18].

Audio guides are often used for telling stories about tourist destinations. Unlike visual displays, audio guides allow for an unobtrusive user experience by not distracting the tourist’s visual attention from the real world (see also [17]). In addition, audio can also be enriched to create immersive experiences including interactive multi-narrative soundscapes [42], 3D audio [6], or audio augmented reality [22].

A particular challenge of using audio, is that tourists must be able to map descriptions (e.g., "the tower on your left") to the real world. This type of information mapping can lead to misunderstandings (due to ambiguity, cultural biases, or different conceptualizations of spatial descriptions) and to tourists getting lost while exploring a city’s panorama. Moreover, classic audio guides are only capable of playing at a specific pace and do not take into account the comprehension speed of their users. This lack of customization often generates confusion and can lead to cumbersome interactions that require stopping and rewinding the audio.

In this paper, we address these challenges with a gaze-based interaction concept called Gaze-Guided Narratives. Gaze-Guided Narratives adapt the audio content played to tourists exploring a panorama from a vantage point based on their real-time gaze. This technology is facilitated by recent progress in pervasive outdoor eye tracking [8, 19] that enables gaze-based interactions with objects in the wild [2] and developments in the smart glasses consumer market (e.g., the Pupil Labs eye tracking add-on for HoloLens [32]).

We believe that Gaze-Guided Narratives will help users find the reference objects (e.g., buildings) in a city panorama (gaze guidance), while adapting the audio content to objects that have been previously looked at (content adaptation). In a controlled lab study (n=60), we investigated the influence of gaze guidance and content adaptation on user experience, cognitive task load, content learning, and spatial learning for three different panoramas projected on a virtual CAVE (Cave Automatic Virtual Environment) environment. In a second study (n=16), we tested the Gaze-Guided Narratives interaction concept in the real world with tourists visiting a popular vantage point in Zürich, Switzerland.

Our contributions are:

- A report of a real-world study demonstrating the feasibility of Gaze-Guided Narratives with 16 tourists in Zürich, Switzerland.
- An implicit gaze-based interaction concept for audio guides, applied to touristic panorama views,
- An empirical evaluation of this interaction concept through a study in a controlled lab environment with 60 participants,
as well as interact with a tourist map [44]. However, these systems did not allow users to interact directly with real 3D objects. To the best of our knowledge, there is no prior research on gaze-based interaction with a city panorama in the real world.

Different types of eye events, such as saccades, fixations [31], or smooth pursuits [50] can be used for eye-based interaction. Here, we use fixation data in order to determine the interaction. An interaction can be designed as either explicit or implicit [51]. With explicit interaction, users intentionally trigger the interaction. During implicit interaction, the system interprets the user’s regular interaction behavior on a higher level (e.g., in terms of activities [9]). We propose an implicit gaze-based interaction concept that adapts content based on previous visual attention and interprets user’s gaze in a wrong position as a failing visual search.

Gaze Guidance
The system we propose takes the role of a guide that helps the user find objects in the environment. Different approaches for gaze guidance have been suggested. Krejtz et al. [29] demonstrated that verbal audio descriptions can be used to guide a person’s gaze to a target. Gaze guidance has also been combined with vibrotactile feedback [25], visual feedback in AR [45], and non-verbal auditory feedback [34]. In addition, subconscious approaches to gaze guidance have been suggested that use image modulations in the user’s periphery [4, 37]. In this research, we use verbal audio descriptions, but make them dependent on the user’s current gaze position. Since the interpretations of verbal descriptions of spatial scenes often depend on the context [13], we conducted a pre-study (see section 3) to determine the parameters for our verbal gaze guidance.

Interactive Narratives and Spatial Narratives
The field of narratology has a history of investigating the structure of literary stories. Over the last decade, narratives and stories are increasingly being told through a variety of media. Indeed, the introduction of interactive narratives has led to the development of formal computational models of narrative [10]. Azaryahu and Foote [3] examined the way historical spaces can operate as a medium for telling spatial narratives. Looking at existing spatial narratives at historical touristic sites, these authors discuss the way the structure of narratives can vary from linear sequences in time and space to more complex non-linear configurations and configurations based on themes and sub-themes that operate over space and time. Although it is easier to construct linear (chronological) stories that maintain a dramatic arc, location-aware narrative systems can introduce non-linear structures depending on the amount of agency that users have to direct the storyline [14].

