Pervasive and Mobile Computing 18 (2015) 55–70

Exploring real world points of interest: Design and evaluation of object-centric exploration techniques for augmented reality

Markus Tatzgern *, Raphael Grasset, Eduardo Veas, Denis Kalkofen, Hartmut Seichter, Dieter Schmalstieg

Institute for Computer Graphics and Vision, Graz University of Technology, Graz, Austria

A B S T R A C T

Augmented reality (AR) enables users to retrieve additional information about real world objects and locations. Exploring such location-based information in AR requires physical movement to different viewpoints, which may be tiring and even infeasible when viewpoints are out of reach. In this paper, we present object-centric exploration techniques for handheld AR that allow users to access information freely using a virtual copy metaphor. We focus on the design of techniques that allows the exploration of large real world objects. We evaluated our interfaces in a series of studies in controlled conditions and compared them to a 3D map interface, which is a more common method for accessing location-based information. Based on our findings, we put forward design recommendations that should be considered by future generations of location-based AR browsers, 3D tourist guides or situated urban planning.

1. Introduction

Mobile devices such as smart phones allow users to access location-based information anywhere and at anytime. For instance, tourists can query information about surrounding points of interest in a foreign city, a task which can also be supported by mobile tourist guides [1]. The information is commonly presented using a spatial representation, such as 2D or 3D maps. 3D maps even allow exploring real world objects freely since they are not bound to the egocentric viewpoint of the user. However, mobile map solutions are not optimally designed for urban exploration [2] and provide limited capabilities to access the data and to relate it to the real world. For instance, users of 3D maps often try to align the virtual viewpoint of the map with their egocentric viewpoint for easier orientation, a strategy that is not well supported by the interface [3]. The only alignment feature 3D maps offer is to align the exocentric top-down view with the general viewing direction of the user. Another issue of currently available 3D maps is that the camera view of the object is often occluded by nearby structures, which is especially problematic in densely built-up areas (Fig. 2).

Augmented Reality (AR) is a natural choice for exploring location-based information of real world objects, because AR overlays information directly into the user's surroundings. For instance, a user can easily access additional information about a building in an urban environment by pointing an AR-enabled mobile phone into its direction. However, in contrast to a map interface, users are limited to the inherent egocentric reference frame of an AR interface, which becomes an obstacle

* Corresponding author. Tel.: +43 3168735085.

E-mail addresses: tatzgern@icg.tugraz.at (M. Tatzgern), raphael.grasset@icg.tugraz.at (R. Grasset), veas@icg.tugraz.at (E. Veas), kalkofen@icg.tugraz.at (D. Kalkofen), seichter@icg.tugraz.at (H. Seichter), schmalstieg@icg.tugraz.at (D. Schmalstieg).

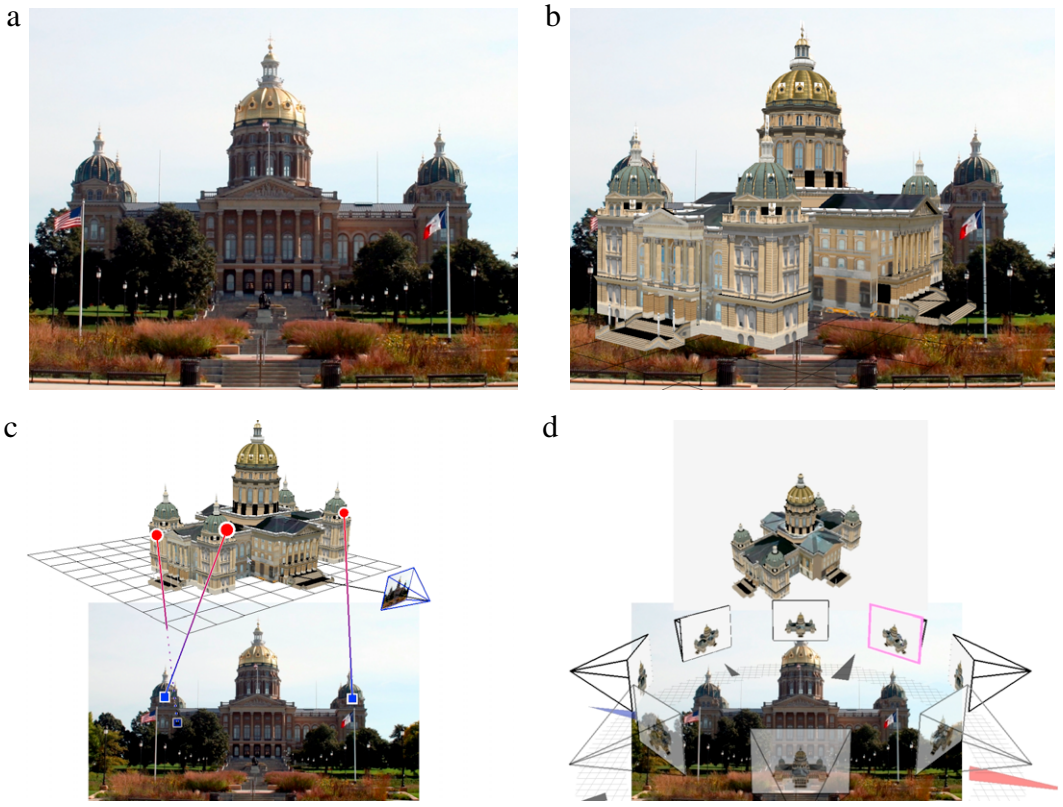

Fig. 1. (a) How can I explore the Iowa State Capitol without physically moving? We present Augmented Reality interfaces using a virtual copy metaphor to access additional views, e.g., (b) uses an in-place, (c) a separated 3D copy with visual links between virtual and real world objects. (d) We also present a spatially separated interface, which uses a 2D copy of the real world object. The available viewpoints are arranged as a circle around the real world object. The current viewpoint is highlighted.

once the user wants to explore objects that are out of reach. The user would need to physically move to a new position, which might be too cumbersome or even impossible.

To deal with these limitations, we introduce object-centric exploration (OCE) techniques for handheld AR, which use a virtual copy metaphor [4] to gain access to distant viewpoints of a real world object in the user's AR view. In contrast to 3D maps, OCE techniques allow a user to focus on the one object he is interested in. OCE techniques also do not suffer from occlusions from neighboring structures, because a virtual copy of only a single object is presented. To present additional viewpoints of this real world object, our OCE interfaces separate the virtual copy (focus) from its real world counterpart and from its surroundings provided by the AR video (context). We consider spatial and temporal techniques for combining focus and context [5]. While the former separates focus and context in space, the latter does so over time, thus removing the context from the interface. Fig. 1 shows spatial OCE techniques that preserve the context by either overlaying the copy on the context (Fig. 1(b)) or separating the copy from the context (Fig. 1(c)).

We explore different designs of OCE interfaces for the exploration of buildings in an urban setting. We perform a series of studies to evaluate our initial designs and the ability of the user to relate the virtual information to the real world. We perform studies under controlled conditions and collect real world experiences with our interfaces in a real world pilot study. Based on the results from the real world pilot study, we evaluate the performance of our designs and compare them to a more common 3D map interface. We summarize our findings in design recommendations that should be considered when developing OCE interfaces for potential application areas such as future generations of location-based AR browsers, 3D tourist guides, or situated urban planning. Relevant real world objects could be annotated with additional information that can easily be explored using OCE interfaces.

2. Related work

In line with Cockburn et al. [5], we classify the related work into spatial and temporal techniques.

Spatial techniques. Hoang and Thomas [6] provide a zoomed view of distant objects to improve interaction accuracy. However, they do not allow free viewpoint selection. Other solutions use a world-in-miniature (WIM) [7] to complement the egocentric view of the user. Bell et al. [8] use a WIM in AR that shares annotations with the real world. This concept is similar to our visual links (Fig. 1(c)), which connect the virtual and real worlds. Unlike shared annotations, in which the

Fig. 2. 3D map. A 3D map (here: Google Earth) allows users to explore surrounding real world objects. However, the user first has to identify the corresponding virtual object in the map and then relate it to his current position. Furthermore, in densely built-up areas, neighboring buildings will cause occlusions of the virtual viewpoint during exploration.

annotation either addresses the real world or the virtual world, visual links always connect both real and virtual worlds. Bane and Höllerer [9] present a WIM interface, in which users are able to seamlessly switch to a copy of an occluded room and interact with this copy. Similar to OCE techniques, the WIM interfaces for AR provide copies of real world objects. However, unlike our interfaces, these interfaces were designed for head-mounted displays (HMD), not handheld devices.

Keil et al. [10] overlay a historical 3D representation of a building on a previously taken picture (context) on a mobile phone, but restrict its viewpoint to the egocentric viewpoint of the picture. Another mode allows users to freely explore the 3D model, but, unlike our interfaces, without providing the context and without seamless transitions between the real world and the virtual copy.

Spatial techniques can also be realized through the use of multi-perspective presentations. Bichlmeier et al. [11] use a mirror to reveal those parts of a table-sized object that face away from the user. Au et al. [12] demonstrate how mirrored views from a live video facilitate orientation in urban environments. Mulloni et al. [13] evaluate different panoramic representations of the users' surroundings by letting them match features in the panorama and the real world. In another paper, Mulloni et al. [14] increase the users' egocentric field of view (FOV) by zooming in on a panoramic presentation of the surroundings. Sandor et al. [15] deform real world buildings to enable users to investigate parts not visible from their current viewpoint. Veas et al. [16] seamlessly integrate an exocentric viewpoint into the egocentric AR view to provide an overview over a large area. While these techniques extend the egocentric viewpoint, they do not allow viewpoint changes for exploring distant objects.

In another work, Veas et al. [17] allow users to transition between live video feeds for exploring outdoor environments. Similarly, Sukan et al. [18] allow users to virtually return to previously captured viewpoints in a table-top application scenario. Both approaches register images of these viewpoints in the real world, thus creating a multi-perspective rendering, which is similar to the ring of images used to navigate viewpoints in our 2D interface. However, their designs are not focused on exploring a single, real world object of interest, but aimed at communicating available viewpoints of the environment [17], or manipulating VR content in AR [18].

Temporal techniques. Bowman et al. [19] present instant teleportation techniques in VR environments and discover that the lack of continuous motion cues causes user to be disoriented. Kiyokawa et al. [20] allow users to seamlessly teleport between a virtual and an augmented collaborative workspace. Seamless transitions are also provided by the MagicBook [21], which allows users to switch between an exocentric AR view on a VR scene and an immersed egocentric VR view on the same scene. In contrast to this previous work, our interfaces switch from an egocentric to an exocentric viewpoint and are designed for exploring real world objects.

Avery et al. [22] and Mulloni et al. [14] switch to exocentric viewpoints to provide an overview on the surroundings. However, the overview is focused on the user's position and does not allow viewpoint changes around a focus object. Sukan et al. [18] and Veas et al. [17] allow switching to already established viewpoints and perform transitions to these viewpoints. In contrast to our 3D interfaces, they provide only access to a discrete number of views.

Spatial cues. When navigating in VR, users need to keep track of the relationship between the egocentric view and spatially or temporally separated views. Since VR lacks natural locomotion and therefore multi-sensory input [23], users require artificial cues to be able to understand the transformation between viewpoints.

Using ideas from cinematography, Woods [24] proposes the concept of visual momentum between two views. Preserving this momentum helps the user to understand changes, such as changes in camera views in a movie. This concept can also be applied to switches between AR and VR views. Spatial awareness can be maintained using spatial cues, location and

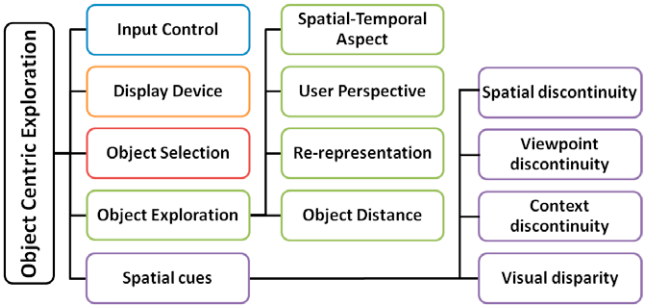

Fig. 3. A hierarchical diagram of the discussed design space for object-centric exploration techniques in AR.

navigation aids. Fitzmaurice et al. [25] discusses some of these aspects for the design of camera control in 3D modeling software. Veas et al. [17] analyze the spatial awareness of a user switching between different egocentric viewpoints. They propose transition techniques to supply missing information between cameras. Other techniques use transitions to enter an egocentric [21] or exocentric view [9,22,16].

In this paper we use smooth transitions [21] as a standard spatial cue to support the user in understanding the relationship between focus and context viewpoints. We use additional cues to communicate the relative position between object and user.

3. Design space

Our goal is to create OCE techniques for mobile AR that enable users to remain at a physical location and explore a real world object from arbitrary viewpoints taken on its virtual copy. For instance, the 3D model of the building in Fig. 1(b) is virtual copy of the physical one.

The design aspects of OCE techniques derive from different steps involved in the exploration of real world objects through a virtual copy in AR. OCE techniques must enable users to identify a selectable real world object and provide means to select and interactively explore this object using appropriate input controls and feedback on a display device. OCE techniques also require spatial cues to facilitate mental linking between the virtual copy and the real world. Fig. 3 outlines the aspects discussed in this section.

Input control. The input control is defined by the input type, the mapping of the input data to programmatic functions, and the feedback to the user in regard to how to perform an action. The input type for mobile AR may be one or more of the following: single and multi-touch, locomotion, sensors and speech input.

Display device. The presentation medium impacts the overall presentation of OCE interfaces. An interface designed for a handheld device, such as a mobile phone, might not work for a HMD. For instance, when using video see-through HMDs, manipulations of the video background also influence the real world view of the user. A user with a handheld device still has an unmodified view on the world by looking past the device.

Object selection. Object selection connects the user's intent with a specific virtual or real entity for subsequent tasks. In AR, not all of the objects of a scene can be selected and interacted with, unless a full, dense and semantically connected 3D reconstruction of the scene is available. Because such a reconstruction is hardly ever available, users require guidance to recognize interactive objects in the AR scene. Hence, this aspect requires selection guidance to highlight [26] which objects are interactive.

Object exploration. To categorize the design of the exploration technique, we identified four aspects that consider the relationship between the real world object and its virtual copy: the type of separation; the user's perspective; the properties of the virtual copy and; the size of the object on the screen.

Spatial–temporal aspect. Additional viewpoints of an object can be spatially or temporally separated from the original AR viewpoint. Spatial techniques preserve the egocentric viewpoint of the user in the video image, and show additional viewpoints of the object. Temporal techniques also provide additional viewpoints, but do not preserve the original viewpoint of the object.

User perspective. In outdoor mobile AR, users explore the world from an egocentric perspective. OCE techniques can provide the user with exocentric viewpoints of an object. In indoor environments such as tabletop setups [21], the user already has an exocentric view of the object. In this case OCE techniques can complement an egocentric perspective.

Virtual copy. The virtual copy of the real object depends on the available data. The object can be represented using different media (3D model or 2D picture) and can be placed either in a 2D image space (e.g., map) or within a 3D coordinate system.

Object distance. The projected size of the real object depends on the screen size and its position relative to the user. The object may either be too large (too close) or too small (too distant) to effectively combine virtual and real views with a spatial technique. In this case, a temporal technique can be employed, which replaces the real object with one taken from a more appropriate viewpoint.

Spatial cues. Spatially or temporally separating the virtual copy from the real world object creates discontinuities between the virtual and the real world. Users have to be able to link both worlds in order to transfer the spatial knowledge gained in one representation to the other representation. Spatial cues facilitate this linking to overcome the following discontinuities.

Spatial discontinuity. Moving the virtual copy out of its original position in the context creates a spatial discontinuity.

Viewpoint discontinuity. Changing the viewpoint of the virtual copy causes a misalignment of this viewpoint with respect to the real world viewpoint.

Context discontinuity. A context discontinuity occurs when the video image of the real world is modified. In extreme cases, an interface zooms in on the object and removes its context completely.

Visual disparity. The degree of visual disparity depends on the quality of the virtual copy. There is no disparity, when the virtual copy perfectly matches the real object. Note that not only the virtual copy can be adapted to become more similar to the real world, but also the representation of the real world can be changed. In this case, the context discontinuity may increase, but at the benefit of decreasing the object discontinuity. Hence, spatial cues can resolve certain discontinuities, but aggravate others.

Spatial cues that address all of these discontinuities are transitions between real and virtual spaces [21,27]. Hence, we use transitions as a standard cue when switching between the copy and the real world. For instance, when switching to the copy, the virtual copy is gradually faded in (addresses visual disparity) and seamlessly rotated to a bird's eye view (addresses viewpoint discontinuity). At the same time, the virtual copy and the context are rearranged on the screen using an animated transition (addresses spatial and context discontinuity).

4. Interface design

In this work, we explore a limited subset of the design space. Our focus is the design of spatial–temporal representations of the interface and evaluating these with respect to the spatial awareness of the user. Therefore, we explore the aspects of *object exploration* and *spatial cues* in detail.

In our designs, we only consider handheld devices as *display device*, because these devices are widely available and a major platform for AR applications. We assume that the mobile device has a large screen, to be able to experiment with screen-space demanding designs. Our interfaces are designed for landscape mode, which is the default mode of currently available AR browsers. Furthermore, we only use a common single-touch interface for *input control*, and we *highlight* selectable objects with a simple frame.

The *user perspective* is defined by our application case, where we focus on large-scale outdoor exploration. Hence, in accordance with the design space, the user perspective is always egocentric and extended by exocentric viewpoints. For the *object distance* we assume the ideal case where the real world object is presented at a sufficient scale so that all of the features relevant to our studies are clearly visible. For the *virtual copy* we assume that we have access to a 3D model of the object. In the following, we refer to the initial view containing only the real world object as AR mode, and to the mode containing the copy of the object as VR mode.

Spatial separation techniques seem to be the most relevant choice for exploring large objects in an outdoor setting, because they preserve the real world context. We expect that spatial separation techniques create an artificial bridge for mapping content in the virtual copy to the real world. To investigate this aspect, we developed a 3D interface and a 2D interface with spatial separation between focus and context.

In the **3D separation interface** (3DSEP) (Fig. 5(b)), a 3D copy is presented, which allows for the continuous exploration of different viewpoints of the object. The user interacts directly with the 3D copy through a virtual orbit metaphor. When entering the VR mode, the copy is viewed from a bird's eye perspective. We integrated common spatial cues into the interface to allow users to mentally link the viewpoint of the copy to the original viewpoint of the context. A grid shows the ground plane of the copy and a camera icon, located in the coordinate frame of the copy, indicates the original egocentric viewpoint relative to the object. A radar icon in the top right shows the same information in a more abstract visualization and from a top-down view. The copy is in the center of the radar, while a dot rotating around the center indicates the camera position relative to the object.