3 GAZE-GUIDED NARRATIVES
Concept
Traditional audio guides follow an information push paradigm. While many tourists appreciate the positive side of getting carefully selected content in a well-designed narrative sequence, traditional audio guides are somewhat inflexible. This can be problematic since the sequence of described objects is fixed and the audio speed does not often match the speed of comprehension and the activities taking place in the real world.

With the Gaze-Guided Narratives concept, we aim at diminishing some of the negative aspects of traditional audio guides. Gaze-Guided Narratives is based on two features:

1) Gaze Guidance: The system interactively helps the user find the objects of interest and only starts playing the content when the object has been found. Here, gaze guidance is provided together with directional cues (see section 2) in the form of verbal audio descriptions such as “building A is far left of what you are looking at” (refer to Figure 1, fixation 3). By providing gaze guidance along with directional cues, we prevent a break between the content parts and the guidance parts of the system while limiting the engagement of additional modalities (e.g., haptics) during interaction.

2) Content Adaptation: The system ensures flexibility regarding the sequence of objects described while maintaining the qualities of a good narrative by adapting the content to previously inspected objects (measured with eye tracking; refer to Figure 1, fixation 6). Content adaptations are made based on previously modeled relations between objects. In the context of tourism, we consider temporal, spatial, and thematic relations. For instance, if building B has been looked at at some point before building A, the system might tell that A has been built before B (temporal), A is located left of B (spatial), or A and B have the same architectural style (thematic). These kinds of adaptations should improve learning about the content and about the position of objects.

Implementation
Mapping Gaze to Objects in the Panorama. The system must be able to determine which object (e.g., building) in the panorama is being looked at. Our implementation is based on that suggested by [2]. A reference image is manually annotated with Areas of Interest (AOIs) around buildings that are relevant for the tourist guide. Geometric information of those AOIs is stored in the JSON format [2]. Objects are detected by matching the front-facing video of a head-mounted eye tracker to the annotated reference image using an ORB [46] feature detector. Fixations are calculated with the I-DT method [48] (thresholds: dispersion 1°, min duration 200ms). When a fixation in an AOI occurs an event is triggered and processed by the interaction logics. Apart from being robust
and fast for detection, this approach has been designed for stationary observers, such as tourists in our scenario.

Interaction Logics. Verbal directional instructions are provided by the system to help users identify buildings. Instructions are relative to the user’s current gaze (i.e., left, far left, right, far right, in front of, behind of) and are based on the thresholds determined by a pre-study (described in the next sub-section). A minimum time interval between each two consecutive instructions is needed in order to provide constant feedback to the user while ensuring that the user has sufficient time for interpreting and reacting to the instruction. We chose 4 seconds based on previous research on tactile navigation [43]. When the user successfully locates the objects in the panorama, the system starts to play the audio information. Based on the history of interaction, the system then decides which audio file (i.e., information with/without content adaptation) to play. While the audio content is playing, the user can explore the view without interruption.

Pre-study: Determining Parameters for Gaze Guidance. We performed a pre-study with 8 subjects in the CAVE to determine how participants interpret the verbal descriptions of directions used for gaze guidance when viewing a city panorama. Using the Zürich panorama as stimulus, participants were asked to fixate on a given starting location and then to fixate (head movements were allowed) on an ending location of their choice based on their interpretation of a directional instruction provided by the experimenter. Participants were tested on four directional instructions (left, far left, right, far right) with three starting locations (the center, the left, and the right of the city view). The vertical dimension (in front of, behind of) was not included in the pre-study. Table 1 presents the results of the saccade distances, which were then used to calculate the thresholds (mean plus one standard deviation) for the directional instructions used in the studies (see Table 2).

4 STUDY 1: CONTROLLED LAB STUDY
The goal of the first study was to evaluate the two main features (i.e., gaze guidance and content adaptation) of the Gaze-Guided Narratives system in relation to traditional audio guides. We designed four conditions in order to test the individual and interaction effects of these features (see Table 3). In condition A, the traditional audio guide serves as baseline. Conditions B and C include either the gaze guidance or the content adaptation features. Condition D simultaneously provides both gaze guidance and content adaptation.

This study focuses on the following research questions: RQ1: Does gaze guidance help participants identify and locate buildings? RQ2: Do Gaze-Guided Narratives improve the acquisition of information? RQ3: Do Gaze-Guided Narratives enhance system usability and reduce cognitive load? RQ4: Do Gaze-Guided Narratives enhance user experience (UX)?