The **2D separation interface** (2DSEP) (Fig. 5(a)) uses images as copy. These images could be pictures taken from the real world object. To avoid visual disparity of the focus between both interfaces, we render them from the same 3D model used in 3DSEP, taken at equidistant positions (45°) on a horizontal circle around the object, with the camera pointing towards its center. The viewpoints are elevated to bird's eye views. The user can replace the zoomed image at the top of the interface by using an explicit one-finger tap, or by swiping over the set of images. The ground plane is rotated upwards around the x-axis so that the images do not occlude each other or the object. In contrast to 3DSEP, 2DSEP does not provide continuous viewpoint updates.

We included corresponding spatial cues from 3DSEP in 2DSEP. We did not include the cues in the rendered images, but only applied them to the image circle, so that we could investigate if the circle is sufficient for users to orient in the interface. We added a grid to visualize the ground plane on which the images are placed and removed its center to avoid occlusions of the real object in the video image. Each image in the circle received a camera icon representation. A radar-like cue is achieved by the relation of the currently selected highlighted image to the image showing the frontal view.

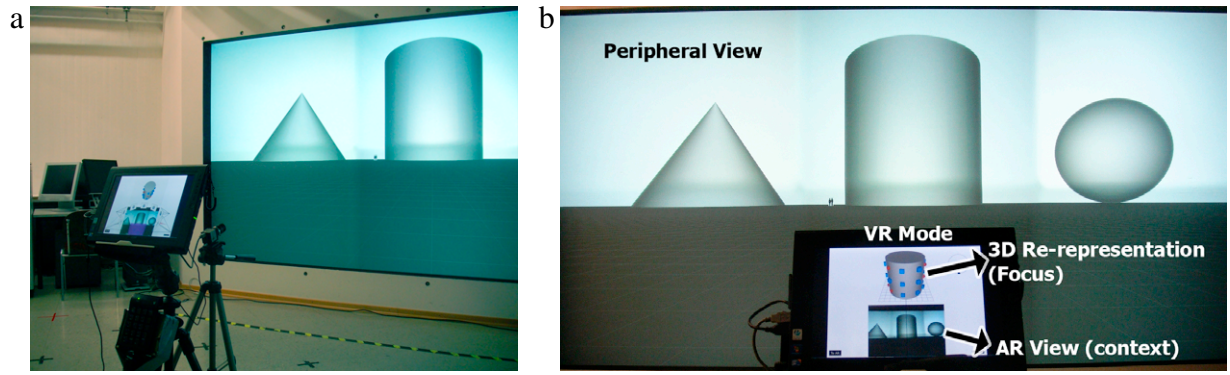

Fig. 4. Indoor apparatus. (a) The apparatus used during the laboratory studies. We placed a tablet PC in front of a back projection wall. Users were seated in front of the tablet PC looking at the wall. (b) The view of an abstract scene from the participant's position showing the peripheral view in the background and the used tablet PC in the foreground. The tablet PC shows the VR mode of the 3DSEP interface.

Aside from these spatial cues, both interfaces provide a smooth transition between AR mode and VR mode to connect these spaces [21]. When entering the VR mode, the video image is scaled down and moved to the bottom of the screen, while the copy is moved to the top of the screen. Spatial separation fully preserves the context at the cost of introducing a spatial offset between focus and context. Spatial discontinuity is alleviated by seamlessly animating the transition of focus and context. Visual disparity is addressed by gradually fading the copy in and out.

To address both viewpoint and spatial discontinuity, we added a switchable spatial cue called visual links (as shown in Fig. 1(c)) to 3DSEP, thus creating interface 3DSEP + L(inks). Links provide a visual connection between the copy and the real world object. By tapping on a location on the 3D copy, a user can create a 2D line to the corresponding location in the video image. The line style is adapted to communicate occlusion with the focus object, and color coded to communicate the end points.

This paper extends on our previous work on OCE techniques [28] by adding a design space for such interfaces and discussing the interfaces with respect to this design space. Furthermore, we performed a study simulating real world conditions to evaluate our interfaces against each other and a more traditional map interface.

5. Laboratory study: abstract scenario

We explored the usage and usability of OCE techniques in a series of user studies. We focused on how users interact with our techniques independent from the semantics and salient content of the real world. Therefore, we evaluated the interfaces using abstract scenes with basic geometric shapes.

5.1. Evaluation testbed

To avoid confounding factors from the real world, we used a simulation testbed for AR. A testbed allows us to present artificial environments and structures with which the participants are not familiar. These scenes can represent real world environments, or can be purely abstract. Testbeds for simulating AR scenarios have already been used to control the registration error [29] or variable lighting conditions [30]. Testbeds were also used to overcome technical limitations of currently available hardware [31].

In our scenarios, a user has already found a real world object of interest and is looking towards it. We assume that the user remains stationary while exploring the object with our interfaces and thus does not require an immersive 360° view of the environment. Therefore, we simulate the peripheral view of the world with a back-projection wall (4×2 m, 4000×2000) pixels) used in daylight conditions (Fig. 4(a)).

We seated participants in front of the wall and mounted the AR device on a tripod in front of them, to simulate holding a handheld device, while at the same time removing the associated physical fatigue. The AR device, a tablet PC (Motion Computing J3500, 12.1″), showed a static snapshot of the environment (1066×800) pixels) that simulated the view through a video camera. Fig. 4(b) shows the view of the participant when seated in front of the wall.

5.2. Experimental design

The following studies are within-subject and share the same experimental design and apparatus (Section 5.1). They differ only in their interface conditions.

Scenario. We rendered a virtual scene consisting of only basic geometric shapes (cone, elliptic cylinder, sphere). The scale and position of these were chosen to resemble real buildings (e.g., elliptic cylinder, 35 m in height, half-axes length $x = 17$ m

Table 1
Questions asked in the studies. All questions except for Q11 use a 5-point Likert scale (1 = strongly agree).

Q1	It was easy to solve the task using the interface.
Q2	I did not feel confident using this interface.
Q3	The interface was intuitive to use.
Q4	I did not like the presentation of the interface.
Q5	The presentation of the images does not reflect the location relative to the object.
Q6	The camera icon helped me orienting.
Q7	The radar icon did not help me orienting.
Q8	I did not like the visual links.
Q9	The links helped me to solve the task.
Q10	The visual links are intuitive.
Q11	Rate how you liked the interface. (1–5; 1 = worst)

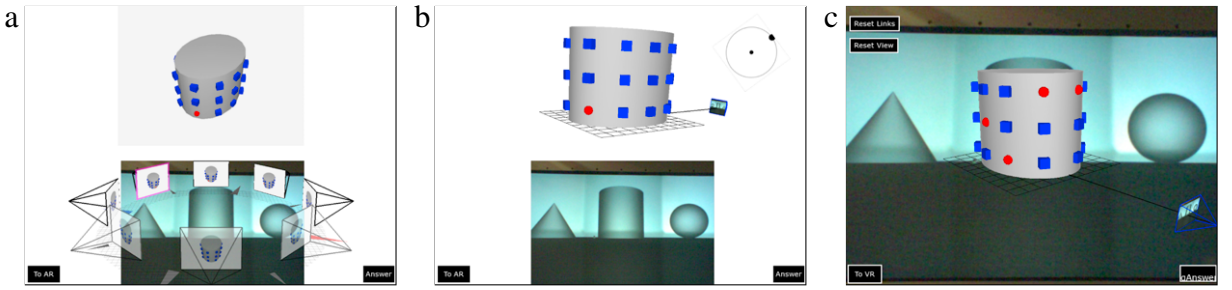

Fig. 5. Interface designs for studies in abstract conditions. Spatially separated interfaces using (a) images (2DSEP) and (b) a 3D copy (3DSEP), as used in the first study. (c) The 3D in-place interface (3DINP + L), as used in the second study.

and $z = 22$ m). The peripheral view was rendered using a virtual camera (60° FOV), placed 120 m from the scene at eye level of the participants. A human scale icon was used as a reference. The AR view was taken with a camera (60° FOV) mounted on the tablet PC.

Tasks. The tasks are representative of interaction with real world 3D objects: (T1) a counting task, where users navigate the copy to find particular figures and count them; (T2) a mapping task, where users search the copy for a single object and point to its location in the peripheral view. For both tasks, the scene included distractors (blue cubes) and targets (red spheres), which were placed on the cylinder in a regular pattern (10 angles, 3 elevations). For T1, five to seven spheres were randomly distributed around the object. For T2, only one sphere was placed at a random location on the grid. The distractors and targets were only visible on the virtual copy. The scene in the peripheral view was not modified.

Procedure. For each task and interface, the participants had one practice trial without time constraints. T1 trials were completed by entering the number of counted spheres on an auxiliary keypad, T2 trials by point-and-click to the location of the sphere in the peripheral view with a laser pointer. Participants completed questionnaires (Table 1) between each interface and task, and after the experiment. We recorded task completion time for both T1 and T2, counting error for T1, and a pointing error for T2. The latter was estimated using a vision-based method, which provided the Euclidean distance for images with resolution 640 × 480.

5.3. First study: varying copy and cues

This explorative study compared our first interface designs (3DSEP, 2DSEP, 3DSEP + L) to evaluate 2D and 3D copy representations and the spatial cues. Fig. 5(a) shows 2DSEP and (b) 3DSEP as used in this study. 3DSEP + L is the same as 3DSEP with the option to create visual links.

Participants. A total of 24 participants (12m/12f), 16–35 years old (mean = 25.9, sd = 4.2), performed 5 repetitions (720 trials) for each task and interface. The presentation order of interfaces and tasks was counterbalanced.

Results. For each interface × task condition and participant, we calculated the mean completion time and error from the 5 repetitions (see Table 2). We performed non-parametric tests, because our sample violated normality. A significant effect of interface on time was only observed for T1 (Friedman, $X^2(2) = 22.3,\ p < 0.001$). A post hoc Wilcoxon signed-rank test with Bonferroni corrected $\alpha = 0.0176$ showed that for T1 3DSEP ($p < 0.001$) and 3DSEP + L ($p < 0.001$) were significantly faster than 2DSEP. Otherwise the performance data revealed no significant effects.

Questionnaires. The questionnaire data is summarized in Table 3, significant effects in Table 4.

Observations. A general strategy, which was applied to all of the interfaces, was panning around the object. Alternatively, in 3DSEP and 3DSEP + L, some participants moved to a top-down view and performed small viewpoint adjustments to look around the edges of the top and locate the spheres. In 2DSEP, participants also adopted the strategy of only selecting viewpoints, which were 180° or 90° offset.

Table 2
Mean completion times in seconds and point errors in pixels, both with SD, for the first and second studies.

	Interfaces	T1	T2	
		Time	Time	Error
Study 1	2DSEP	22.8 (7.3)	18.3 (7.3)	41.5 (28.0)
	3DSEP	15.1 (4.5)	16.7 (7.6)	48.1 (26.3)
	3DSEP + L	16.9 (5.6)	20.5 (9.6)	39.4 (21.2)
Study2	3DSEP + L	19.7 (8.1)	25.5 (12.9)	22.5 (9.9)
	3DINP + L	20.8 (7.7)	25.7 (12.9)	23.4 (7.8)

Table 3
Questionnaire data with mean and SD (rounded) for studies with abstract content (1 and 2) and using real world content (3).

	Study 1						Study 2				Study 3			
	T1			T2			T1		T2		T3			
	2DSEP	3DSEP	3DSEP + L	2DSEP	3DSEP	3DSEP + L	3DINP + L	3DSEP + L	3DINP + L	3DSEP + L	MAP	3DTMP	3DINP	3DSEP
Q1	2.8 (1.1)	1.4 (0.5)	1.3 (0.5)	2.2 (0.9)	1.7 (0.8)	1.4 (0.6)	1.7 (0.5)	1.6 (0.5)	1.5 (0.5)	1.8 (0.9)	3.0 (1.2)	1.4 (0.5)	1.6 (0.6)	1.3 (0.5)
Q2	3.4 (1.1)	4.3 (0.8)	4.6 (0.6)	3.8 (1.0)	4.0 (1.1)	4.2 (1.0)	4.0 (1.0)	4.1 (0.9)	4.3 (0.7)	4.3 (0.8)	3.1 (1.2)	4.4 (0.6)	4.3 (1.1)	4.6 (0.6)
Q3	2.3 (1.0)	1.6 (0.5)	1.5 (0.7)	2.0 (0.8)	1.6 (0.7)	1.6 (0.6)	2.1 (0.8)	2.2 (0.8)	1.6 (0.7)	1.6 (0.7)	2.9 (0.9)	1.5 (0.8)	1.3 (0.5)	1.3 (0.5)
Q4	3.1 (0.9)	4.0 (0.9)	4.2 (0.7)	3.8 (0.9)	3.8 (1.1)	4.0 (0.9)	4.1 (0.7)	3.5 (1.2)	4.4 (0.7)	4.1 (0.9)	3.5 (1.1)	4.1 (1.0)	3.9 (0.8)	3.9 (0.9)
Q5	3.7 (0.9)			3.8 (0.8)										
Q6		2.1 (1.4)			2.1 (1.3)		2.6 (1.1)	2.7 (1.0)	2.8 (1.0)	2.4 (1.0)				
Q7		2.8 (1.2)			3.3 (1.2)									
Q8			3.6 (1.3)			3.9 (1.0)	3.6 (1.0)	4.0 (1.0)	4.4 (0.7)	4.3 (0.7)				
Q9			3.0 (1.6)			2.2 (1.4)	2.7 (1.6)	2.3 (1.4)	1.3 (0.7)	1.5 (0.8)				
Q10			1.8 (0.7)			1.9 (1.1)	2.7 (0.9)	2.1 (1.0)	1.6 (0.7)	1.8 (0.6)				
Q11	3.0 (1.4)	4.5 (0.5)	3.8 (1.3)	3.3 (1.3)	4.1 (0.8)	4.3 (1.0)	4.1 (1.1)	4.0 (1.1)	4.5 (0.9)	3.7 (1.0)	2.1 (1.0)	3.9 (1.0)	4.4 (0.6)	3.8 (0.9)

Table 4
Significant effects in questionnaire data. Study 1 was tested with Friedman (not reported) and Wilcoxon signed-rank tests with Bonferroni corrected $\alpha = 0.0167$; Study 2 with Wilcoxon signed-rank tests with $\alpha = 0.05$.

	Task	Study 1			Study 2
		2DSEP&3DSEP	2DSEP&3DSEP + L	3DSEP&3DSEP + L	3DINP + L&3DSEP + L
Q1	T1	$\mathbf{p < 0.001}$	$\mathbf{p < 0.001}$	$p = 0.317$	$p = 0.564$
	T2	$\mathbf{p = 0.011}$	$\mathbf{p = 0.001}$	$p = 0.07$	$p = 0.257$
Q2	T1	$\mathbf{p = 0.002}$	$\mathbf{p = 0.001}$	$p = 0.059$	$p = 0.783$
	T2	$p = 0.285$	$p = 0.026$	$p = 0.096$	$p = 0.564$
Q3	T1	$\mathbf{p = 0.004}$	$\mathbf{p = 0.003}$	$p = 0.527$	$p = 0.317$
	T2	$\mathbf{p = 0.012}$	$p = 0.032$	$p = 0.705$	$p = 1.0$
Q4	T1	$\mathbf{p = 0.002}$	$\mathbf{p < 0.001}$	$p = 0.234$	$p = 0.109$
	T2	$p = 0.666$	$p = 0.119$	$p = 0.238$	$p = 0.102$
Q11	T1	$\mathbf{p = 0.001}$	$p = 0.067$	$p = 0.035$	$p = 0.713$
	T2	$\mathbf{p = 0.006}$	$\mathbf{p = 0.008}$	$p = 0.512$	$\mathbf{p = 0.029}$

During T2, participants either pointed directly at the periphery after finding the sphere, or rotated back to the frontal view. In 3DSEP + L, 54% of the participants used the visual links to highlight the sphere and find it in the spatially separated video image. In 2DSEP, 63% used the small images to quickly find the sphere in the small images. Participants either clicked directly on the corresponding viewpoint, or in some cases (16%) pointed directly at the periphery.

Discussion. In general, both 3D interfaces were preferred over the 2D interface (Q11). Questionnaire data and feedback collected from the participants support this result. For instance, participants perceived that solving the tasks was easier with the 3D interfaces (Q1). The interview also revealed that participants had difficulties with orientation using the discrete image switches in 2DSEP. Participants found this especially challenging in T1, where they had problems keeping track of multiple neighboring spheres. This issue is reflected by the significantly higher confidence (Q2) and intuitiveness (Q3) when using

the 3D interfaces for T1. It may also be the reason why participants rated the presentation of 2DSEP significantly lower only for T1 (Q4), while for T2 there was no significant effect in presentation. Also the effect on intuitiveness (Q3) diminishes in T2, although 3DSEP was still perceived as more intuitive than 2DSEP.

For T1, only 3DSEP was significantly preferred (Q11) over 2DSEP. This is reasonable, given that participants did not require the visual links to solve this task. It is also reflected by Q9, where links were more helpful for T2 than T1. In general, visual links were well received (Q8) and found to be intuitive (Q10). Nevertheless, only 58% of the participants used the links, because, according to their feedback, the task could easily be solved without them. We did not find any significant difference in point error between 3DSEP and 3DSEP + L. However, when dividing the trials of 3DSEP + L and 3DSEP into those with ($n = 68$, mean $= 31.2$, sd $= 14.4$) and those without ($n = 172$, mean $= 48.8$, sd $= 40.58$) visual link usage, the results indicate that participants made less errors when they used links.

The interviews showed that the camera icon was a strong cue for communicating the starting point of rotation. Based on the interviews, we believe that participants were unsure when rating the radar cue, which is also reflected by the trend of neutral answers for the radar cue (Q7). The arrangement of images in 2DSEP was well perceived (Q6). Participants also stated that it provided a good overview of the object in T2, because the single red sphere was very salient.

5.4. Second study: varying spatial separation

Since 3DSEP and 3DSEP + L were the preferred interfaces and both performed better during exploration task (T1), we focused on investigating 3D interfaces further. We kept 3DSEP + L as representative 3D mode, because the visual links showed value as spatial cue in the mapping task (T2). Based on our observations, we introduced a reset button, which automatically realigns copy and context viewpoint. We also removed the radar cue from the interface. Aside from these changes 3DSEP + L corresponded to the same interface as used in the first study (Fig. 5(b)).

In this study, we explored two variations of spatial separation. We created an in-place interface (3DINP + L) that is similar to 3DSEP + L, but which has the copy overlaying the real-world object (Fig. 5(c)). We included the visual links in 3DINP + L, even though their end points are occluded by the 3D copy. Our assumption was that participants would need to switch between AR and VR modes to remove the occlusion and to mentally connect focus and context.

Participants. Twelve participants (6m/6f), aged between 19 and 30 (mean $= 24.7$,sd $= 3.3$), performed 5 repetitions (240 trials) of each task and interface. The presentation order of interfaces and tasks was counterbalanced.

Results. In the analysis, we used the same methods and statistical tests as in the previous study. Time and error measurements are summarized in Table 2. We performed non-parametric tests, because our sample violated normality. Statistical analysis did not reveal any significant effects.

Questionnaires. The questionnaire data is summarized in Table 3, significant effects in Table 4.