Methodology
Participants. Sixty participants (28 females) were recruited for the study. The age of the participants ranged from 18 to 55 years (M = 23.3, SD = 6.4). All participants were native German speakers and were not tourists at the time of the experiment. Participants were recruited via the DeScil (Decision Science Lab) participant recruitment platform of ETH Zürich and were required to have normal or corrected to normal vision (with contact lenses) to participate.

Ethics statement. Written informed consent was obtained from all participants prior to starting the experiment. The participants were paid 30 CHF per hour and were told that they were allowed to abort the experiment at any time.

Materials. The experiment was conducted in a controlled CAVE environment with three large projection walls. A seat was placed at the center of the room at a distance of 1.8 meters from the front facing projection wall. The lights were turned off during the experiment. Participants wore SMI Eye Tracking Glasses (120Hz) and a Bluetooth SONY MDR-ZX770BN headphone during the experiment. The software modules provided by the eye tracking vendor were used for calibration and for recording the front-facing video image frames and raw gaze data.

Table 1: Pre-study for gaze guidance: Mean and standard deviations for saccade distances in visual angles for directional instructions.

<table>
<thead>
<tr>
<th>Directional Instruction</th>
<th>Far Left</th>
<th>Left</th>
<th>Right</th>
<th>Far Right</th>
</tr>
</thead>
<tbody>
<tr>
<td>Left</td>
<td>−24.51°</td>
<td>42.66°</td>
<td>62.43°</td>
<td></td>
</tr>
<tr>
<td>(−6.95°)</td>
<td>(25.22°)</td>
<td>(21.56°)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Center</td>
<td>−55.44°</td>
<td>−39.45°</td>
<td>38.76°</td>
<td>58.47°</td>
</tr>
<tr>
<td>(−14.64°)</td>
<td>(−5.20°)</td>
<td>(6.35°)</td>
<td>(18.82°)</td>
<td></td>
</tr>
<tr>
<td>Starting Location</td>
<td>−61.53°</td>
<td>−44.92°</td>
<td>33.66°</td>
<td>Not tested</td>
</tr>
<tr>
<td>Right</td>
<td>(−13.85°)</td>
<td>(−17.00°)</td>
<td>(7.57°)</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Thresholds used for directional instructions in gaze guidance.

<table>
<thead>
<tr>
<th>Directional Instruction</th>
<th>Far Left</th>
<th>Left</th>
<th>Right</th>
<th>Far Right</th>
</tr>
</thead>
<tbody>
<tr>
<td>Left</td>
<td>&lt; 0°</td>
<td>0° to 67.88°</td>
<td>&gt; 67.88°</td>
<td></td>
</tr>
<tr>
<td>Center</td>
<td>&lt; −44.65°</td>
<td>−44.65° to 0°</td>
<td>0° to 45.11°</td>
<td>&gt; 45.11°</td>
</tr>
<tr>
<td>Right</td>
<td>&lt; −61.92°</td>
<td>−61.92° to 0°</td>
<td>&gt; 0°</td>
<td>Not Used</td>
</tr>
</tbody>
</table>
were calculated on a scale ranging from zero to one hundred.

were generated by an online text-to-speech reader [38] and written in English and translated into German. Audio files were defined by polygons bounding the buildings without AOIs was created based on the official web-pages of each building (if available) and Wikipedia [57]. The content was panoramas.

For each panorama, we selected four different buildings as target tourist attractions. The AOIs around these buildings were defined by polygons bounding the buildings without buffer. To ensure better control of the experiment, the sequence of AOIs was fixed and randomly generated before the creation of the audio content. The content for the different AOIs was created based on the official web-pages of each building (if available) and Wikipedia [57]. The content was chosen such that the audios for different AOIs were similar in length (approximately 1 minute). The scripts were originally written in English and translated into German. Audio files were generated by an online text-to-speech reader [38] and played through the headphone.

Participants completed the Santa Barbara Sense of Direction Scale (SBSODS) [21]. The SBSODS is a self-report measure of sense of direction ranging from zero (low) to seven (high). In order to measure cognitive load and system usability, participants completed the NASA Task Load Index (NASA-TLX) [20] and the System Usability Scale (SUS) [7], respectively. The scores for the NASA-TLX and SUS scores were calculated on a scale ranging from zero to one hundred.