Observations. As in previous studies, participants either panned around the object or used a top-down view to solve the tasks. For T2, 92% of the participants used visual links for both interfaces. In 3DSEP + L, 42% switched back to AR to increase the size of the video image. As expected, in 3DINP + L the majority of participants (67%) switched back to AR to resolve occlusions of the link endpoints.

Discussion. Participants generally found the tasks easy to solve (Q1), felt confident with the interfaces (Q2) and found them to be intuitive (Q3). For T2, participants significantly preferred 3DINP + L over 3DSEP + L (Q11). The lack of significance for T1 can be explained by the comments of participants who stated that they only focused on the 3D copy, and did not consider the video background for this task. In the interview, participants stated that they preferred 3DINP + L because it was a more natural and intuitive approach to not separate focus from context. They also mentioned the increased size of the object in 3DINP + L. This is reflected by the higher values in presentation for 3DINP + L (Q4).

Interestingly, the visual links still served as orientation cue in 3DINP + L, even though they penetrated the copy and the endpoints were occluded. Participants noted that visual links showed the misalignment between the copy and the context. As before, trials in which links were used showed smaller point errors ($n = 97$, mean $= 19.6$, sd $= 10.6$) than trials without visual links ($n = 23$, mean $= 36.9$, sd $= 26.4$), which underlines their value as spatial cue.

6. Evaluation: real-world scenario

In the previous studies, we focused on general properties of our interfaces and avoided confounding factors from real world scenes by using only abstract scenarios. In the following study, we introduce the real world into our interface design. We first performed a pilot study in a real world setting to collect qualitative feedback and identify issues with the interfaces. Afterwards, we performed a more thorough and controlled study in our laboratory testbed.

6.1. Pilot study: real-world setting

We conducted a study in a popular urban area of our city center with the 3DINP + L and 3DSEP + L interfaces. Fig. 6 shows the 3DSEP + L interface with one of the target buildings.

Task and methodology. Participants had to find and point to the real world location of a sphere located on the copy of the focus object. Participants were bound to a fixed location, but could rotate with the mobile device (InertiaCube3 sensor). The

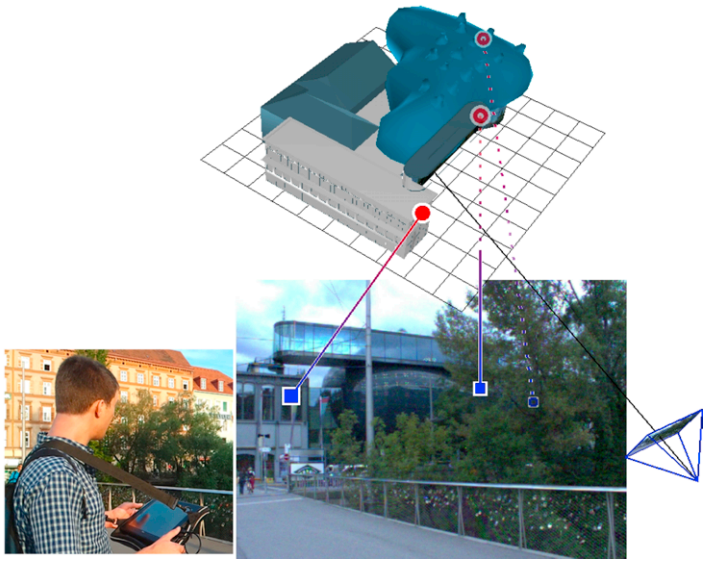

Fig. 6. Pilot study. A spatial technique (3DSEP + L) applied in a real urban environment. The small inset shows a participant using our system.

task was repeated with three visible distinctive cultural buildings located in varying distance around the participant: an art gallery (40 m), a building floating on the river (200 m), and a tower (370 m). Pointing was estimated roughly by visual and verbal assessment. After the experiment, participants completed a questionnaire.

Participants. Ten participants (7m/3f) aged between 16 and 32 (mean = 24.2, sd = 4.3) participated. They were recruited among local pedestrians and familiar with the surroundings.

Discussion. All participants were able to solve the task easily and generally gave positive feedback. All of them could imagine to use such an interface as a tourist, for exploring unknown landmarks and sights (5-point Likert, 1 = strongly agree: mean = 1.3, sd = 0.48). Visual links were regarded useful as orientation cue. In contrast to the previous study, we did not find any significant difference in preference between 3DINP + L (mean = 4.0, sd = 0.82) and 3DSEP + L (mean = 3.8, sd = 1.1). Participants who preferred 3DINP + L again stated that it was more intuitive and natural; the ones who preferred 3DSEP + L stated that it provided a better overview and that the copy was clearly visible due to the spatial separation from the video context. Hence, a main issue seems to be the visual interference of the copy with the real world.

6.2. Experimental design

Although users preferred 3DINP + L in the laboratory setup, we could not reproduce this when deploying the interfaces to a real world setting. According to participants' feedback, the main issue was the visual interference between video background and the 3D copy in 3DINP + L. Therefore, we decided to investigate the influence of different real world scenes on our interface design. We used the apparatus described in Section 5.1.

Condition: interfaces (4). The studied OCE interfaces only differ in terms of the spatial–temporal aspect. We reused the spatially separated 3D interface (3DSEP) without visual links (Fig. 7(c)) and also the in-place interface, from which we removed the links (3DINP) (Fig. 7(b)). We also developed a temporal interface (3DTMP) which similar to 3DINP shows the focus object registered to the real object, but does not preserve the video when switching to the VR mode (Fig. 7(a)). Hence, this interface not only exhibits high visual contrast to the background similar to 3DSEP, but also exhibits the natural behavior of 3DINP and its increased size of the 3D object.

As a baseline condition for our interfaces, we included a 3D map interface (MAP) (Fig. 7(d)). The map shows buildings and terrain without additional contextual information from the real world, such as trees and cars. The view is centered on the location of the user, which is indicated with a blue cylinder, and oriented towards the real focus object. The user can translate, zoom and rotate the view on the map. The rotation and zooming center is defined by the screen center and indicated by a gray cylinder.

We did not include the visual links in this study, because they may be a confounding factor due to the clutter added to the presented real world scenes. Furthermore, in MAP and 3DTMP, the end point of visual links do not connect to a visible real world context.

Condition: scene complexity (3). We prepared three artificial scenes with a 3D city modeling software (Esri CityEngine). Unlike when using realistic pictures created with image-based modeling techniques relying on photographs, this approach allowed us to have control over the presented scenes. It also removed the effect of scene knowledge from the study, because participants were not familiar with the buildings or their locations.

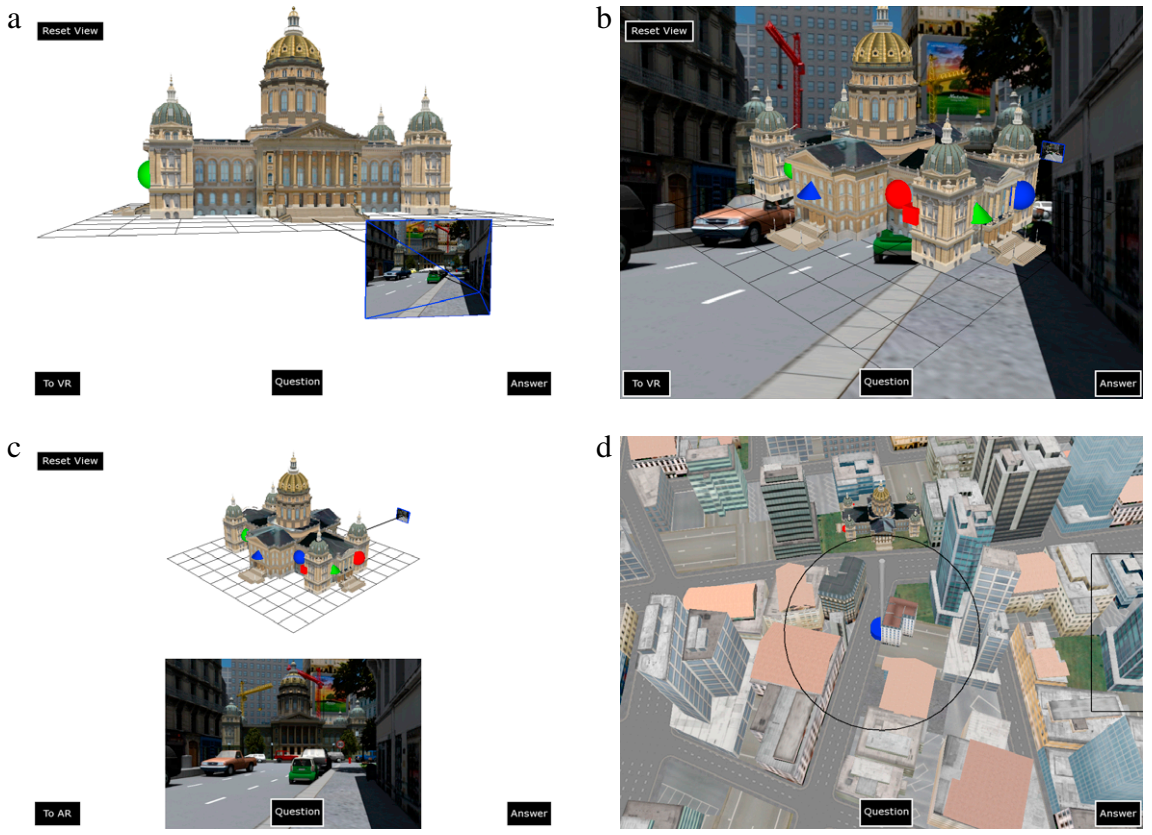

Fig. 7. Interface designs for study in real scenarios. (a) Temporal (3DTMP), (b) in-place (3DINP) and (c) separation interface (3DSEP) applied to the most complex scene. Note the lack of contrast between virtual copy and context in (b) and the reduced size of the virtual copy in (c). (d) A map interface (MAP) applied to the most complex scene. The map is centered on the user's position and oriented in viewing direction. The circle in the center indicates the interaction area for the rotation, the rectangle on the right the one for the zoom. The translation interaction area is located outside of the other areas.

We rendered two views of each scene using a third party software (Lightwave'12): one simulated the periphery, one the AR view. The periphery was rendered using a camera (100° FOV) in a resolution that matched the projection wall. The AR view was rendered with a camera (60° FOV) in a resolution that matched the one of the display devices. All views were taken at the eye level of the participant.

The generated scenes exposed different degrees of real world complexity. In contrast to Lee et al. [30], who define complexity with different levels of visual realism, we define it as the number of unique objects classes, the geometric complexity of the focus buildings and the density of the neighboring buildings. Fig. 8 shows the peripheral views, the outtake for the AR view (red rectangle), and the buildings with an outlined silhouette.

Task. We included the context in the task and asked the participants to find a target object of a certain color and shape, which was close to an object visible in the context (e.g., certain car, street sign, billboard). We used a total of nine target objects: 3 cubes, 3 cones, 3 spheres. Each of the objects of a shape were colored in either red, green or blue. We arranged the shapes at a medium height on the copy of the focus. To force participants to navigate the copy, we placed the objects only sideways and in the rear (Fig. 7). The colors were randomized and the shapes were placed pseudo-randomly such that the answer to the posed question had a clear solution. The position of the object in question was consistent among the trials between participants.

The question was: "Which color does the *shape* closest to the *object* have?" *Shape* refers to one of the 3 shape types, *object* to an object visible in the periphery as well as the AR view. To avoid that participants learn the location of *objects*, we varied their placement in the scene between each interface.

Procedure. The interface condition was counterbalanced, scenes in each interface condition were presented in random order. Before using an interface, participants solved a trial task without time constraint in a trial scene. Participants finished all scenes with an interface, before progressing to the next. The question was shown at the bottom of the screen when starting the task. It disappeared when interaction started and reappeared when the corresponding button was pressed. The task finished when participants selected a color on screen. We recorded task completion time and color error.

When using MAP, the participants had to take the mobile device in their hands to simulate map usage behavior. In the other interface conditions, the device was mounted in front of the participants to simulate an AR view. Participants completed questionnaires after each interface (4) and after finishing the study (1).

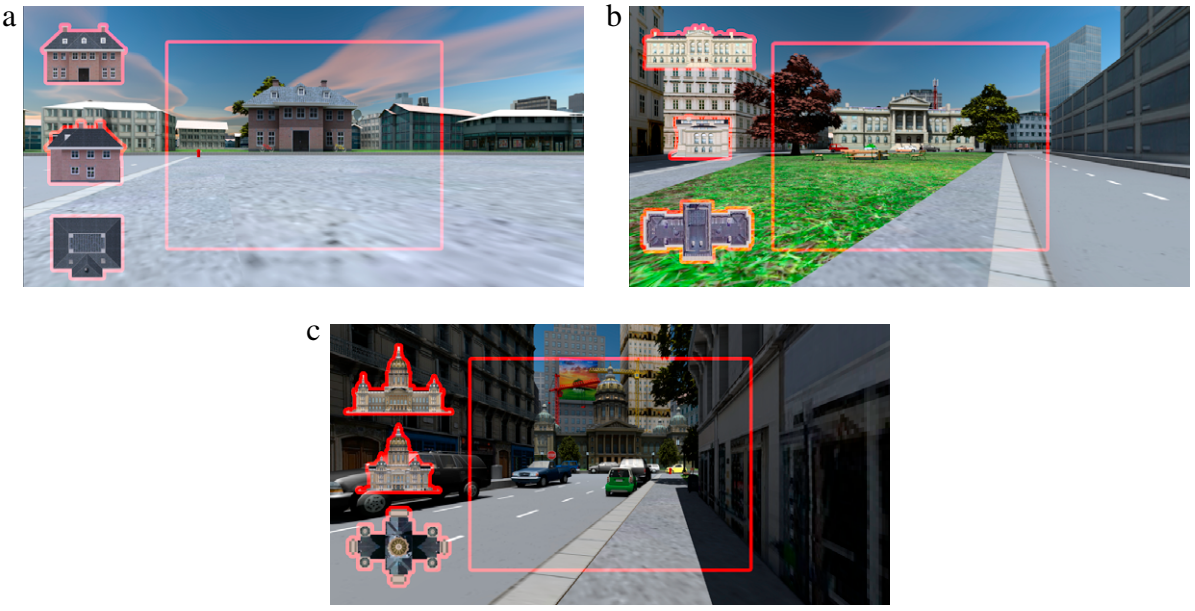

Fig. 8. Scenes with (a) low, (b) medium and (c) high complexity. The estimation of complexity considers the number of unique object classes, the density of neighboring buildings and the geometric complexity of the focus buildings. To the left, the silhouette of the buildings is outlined. The center rectangle is the content visible in the AR view. (Best viewed digitally and zoomed.)

Table 5
Third study. Mean completion times and SD in seconds.

		Low	Medium	High
	Mean (SD)	15.36 (4.43)	18.51 (4.67)	22.57 (5.9)
MAP	30.90 (9.45)	20.69 (7.77)	31.56 (13.85)	38.34 (12.12)
3DTMP	14.92 (3.68)	13.66 (4.51)	14.16 (2.80)	16.43 (4.51)
3DINP	15.67 (6.75)	13.74 (6.95)	14.32 (5.32)	18.48 (9.20)
3DSEP	14.87 (4.56)	13.35 (4.04)	14.05 (4.32)	17.04 (6.36)

Hypotheses. *H1*. Our first hypothesis was that the interfaces outperform each other in terms of task completion time as follows: MAP $<$ 3DTMP $<$ 3DINP $<$ 3DSEP. We considered that MAP is not designed for object-centered exploration and requires the most interaction effort. Furthermore, in MAP the investigated buildings are occluded by neighboring buildings. 3DINP and 3DSEP outperform 3DTMP, because in 3DTMP participants cannot use the video context in the VR mode and must look into the periphery. 3DSEP outperforms 3DINP, because the video is not occluded by the copy. *H2*. Our second hypothesis was that the scene complexity has a negative influence on the task completion time. We believe that more complex scenes lead to higher task completion times.

Participants. The experiment followed a within-participants design, with five repetitions for each interface and task (960 trials). A total of 16 participants (14m/2f), 24–46 years old (mean $=$ 30.06, sd $=$ 5.62), took part in the study. The participants were recruited from university staff and on the campus.

6.3. Results

For each interface $\times$ scene complexity condition and participant, we calculated the mean of time and error from the 5 repetitions. Based on this we calculated the mean of the scene complexity and interface conditions for each participant. The values are summarized in Table 5 and Fig. 9. We performed non-parametric tests, because our sample violated normality. We do not report on the error, because it was practically non-existent.

Interface. A significant effect of interface on time was observed (Friedman, $X^2(3) = 18.075$, $p < 0.001$). A post hoc Wilcoxon signed-rank test with Bonferroni corrected $\alpha = 0.0083$ showed that all AR interfaces were significantly faster than MAP ($p < 0.001$). This significant effect between MAP and AR interfaces was present in each scene complexity condition. For better readability, these test results are presented in Table 6.

Based on these results, we *partially accept H1*. The AR interfaces outperform MAP in terms of task completion times. The performance of the task completion time is also consistent between all scenes and thus applies to different scene complexities.

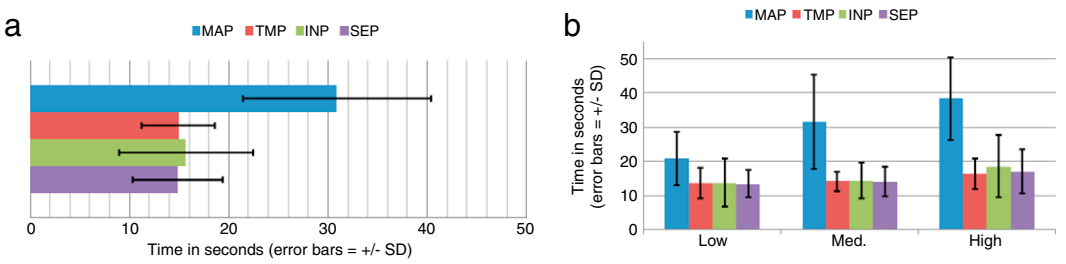

Fig. 9. Third study. Task completion time (a) per interface condition and (b) per interface and scene.

Table 6
Third study. Significant effects between interfaces per scene, tested with Friedman ($\alpha = 0.05$) and Wilcoxon signed-rank tests with Bonferroni corrected $\alpha = 0.0083$.

	Low $X^2(3) = 17.1$ $\mathbf{p < 0.001}$	Med. $X^2(3) = 26.8$ $\mathbf{p < 0.001}$	High $X^2(3) = 29.7$ $\mathbf{p < 0.001}$
MAP&3DTMP	$\mathbf{p = 0.006}$	$\mathbf{p < 0.001}$	$\mathbf{p < 0.001}$
MAP&3DINP	$\mathbf{p = 0.001}$	$\mathbf{p = 0.001}$	$\mathbf{p < 0.001}$
MAP&3DSEP	$\mathbf{p = 0.002}$	$\mathbf{p < 0.001}$	$\mathbf{p < 0.001}$
3DTMP&3DINP	$p = 0.836$	$p = 0.918$	$p = 0.535$
3DTMP&3DSEP	$p = 0.796$	$p = 1.0$	$p = 0.836$
3DINP&3DSEP	$p = 0.959$	$p = 0.642$	$p = 0.569$

Table 7
Third study. Significant effects between scenes (Low = L, Med. = M, High = H) per interface tested with Friedman ($\alpha = 0.05$) and Wilcoxon signed-rank tests with Bonferroni corrected $\alpha = 0.0167$.