Participants were also asked to complete a 6 factor User Experience Questionnaire (UEQ, [33]). This questionnaire measured the attractiveness (i.e., overall impression of the system), perspicuity (i.e., how easy it is to get familiar with the system), efficiency, dependability (i.e., whether users feel in control during the interaction), stimulation (i.e., if the user is excited and motivated to use the product) and novelty (i.e., the level of creativeness). Items in UEQ range from -3 to +3 with negative ratings representing a negative user experience.

At the end of each trial, participants completed a questionnaire with one multiple choice (e.g., “According to the audio, which of the following is the newest building?”) and six “true/false or unknown” (e.g., “Does Roppongi Hills Mori Tower have a museum in it?”) content-related questions specific to each panorama. Participants also completed a questionnaire (7-point Likert scale each) about their familiarity with the panorama and overall system helpfulness. Specifically, participants were asked whether the system was helpful in identifying the buildings, their familiarity with the cities presented in the experiment, their familiarity with the panorama they interacted with during the experiment and their familiarity with the content presented by the audio guide. All questionnaires were completed via paper and pencil.

Procedure. Upon arrival, participants were given an information sheet that described the experiment and were asked to sign the consent form. Participants then completed a short demographic questionnaire and the SBSODS before being asked to sit on the chair at the center of the CAVE. At this stage, the experimenter helped participants to adjust the eye tracker and headphones to their head. The eye tracker was 3-point calibrated at the beginning of each trial.

On each trial, the system started by familiarizing the participant with the sound effects for the audio transitions and the target buildings. For all conditions, the system first directed the participant’s gaze location to the target building with one initial instruction (see row “a” in Table 4). When transitioning from one building to another, exactly one description of the relative location to the previous building was provided (see row “d” in Table 4). Only the system with gaze guidance provided additional verbal feedback until the participant successfully located the building (refer to rows “b” and “e” in Table 4).

At the end of each trial, participants completed the UEQ, the content-related questions, and the questionnaire about familiarity and system helpfulness. In addition, participants were asked to draw a sketch map within an empty drawing canvas on a tablet. The canvas size had the same resolution as the image of the target environment. Subjects had to draw rectangles based on their memory for building locations. A
Table 4: Example for audio provided (with gaze guidance) and user actions for the Tokyo panorama.

<table>
<thead>
<tr>
<th>System</th>
<th>User</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>Now you can spend some time to look around. In the following part, you will hear information about the panorama of Tokyo... [...].. Now please look straight ahead. Zojoji Temple is on your right hand side. Explores the panoramas freely and then tries to locate Zojoji Temple</td>
</tr>
<tr>
<td>b</td>
<td>It is on the right of what you are looking at... [ more gaze-guidance until AOI found ] Searches and finds Zojoji Temple</td>
</tr>
<tr>
<td>c</td>
<td>[sound effect for target found] Zojoji Temple .. [ content ] Listens to the content</td>
</tr>
<tr>
<td>d</td>
<td>We have now finished talking about Zojoji Temple. Now please look at Reiyukai Temple. You will find it to the far left of the Zojoji Temple. Follows the instruction and search for the next building</td>
</tr>
<tr>
<td>e</td>
<td>It is on the right of what you are looking at... [ more gaze-guidance until AOI found ] Searches and finds Reiyukai Temple</td>
</tr>
<tr>
<td>f</td>
<td>.......... Continues to use the system</td>
</tr>
<tr>
<td>g</td>
<td>This is the end of this part, please fill in the questionnaire provided. Fills in post-session questionnaire</td>
</tr>
</tbody>
</table>

list with building names was provided during the sketch map task. After the last trial, participants filled in the NASA-TLX and the SUS questionnaires.

Study Design and Analysis. We adopted a between-subject design with participants randomly assigned to one of the four experimental conditions. Each participant completed three trials (one for each city panorama) in a pseudo-random order leading to 180 trials in total. The experiment required approximately 45 minutes to be completed. All statistical analyses were performed with SPSS.

Measurements. We were interested in how long it takes for participants to identify the building after the audio of the previous building has finished. The commonly used eye tracking measure time to first fixation [23] is not suitable in this case because the participant might have a random fixation on the building while still performing the search. Instead, we attempted to find the first fixation on the target AOI which falls into a phase of focal attention. We distinguished phases of focal and ambient attention using coefficient \( K \) [30], which is calculated by subtracting the standardized (z-score) fixation duration from the standardized amplitude of the subsequent saccade. The Time to Object Identification (TTOI) was defined as the time between the end of the last instruction ("d" in Table 4) and the first fixation on the target AOI in a phase of focal attention.

Sketch maps were analyzed using bidimensional regression (BDR) [54]. BDR quantifies the relationship between two sets of coordinates. The returned squared regression coefficient \( R^2 \) can be used as a similarity measure between the 2D configuration on the sketch map and that of the panorama image.

Results

Familiarity and Sense of Direction. Participants were more familiar with the Zürich panorama and its contents played through audio guide compared to New York and Tokyo. A one-way repeated-measures ANOVA revealed significant differences in the level of familiarity with the three panoramas (\( F_{2,112} = 232.630, p < .001 \)). Post-hoc tests with Bonferroni correction further indicated that participants were more familiar with the Zürich panorama (6.9 ± 0.8) compared to the New York (3.4 ± 1.9, \( p < .001 \)) and the Tokyo (1.6 ± 1.4, \( p < .001 \)) panoramas. Participants were also more familiar with the New York panorama compared to Tokyo (\( p < .001 \)). A one-way repeated-measures ANOVA also revealed significant differences in familiarity with the audio contents in different panoramas (\( F_{2,112} = 55.839, p < .001 \)). Here again, post-hoc tests with Bonferroni correction showed that participants were more familiar with Zürich (3.3 ± 1.5) than New York (2.6 ± 1.5, \( p = .007 \)) and Tokyo (1.4 ± 0.9, \( p < .001 \)). Participants were also more familiar with the audio contents of New York compared to the contents of Tokyo (\( p < .001 \)). With regards to sense of direction, the mean SBSODS for all participants was 4.81 (SD = 0.83). Results from a one-way ANOVA revealed no significant differences in the self-reported sense of direction for participants in the four experimental conditions (\( F_{3,56} = 0.492, p = .689 \)).

Content-Related Questions. After each trial, participants answered seven content-related questions. The total number of correctly answered questions in each trial was counted (see Figure 2). We conducted a 3 (city) x 4 (condition) mixed factorial ANOVA on the number of total correct answers across trials. Results revealed a significant main effect for city (\( F_{2,112} = 18.26, p < .001 \)). Pairwise comparisons with Bonferroni correction further revealed that the average number of correct answers in Zürich (3.1 ± 1.4) and New York (3.6 ± 1.6) was lower compared to Tokyo (4.6 ± 1.6). There were no other main effects or interactions (\( p > .59 \)).

Sketch Maps. A 3 (city) x 4 (condition) mixed factorial ANOVA on the results of the BDR (see example of a sketch map in Figure 3) revealed that there was a significant main effect of city (\( F_{2,112} = 28.793, p < .001 \)) and condition (\( F_{3,56} = \))
Additional pairwise contrasts with Bonferroni correction revealed that these differences were driven by the interaction between condition A (0.438 ± 0.416) and D (0.829 ± 0.310) for the Tokyo panorama (p = .036).

User Experience Ratings. The UEQ results are illustrated in Figure 4. In general, a trend can be observed in which the conditions with gaze guidance (C and D) were rated higher compared to the conditions without gaze guidance (A and B). We conducted a 3 (city) by 4 (condition) mixed ANOVA for each of the six factors in the UEQ questionnaire. Results revealed a significant main effect for city in terms of attractiveness (F_{2,112} = 4.185, p = .018), perspicuity (F_{1,81,101.47} = 4.185, p < .001), efficiency (F_{2,112} = 7.331, p = .001), dependability (F_{2,112} = 17.240, p < .001) and novelty (F_{1,70,95.34} = 6.876, p = .002). The between-subjects test revealed a significant main effect of condition for perspicuity (F_{3,56} = 5.013, p = .004) and dependability (F_{3,56} = 4.070, p = .011). Post-hoc tests with Bonferroni correction indicated that condition D had higher scores in perspicuity compared to conditions A (p = .033) and B (p = .037). Condition D also had higher scores in controllability when compared to condition B (p = .024). Results of the ANOVA also revealed a significant condition by city interaction for efficiency (F_{6,112} = 2.208, p = .047) and novelty (F_{5,11,95.34} = 2.364, p = .035). No significant difference was found in the post-hoc test analysis with Bonferroni correction.