		L&M	L&H	M&H
MAP	$X^2(2) = 22.875$, $\mathbf{p < 0.001}$	$\mathbf{p = 0.001}$	$\mathbf{p = 0.001}$	$p = 0.026$
3DTMP	$X^2(2) = 14$, $\mathbf{p < 0.05}$	$p = 0.148$	$\mathbf{p = 0.001}$	$\mathbf{p = 0.003}$
3DINP	$X^2(2) = 16.625$, $\mathbf{p < 0.001}$	$p = 0.163$	$\mathbf{p = 0.001}$	$\mathbf{p = 0.005}$
3DSEP	$X^2(2) = 7.875$, $\mathbf{p < 0.05}$	$p = 0.5$	$\mathbf{p = 0.003}$	$\mathbf{p = 0.007}$

Table 8
Third study. Significant effects in questionnaire data. The data was tested with Friedman (not reported) and Wilcoxon signed-rank tests with Bonferroni corrected $\alpha = 0.0083$.

	Q1	Q2	Q3	Q4	Q11
MAP&3DTMP	$\mathbf{p = 0.001}$	$\mathbf{p = 0.003}$	$\mathbf{p = 0.001}$	$p = 0.119$	$\mathbf{p = 0.002}$
MAP&3DINP	$\mathbf{p = 0.003}$	$p = 0.018$	$\mathbf{p = 0.001}$	$p = 0.185$	$\mathbf{p = 0.001}$
MAP&3DSEP	$\mathbf{p = 0.001}$	$\mathbf{p = 0.002}$	$\mathbf{p < 0.001}$	$p = 0.142$	$\mathbf{p = 0.001}$
3DTMP&3DINP	$p = 0.317$	$p = 0.739$	$p = 0.527$	$p = 0.559$	$p = 0.070$
3DTMP&3DSEP	$p = 0.414$	$p = 0.157$	$p = 0.206$	$p = 0.518$	$p = 0.869$
3DINP&3DSEP	$p = 0.096$	$p = 0.206$	$p = 0.655$	$p = 0.952$	$p = 0.078$

Scene. A Friedman test between scene complexities revealed a significant effect on completion time ($X^2(2) = 22.875$, $p < 0.001$). A post hoc Wilcoxon signed-rank test with Bonferroni corrected $\alpha = 0.0167$ showed significant differences between scenes A and B ($p = 0.003$), A and C ($p = 0.001$) and B and C ($p = 0.002$). For better readability, the significant effects between the scenes of each interface are summarized in Table 7.

Based on these results we *accept H2*. Scene complexity had an overall negative impact on the task completion time of all interfaces.

Questionnaires. The questionnaire data is summarized in Table 3, significant effects in Table 8. Participants were asked to rank the scenes according to their perceived complexity, and rated the scenes as follows: high as high (100%), medium as medium (94%), low as low (94%). One participant switched medium and low.

6.4. Discussion

Participants significantly preferred the AR interfaces over MAP (Q11) and all AR interfaces performed better than a traditional 3D map interface. Even 3DTMP that did not preserve the video context in the VR mode performed better, which indicates that the transition between AR and VR and the simple camera cue are sufficiently strong cues for connecting real and virtual worlds. A major limiting factor of MAP was occlusions from neighboring buildings in scenes of higher complexity.

This was also supported by corresponding statements in the interview. Generally, the task was significantly easier to solve with the AR interfaces (Q1). Another factor which might have negatively influenced MAP performance is that the objects referred to in the task were only visible in the periphery and not in the map itself. For this reason, in MAP, participants were forced to redirect their views between the mobile device and the periphery, while in the AR conditions the relevant objects were presented in the context displayed on the mobile device either sequentially (3DTMP) or parallelly (3DINP, 3DSEP).

Contrary to what we expected, we did not find any differences between the AR interfaces. Participants could quickly find the queried object in the context at the beginning of the task. In the VR mode, participants could always look into the periphery by glancing past the mobile device mounted in front of them. This may have been sufficient to also achieve good performance in 3DTMP, where no context was available after switching to the VR mode. When using an HMD without the option of looking at the periphery directly after switching to the VR mode, 3DTMP may perform differently.

The participants' comments regarding the interfaces were in line with those of previous studies. Participants noted the good visibility of the copy in 3DTMP and 3DSEP and the intuitive arrangement of focus and context in 3DINP. Although there is no significant preference (Q11) between the AR interfaces, there is again a trend towards preferring an in-place interface (Q11 in Tables 3 and 8).

Q3 revealed that the AR interfaces were significantly more intuitive than MAP. One reason for this rating might be that the interface was not well suited to the task of object-centric exploration. Another reason may be that MAP used a single touch interface instead of a more common multi-touch interface. A multi-touch interface might improve the intuitiveness and even the performance of MAP, but will still exhibit problems with occluding structures. Hence, we are confident that our results also hold up against MAP interfaces available on current mobile devices.

The performance and questionnaire data confirmed our estimation of the scene complexity. Performance generally degraded with increasing scene complexity, and the ranking of scene complexity by the participant was in line with our estimation. The impact on performance can be attributed to different factors in the AR and MAP conditions. In MAP, our observations and the interviews revealed that the density of neighboring buildings had a major impact on performance. In the AR conditions, the decrease in performance can mainly be attributed to the geometric complexity of the focus object. In the AR conditions, there was only a significant effect between the scene with highest complexity and all other scenes (Table 7). In contrast to the buildings in the other scenes, the building in the most complex scene exhibited concavities, which occluded the queried shapes and thus required more navigational effort.

The map interface can be classified within the frame of our design space. It is a temporal interface, which does not provide any strong cues to resolve the viewpoint discontinuity between the egocentric viewpoint of the user and the exocentric map view. There is no seamless transition between these views, but similar to the camera icon in the AR interfaces, the position indicator of the map interface represents the current viewpoint of the user. The map interface exhibits a large context discontinuity, because the real world context is replaced with only a virtual representation showing buildings and terrain. In comparison, the context in the three AR interfaces shows only an egocentric 2D view, but is richer in information because it contains details such as cars or street signs.

7. Design recommendations

In the following, we put forward design recommendations for OCE techniques based on the findings of the previous studies. We also outline potential future research directions. The recommendations are structured based on the aspects outlined in the design space presented in Section 3.

Spatial–temporal aspect. We did not find performance differences between the in-place and the spatially separated interface. However, under controlled laboratory conditions, participants significantly preferred the in-place interface (3DINP + L), because the arrangement of focus and context was more natural. Occlusion of the context by the copy did not seem to be an issue. In the real world setting, we did not find a significant preference, because participants also preferred 3DSEP + L, because of the higher contrast between the copy and the white background. Therefore, an in-place interface may have to adapt the *context* to always achieve a good contrast to the overlaid focus object (e.g., desaturation of video background).

Generally, participants used the interfaces to get a quick *overview* of the focus object. Using 3D interfaces, participants switched to a top-down view to quickly look at the different sides of the object. In 2DSEP, participants used the small images as multi-perspective visualization to quickly identify the view containing the queried red sphere in T2. Therefore, OCE techniques should offer modes to explicitly get an overview over an object. When using images as overview, the relevant items on the object should be emphasized in an authoring step beforehand (e.g., by labels), due to the small size of the images, especially on mobile phones.

Input control. An overview can easily provide *shortcut navigation* to quickly access viewpoints. In 2DSEP, participants used the small images to quickly navigate between viewpoints in non-sequential order by accessing 90° and 180° offset views. This is also a main motivation for similar interfaces, such as SnapAR [18] or the one of Veas et al. [17].

Based on our experience with MAP, a simplified *camera navigation* model with few degrees of freedom (e.g., orbit metaphor) was sufficient for the investigated structures. However, future designs should also consider zooming and the exploration of more complex structures, which require more sophisticated navigation metaphors. For instance, the HoverCam [32] allows to explore complex objects with few degrees of freedom.

Virtual copy. Our findings are in line with Bowman et al. [19], who found out that instant teleportation causes disorientation. In our studies, continuous 3D viewpoint changes outperformed discrete 2D switches. Therefore, a 3D interface is the most sensible choice for presenting viewpoint changes to the user.

Spatial cues. When designing our OCE interfaces we focused on the presentation of a single point of interest and provided only limited contextual information through the video background (3DSEP, 3DINP). In one design we even removed the context completely and presented only the virtual copy and the camera icon (3DTMP). Participants could easily solve the given tasks with 3DTMP, the most basic OCE interface design. All OCE interfaces also outperformed the map interface, which also showed structures of surrounding buildings and thus provided more contextual information about surrounding structures. Hence, the *transition* between AR and VR modes and the *camera icon* seems to be sufficiently strong cues to connect separated views and address spatial, viewpoint and context discontinuity. The radar icon on the other hand was not considered helpful. This indicates that cues should be connected directly to the spatial reference frame of the copy.

Based on the collected data and participants' feedback, we can say that *visual links* are a valuable spatial cue. We designed visual links with spatial separation between focus and context in mind (3DSEP + L). However, participants considered spatial separation as inferior to a more natural in-place interface due to the smaller size of the zoomed focus object. Hence, the links could be redesigned to better support in-place interfaces. On the other hand, we observed that participants mainly used links during the mapping task (T2). An intermediate mode could be introduced that switches from in-place to a spatially separated presentation, when users want to map information back to the real world.