System Usability and Cognitive Load. A one-way ANOVA comparing the four conditions in terms of system usability (SUS) revealed no significant differences between conditions (see Figure 5). With regards to cognitive load (NASA-TLX), results of a one-way ANOVA revealed a significant difference between conditions (F_{3,56} = 3.752, p = .016). Post-hoc tests with Bonferroni correction revealed that condition A had higher scores in cognitive load compared to condition D (p = .035). No significant differences were found between the other pairs of conditions.

Helpfulness in Identifying Building. Participants were asked in the post-trial questionnaire whether the system helped them to identify the buildings they were looking for. Results (see Figure 6) from a 3 (city) x 4 (condition) mixed ANOVA revealed a main effect of condition (F_{2,112} = 3.633, p = .015). Additional pairwise contrasts with Bonferroni correction revealed that these differences were driven by the interaction between condition A (0.438 ± 0.416) and D (0.829 ± 0.310) for the Tokyo panorama (p = .036).

Average Time to Object Identification. Figure 7 presents the average TTOI for the different conditions and cities. In general, conditions with gaze guidance tended to have longer TTOI than conditions without gaze guidance (i.e., condition C vs A and D vs B). Meanwhile, conditions with content adaptation also tended to have longer TTOI than that without content adaptation (i.e., condition B vs A and D vs C).

Discussion

Results from the experiment provide interesting insights on the interaction with Gaze-Guided Narratives. Critically, we found that:

**RQ1: Does gaze guidance help participants identify and locate buildings?** As expected, the conditions with gaze guidance (C and D) were rated as more helpful than conditions without (A and B). This result indicates that our approach to gaze guidance was accepted by users. Surprisingly, these effects were not apparent for the Zürich panorama.

Results of the eye tracking/TTOI analysis provide a more comprehensive understanding of the helpfulness of gaze guidance. Given that conditions with gaze guidance (C and D) require additional time for providing directional instructions, we did not expect to find differences in terms of TTOI when compared to conditions without gaze guidance (A and B). Findings were consistent with this expectation and revealed that it took approximately 7 seconds longer to identify the object in condition D compared to condition A.

Overall, gaze guidance appears to be more helpful in guiding users to identify buildings, albeit slower. In the context

![Figure 3: A sketch map drawn by a participant in study 1 (condition D, Zürich).](image-url)
of tourism, we believe that this delay is acceptable as tourists exploring a panorama are typically not under time pressure.

**RQ2:** Do Gaze-Guided Narratives improve the acquisition of information? Results from this study show that the number of correctly answered content-related questions did not significantly differ between different conditions. Interestingly, we found differences in correct answers between cities. Here, participants were better in answering these questions for Tokyo compared to New York and Zürich. We believe that these differences may be related to participants paying more attention to panoramas they were the least familiar with.

Results for the sketch map task revealed that participants were more accurate in drawing a sketch map of Zürich, followed by New York and then by Tokyo. This result was expected given their familiarity with these cities. However, sketch map accuracy was also significantly higher for condition D than condition A for the Tokyo panorama. This result is particularly interesting since it shows that Gaze-Guided Narratives is capable of assisting participants in the acquisition of spatial knowledge in unfamiliar cities.

Taken together, the results for the content-related questions and sketch maps, it seems that Gaze-Guided Narratives can be particularly helpful in unfamiliar panoramas – which is a common case in tourism.

**RQ3:** Do Gaze-Guided Narratives enhance system usability and reduce cognitive load? With gaze guidance, users were expected to have lower workload during search and more attention for listening to the audio guides. System usability was thus expected to improve by adding gaze guidance. Surprisingly, results revealed no significant differences in system usability between the four conditions. One reason for this may be that most of the participants complained that the text-to-speech voice used in the audio was unnatural. Since all conditions adopted the same text-to-speech, we suspect that negative effects caused by this limitation may counteract the potential positive effects of gaze guidance and content adaptation. In order to further investigate this issue, we used a native speaker to record the audios for the study in the real world (study 2, see section 5).

Although gaze guidance and content adaptation did not bring significant improvements in terms of system usability, they were capable of reducing the cognitive load when both features were present (condition D). This is an important finding since low cognitive load leaves free capacity that can be allocated for other tasks including the acquisition of information (see above, RQ2). Moreover, low cognitive workload means less fatigue and may allow tourists to stay attentive for longer periods of time.

**RQ4:** Do Gaze-Guided Narratives enhance user experience (UX)? Results revealed that condition D was easier to get familiar with than conditions A and B (perspicuity), and that it provided more controllability than condition B (dependability). The result on dependability was expected because gaze guidance does not overwhelm the user by pushing information in an uncontrolled speed.

A general trend can be observed in which condition D obtained higher scores compared to the baseline (condition A) in all six factors of the UEQ. However, no significant differences were found between the conditions in terms of attractiveness, efficiency, stimulation and novelty. Here again, this may be related to the quality of the text-to-speech voice in diminishing the positive effects brought on by the gaze guidance and content adaptation features of the system.

## 5 STUDY 2: REAL WORLD STUDY

We performed a study with tourists in order to investigate the feasibility of Gaze-Guided Narratives while exploring a city panorama in the real world (RQ5). In this experiment we focused solely on the classic audio guide (condition A) and the full Gaze-Guided Narratives system (condition D).

**Methodology**

**Participants.** Sixteen participants (8 females, M = 30.1, SD = 6.5, range 22-44) with normal or corrected to normal vision (with contact lenses) were recruited. All participants were native Chinese speaking tourists that were visiting Zürich for the first time and were not familiar with the testing site prior to the experiment. The average time spent by participants in Zürich prior to testing was 2.1 days (SD = 1.4; range 1 - 6).
The experiment procedure was similar to study 1 (see section 4) except that participants completed only one trial. In addition, the pre-study questionnaire included a question about the length of stay in Zürich, and the post-study questionnaire contained an additional question on whether the participants were interested in the content they listened to.

**Study Design, Analysis and Measurements.** We adopted a between subject design with participants randomly assigned to either condition A or D. Participants from both groups were exposed to the real-world panorama and required approximately 25 minutes to complete the experiment. All statistical analyses were performed with SPSS. Similar to experiment 1, participants completed the SBSODS, SUS and the NASA-TLX. Participants also answered content questions, familiarity questions and were asked to draw a sketch map. Sketch maps were analyzed using bidimensional regression and the average TTOI was calculated from the gaze data.

**Results**

**Familiarity and Sense of Direction.** Separate one-way ANOVAs were conducted to investigate differences between participants in conditions A and D in terms of their familiarity with the city of Zürich, their familiarity with the contents of the view, their level of interest with the contents in the view and their self-reported sense of direction (SBSDOS). Results revealed that participants in these two conditions did not differ in terms of familiarity with the city of Zürich ($F_{1,14} = 1.960, p = .183$), familiarity with the contents of the view ($F_{1,14} = 2.483, p = .137$), their interest in the contents of the view ($F_{1,14} = .360, p = .558$) and self-reported sense of direction ($F_{1,14} = .381, p = .547$).

**User Experience Ratings.** Table 5 presents results of the UEQ questionnaire. In general, a trend can be observed in which condition D obtains higher scores than condition A in all six user experience factors.

**System Usability and Cognitive Load.** The mean and standard deviation of SUS and NASA-TLX scores are presented in Table 6. The average SUS score for condition D ($M = 82.81, SD = 12.06$) is almost 15% higher than that of condition A ($M = 76.19, SD = 30.46$). The large standard deviation in the average SUS scores for condition A is related to an outlier who rated the SUS scores as 10.

**Content Related Questions and Sketch Map.** Similar to study 1, participants answered seven content-related questions and had to draw a sketch map without looking at the panorama. The number of correctly answered content-related questions and the $R^2$ values returned by BDR are presented in Table 6. Participants in condition D answered a higher number (4.13 questions) of correct questions and were more accurate ($R^2$:...
Table 5: Mean scores and standard deviations of UEQ factors in study 2 (-3/+3 = neg./pos. experience)

<table>
<thead>
<tr>
<th></th>
<th>Condition A</th>
<th>Condition D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attractiveness</td>
<td>1.46 (0.77)</td>
<td>2.17 (0.87)</td>
</tr>
<tr>
<td>Perspicuity</td>
<td>1.72 (1.08)</td>
<td>2.19 (1.16)</td>
</tr>
<tr>
<td>Efficiency</td>
<td>1.03 (0.69)</td>
<td>1.88 (0.76)</td>
</tr>
<tr>
<td>Dependability</td>
<td>1.41 (1.03)</td>
<td>1.88 (1.16)</td>
</tr>
<tr>
<td>Stimulation</td>
<td>0.84 (0.74)</td>
<td>1.91 (0.95)</td>
</tr>
<tr>
<td>Novelty</td>
<td>0.94 (1.18)</td>
<td>2.19 (0.97)</td>
</tr>
</tbody>
</table>