Although we tested our initial designs in a laboratory setting, the experiences gathered from the real world pilot study make us confident that OCE interfaces are also feasible and practical in real world conditions. Future work will investigate the interfaces in real world conditions in more depth. Our current designs only considered the ideal *object distance* to the real world object. Future designs will investigate situations, where the interface needs to zoom the focus object, because it is either too close or too distant.

8. Conclusion

In this paper, we presented a design space for OCE interfaces, which allows users to explore distant real world objects. Relevant application areas for the presented techniques are mobile tourism and urban planning. They can also provide guidance for supporting the exploration of real objects in the next generation of AR browsers.

We explored different designs in abstract settings and compared their performance in real world conditions with a more common 3D map interface. All of our OCE interfaces, even in a basic design, outperformed the map interface. However, OCE interfaces can easily be combined with 2D or 3D maps to facilitate accessing contextual map information about real world structures. While this does not solve the problem of occluding structures in 3D maps, it may facilitate mentally linking the real world with the spatial map representation.

Based on our findings we provide a set of design recommendations, which we hope will inspire other researchers to explore the design of OCE techniques further.

Acknowledgments

This work was funded by the Christian Doppler Laboratory for Handheld Augmented Reality, the Austrian Science Fund (FWF) under contract P-24021 and the FP7-ICT EU Project No. 601139 CultAR.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.pmcj.2014.08.010.

References

[1] W. Schwinger, C. Grün, B. Pröll, W. Retschitzegger, A. Schauerhuber, Context-awareness in mobile tourism guides—a comprehensive survey, Rapport Technique. Johannes Kepler University Linz.
[2] M. Baldauf, P. Fröhlich, K. Masuch, T. Grechenig, Comparing viewing and filtering techniques for mobile urban exploration, J. Locat. Based Serv. 5 (1) (2011) 38–57.
[3] A. Oulasvirta, S. Estlander, A. Nurminen, Embodied interaction with a 3D versus 2D mobile map, Pers. Ubiquitous Comput. 13 (4) (2009) 303–320. http://dx.doi.org/10.1007/s00779-008-0209-0. URL: http://dx.doi.org/10.1007/s00779-008-0209-0.
[4] J.S. Pierce, B.C. Stearns, R. Pausch, Voodoo Dolls: seamless interaction at multiple scales in virtual environments, in: I3D, 1999, pp. 141–145. http://dx.doi.org/10.1145/300523.300540. URL: http://doi.acm.org/10.1145/300523.300540.
[5] A. Cockburn, A. Karlson, B.B. Bederson, A review of overview+detail, zooming, and focus+context interfaces, ACM Comput. Surv. 41 (1) (2009) 2:1–2:31. http://dx.doi.org/10.1145/1456650.1456652. URL: http://doi.acm.org/10.1145/1456650.1456652.
[6] T.N. Hoang, B.H. Thomas, Augmented viewport: an action at a distance technique for outdoor AR using distant and zoom lens cameras, in: ISWC, 2010, pp. 1–4. http://dx.doi.org/10.1109/ISWC.2010.5665865.
[7] R. Stoakley, M.J. Conway, R. Pausch, Virtual reality on a WIM: interactive Worlds in miniature, in: CHI, 1995, pp. 265–272. http://dx.doi.org/10.1145/223904.223938. URL: http://dx.doi.org/10.1145/223904.223938.
[8] B. Bell, T. Höllerer, S. Feiner, An annotated situation-awareness aid for augmented reality, in: UIST, 2002, p. 213. http://dx.doi.org/10.1145/571985.572017. URL: http://portal.acm.org/citation.cfm?id=572017&CFID=111526848&CFTOKEN=81088077.
[9] R. Bane, T. Höllerer, Interactive tools for virtual X-ray vision in mobile augmented reality, in: ISMAR, 2004, pp. 231–239. http://dx.doi.org/10.1109/ISMAR.2004.36.

[10] J. Keil, M. Zöllner, M. Becker, F. Wientapper, T. Engelke, H. Wuest, The house of Olbrich—an augmented reality tour through architectural history, in: ISMAR AMH, 2011, pp. 15–18.
[11] C. Bichlmeier, S.M. Heining, M. Rustaee, N. Navab, Laparoscopic virtual mirror for understanding vessel structure evaluation study by twelve surgeons, in: ISMAR, 2007, pp. 1–4. URL: http://portal.acm.org/citation.cfm?id=1514348.
[12] C.E. Au, V. Ng, J.J. Clark, Mirrormap: augmenting 2D mobile maps with virtual mirrors, in: MobileHCI, 2011, pp. 255–264. http://dx.doi.org/10.1145/2037373.2037413. URL: http://doi.acm.org/10.1145/2037373.2037413.
[13] A. Mulloni, H. Seichter, A. Dünser, P. Baudisch, D. Schmalstieg, 360° panoramic overviews for location-based services, in: CHI, 2012, pp. 2565–2568. http://dx.doi.org/10.1145/2207676.2208645. URL: http://doi.acm.org/10.1145/2207676.2208645.
[14] A. Mulloni, A. Duenser, D. Schmalstieg, Zooming interfaces for augmented reality browsers, in: MobileHCI, 2010, pp. 161–169. URL: http://data.icg.tugraz.at/~dieter/publications/Schmalstieg_192.pdf.
[15] C. Sandor, A. Cunningham, U. Eck, D. Urquhart, G. Jarvis, A. Dey, S. Barbier, M. Marner, S. Rhee, Egocentric space-distorting visualizations for rapid environment exploration in mobile mixed reality, in: VR, 2010, pp. 47–50.
[16] E. Veas, R. Grasset, E. Kruijff, D. Schmalstieg, Extended overview techniques for outdoor augmented reality, IEEE Trans. Vis. Comput. Graphics 18 (4) (2012) 565–572. http://dx.doi.org/10.1109/TVCG.2012.44. URL: http://dx.doi.org/10.1109/TVCG.2012.44.
[17] E. Veas, A. Mulloni, E. Kruijff, H. Regenbrecht, D. Schmalstieg, Techniques for view transition in multi-camera outdoor environments, in: GI, 2010, pp. 193–200. URL: http://dl.acm.org/citation.cfm?id=1839214.1839248.
[18] M. Sukan, S. Feiner, B. Tversky, S. Energin, Quick viewpoint switching for manipulating virtual objects in hand-held augmented reality using stored snapshots, in: ISMAR, IEEE Computer Society, 2012, pp. 217–226.
[19] D.A. Bowman, D. Koller, L.F. Hodges, Travel in immersive virtual environments: an evaluation of viewpoint motion control techniques, in: VRAIS'97, 1997, pp. 45–52. URL: http://dl.acm.org/citation.cfm?id=523977.836072.
[20] K. Kiyokawa, H. Takemura, N. Yokoya, A collaboration support technique by integrating a shared virtual reality and a shared augmented reality, in: SMC, 1999, pp. 48–53. http://dx.doi.org/10.1109/ICSMC.1999.816444.
[21] M. Billinghurst, H. Kato, I. Poupyrev, The MagicBook—moving seamlessly between reality and virtuality, IEEE Comput. Graph. Appl. 21 (3) (2001) 6–8. http://dx.doi.org/10.1109/38.920621.
[22] B. Avery, C. Sandor, B.H. Thomas, Improving spatial perception for augmented reality X-ray vision, in: VR, 2009, pp. 79–82. http://dx.doi.org/10.1109/VR.2009.4811002.
[23] D.A. Bowman, E. Kruijff, J.J. LaViola, I. Poupyrev, 3D User Interfaces: Theory and Practice, Addison Wesley Pub. Co., Inc., 2004.
[24] D.D. Woods, Visual momentum: a concept to improve the cognitive coupling of person and computer, Int. J. Man–Mach. Stud. 21 (3) (1984) 229–244. http://dx.doi.org/10.1016/S0020-7373(84)80043-7. URL: http://dx.doi.org/10.1016/S0020-7373(84)80043-7.
[25] G. Fitzmaurice, J. Matejka, I. Mordatch, A. Khan, G. Kurtenbach, Safe 3D navigation, in: I3D'08, ACM, 2008, pp. 7–15. http://dx.doi.org/10.1145/1342250.1342252. URL: http://doi.acm.org/10.1145/1342250.1342252.
[26] M. Trapp, C. Beesk, S. Pasewaldt, J. Döllner, Interactive rendering techniques for highlighting in 3D geovirtual environments, in: 3D GeoInfo Conf., 2010.
[27] R. Grasset, M. Billinghurst, A. Dünser, Moving between contexts—a user evaluation of a transitional interface, in: ICAT, 2008.
[28] M. Tatzgern, R. Grasset, E. Veas, D. Kalkofen, H. Seichter, D. Schmalstieg, Exploring distant objects with augmented reality, in: JVRC'13, 2013, pp. 49–56. http://dx.doi.org/10.2312/EGVE.JVRC13.049-056.
[29] C. Lee, S. Bonebrake, T. Hollerer, D. Bowman, The role of latency in the validity of ar simulation, in: VR, 2010, pp. 11–18. http://dx.doi.org/10.1109/VR.2010.5444820.
[30] C. Lee, G.A. Rincon, G. Meyer, H. Tobias, D. Bowman, The effects of visual realism on search tasks in mixed reality simulation, in: VR, 2013.
[31] D. Baricevic, C. Lee, M. Turk, H. Tobias, D. Bowman, Hand-held AR magic lenses with user-perspective rendering, in: ISMAR, 2012.
[32] A. Khan, B. Komalo, J. Stam, G. Fitzmaurice, G. Kurtenbach, HoverCam: interactive 3D navigation for proximal object inspection, in: I3D, 2005, pp. 73–80. http://dx.doi.org/10.1145/1053427.1053439. URL: http://doi.acm.org/10.1145/1053427.1053439.