Table 6: Mean and standard deviations of SUS (0 - 100), NASA-TLX scores (0 - 100), number of correct answers (0 - 7), BDR ($R^2$) (0 - 1) and TTOI (in sec) in study 2

<table>
<thead>
<tr>
<th></th>
<th>Condition A</th>
<th>Condition D</th>
</tr>
</thead>
<tbody>
<tr>
<td>SUS</td>
<td>67.19 (30.46)</td>
<td>82.81 (12.06)</td>
</tr>
<tr>
<td>NASA-TLX</td>
<td>40.83 (15.77)</td>
<td>25.42 (12.92)</td>
</tr>
<tr>
<td>Number of Correct Answers</td>
<td>2.86 (1.36)</td>
<td>4.13 (2.10)</td>
</tr>
<tr>
<td>BDR ($R^2$)</td>
<td>0.88 (0.16)</td>
<td>0.96 (0.04)</td>
</tr>
<tr>
<td>Average TTOI</td>
<td>13.02 (17.87)</td>
<td>19.85 (14.37)</td>
</tr>
</tbody>
</table>

Average Time to Object Identification. Results show that participants in condition A were faster (13.02 sec) compared to participants in condition D (19.85 sec) (see Table 6).

Open Comments from Participants. After the experiment, some participants expressed their positive attitude towards Gaze Guided Narratives by asking when it will be publicly available. Some of them suggested that a slower audio speed might have helped them capture important information.

Discussion

**RQ5: Is it feasible to use Gaze-Guided Narratives in the real world?** Comparing the results on system usability between the two studies, we find that the system with Gaze-Guided Narratives (condition D) obtained an average SUS score of 82.81 (SD = 12.06) in the real world and 69.33 (SD = 12.31) in the CAVE (see section 4). An average SUS score above 80 has been suggested to be considered as “pretty good”, while a score of 66 could be considered as average [56]. One possible explanation for this difference could be the higher degree of immersion in the real world. It is also possible that replacing text-to-speech with a voice recording had an influence on system usability. Overall, the SUS score results indicate that participants were capable of using the system in the real world without loss in system usability.

Inferential statistics were not conducted in study 2 because of the limited number of participants (16 participants). However, results for the NASA-TLX, UEQ and TTOI are consistent with those obtained in study 1. Together, these findings suggest that applications of Gaze-Guided Narratives are not limited to virtual environments. Interestingly, study 2 obtained higher scores for all six factors in UEQ, for both condition A and D. This general trend may also be caused by the introduction of a more natural voice in the audio.

**Missing Gaze Data.** The challenge of achieving a high tracking quality has always been considered as a core research question in pervasive eye tracking [8] and eye tracking “in the wild” [19]. Although the tracking quality was not our main research focus and often determined by the commercial eye tracking hardware, the feasibility of using our system in the real world (RQ5) may be nevertheless influenced by it. Sunlight can disturb the reflective properties of the infrared light used by the eye tracker. In the CAVE environment with controlled lighting, the average percentage of missing gaze data for 180 trials was 2.14%, which is 9.66% less than that in the outdoor environment for 16 trials of data (11.8%). Although the percentage of missing gaze data in the real world was larger than that in a controlled environment, this did not seem to affect the SUS scores, and no participant complained about interruptions or unresponsive system behavior.

6 CONCLUSION AND OUTLOOK

We have proposed Gaze-Guided Narratives as an implicit gaze-based interaction concept for audio guides, which is particularly suited for touristic panorama views. The concept was evaluated through an empirical controlled lab study with 60 participants. Results revealed that the Gaze-Guided Narratives system obtained better UX, lower cognitive load, and better performance in a mapping task compared to a classic audio guide. We demonstrated the feasibility of our approach “in the wild” with a real-world study with 16 tourists.

Although the results of our study are promising, they represent only the start of gaze-based interaction for tourist assistance. Next steps include gaze-based tourist recommendations, interest detection, and a support of touristic activities beyond panorama exploration including wayfinding or shopping. These directions will be accelerated by further developments and pervasive eye tracking technology. Finally, Gaze-Guided Narratives could be applied beyond tourism scenarios, such as for mobile learning [47] of place-related content or for the creation of dynamic story lines in location-based games [39].

ACKNOWLEDGMENTS

This work is supported by an ETH Zürich Research Grant [ETH-38 14-2].
REFERENCES


