Interacting with Computers 24 (2012) 450–460

Contents lists available at SciVerse ScienceDirect

Interacting with Computers

journal homepage: www.elsevier.com/locate/intcom

Analytic review of usability evaluation in *ISMAR*

Zhen Bai *, Alan F. Blackwell

Computer Laboratory, University of Cambridge, William Gates Building, 15 JJ Thomson Avenue, Cambridge CB3 0FD, United Kingdom

ARTICLE INFO

Article history:
Received 22 October 2011
Received in revised form 25 July 2012
Accepted 25 July 2012
Available online 2 August 2012

Keywords:
Augmented Reality
Usability evaluation
User experience
Collaboration

ABSTRACT

There has been a rapid increase in research evaluating usability of Augmented Reality (AR) systems in recent years. Although many different styles of evaluation are used, there is no clear consensus on the most relevant approaches. We report a review of papers published in International Symposium of Mixed and Augmented Reality (*ISMAR*) proceedings in the past decade, building on the previous work of Swan and Gabbard (2005). Firstly, we investigate the evaluation goal, measurement and method of *ISMAR* papers according to their usability research in four categories: performance, perception and cognition, collaboration and User Experience (UX). Secondly, we consider the balance of evaluation approaches with regard to empirical–analytical, quantitative–qualitative and participant demographics. Finally we identify potential emphases for usability study of AR systems in the future. These analyses provide a reference point for current evaluation techniques, trends and challenges, which benefit researchers intending to design, conduct and interpret usability evaluations for future AR systems.

1. Introduction

Augmented Reality refers to a class of technologies that superimpose virtual information over a view of the physical world. It has been applied in various domains such as manufacturing, medicine, maintenance, education and entertainment. By extending users' knowledge space without switching between real and virtual contexts, AR provides distinct user interfaces (UIs) and interactions from conventional computer systems. In Azuma's survey (Azuma, 1997), he described the three main characteristics of AR as: (1) combines real and virtual; (2) interactive in real time; (3) registered in 3D space. Early research in Mixed and Augmented Reality focused on technical factors such as display, tracking, rendering, calibration and 3D reconstruction. These factors have been thoroughly studied and in many cases resolved by the constant efforts in advanced computer vision and graphics approaches as well as the increase of hardware capacity (e.g. powerful computation, high resolution display and more accurate sensors) and the appearance of new technologies (e.g. Microsoft Kinect and high end mobile handheld).

As the base technologies have become more mature and robust, research priorities have shifted toward the design of effective and usable applications. Since 1998, the International Symposium of Mixed and Augmented Reality (*ISMAR)* (including forerunner events *IWAR*/*ISAR* and *ISMR*) has become the leading international conference exclusively focused on AR. As the leader in this field, it provides an annual snapshot of state-of-the-art in AR technologies and applications. *ISMAR* is a highly selective and high impact conference, such that the range of topics addressed in papers accepted for presentation can be used as a sample reflecting the research concerns considered of most relevance by leaders in the AR field. Of particular interest to our own concerns, if we consider the proportion of published *ISMAR* papers that have included usability evaluations over the past decade (Fig. 1), we see that whereas in 2001 only one out of 19 papers reported a user evaluation, there has been a steady increase since then, reaching a peak of 44% in 2008.

Usability studies can help to identify design flaws in application concepts at earlier phases of development. In addition, by gathering user feedback, researchers can improve their understanding of the mental models (Carroll, 2003) by which users will interpret novel AR spaces. This can help AR researchers to understand users' real needs and expectations, suggest specific improvements, and inform design guidelines. However user benefits must ultimately be assessed in the context of an actual AR application design, rather than individual technical components. These factors motivate our detailed review of recent developments in usability evaluation.

We build on some relevant surveys that have previously been reported within the AR field. Swan and Gabbard (2005) reviewed user-based experiments in AR publications between 1992 and 2004, drawing from *ISMAR*, *International Symposium on Wearable Computers*, *IEEE Virtual Reality* and *Presence*. They identified three types of "experiment" relevant to AR research concerns: "human perception and cognition", "user task performance and interaction techniques", and "user interaction and communication between

* Corresponding author. Tel.: +44 1223763626.
E-mail addresses: zhen.bai@cl.cam.ac.uk (Z. Bai), alan.blackwell@cl.cam.ac.uk (A.F. Blackwell).

0953-5438/$ - see front matter © 2012 British Informatics Society Limited. All rights reserved.
http://dx.doi.org/10.1016/j.intcom.2012.07.004

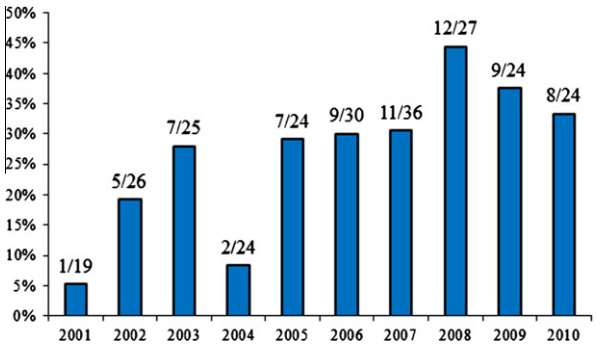

Fig. 1. *ISMAR* publications with usability evaluation.

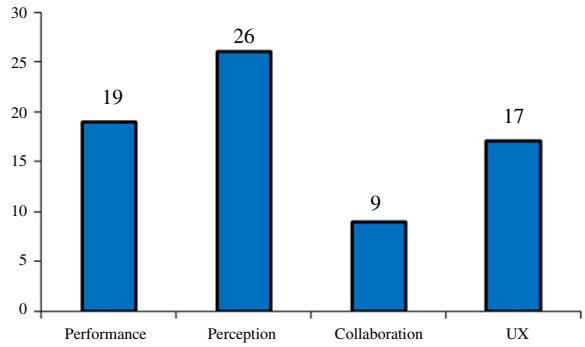

Fig. 2. Number of *ISMAR* publications in four usability evaluation focus categories.

multiple collaborating users". They found 14 HCI-related *ISMAR* publications, nine of which reported empirical user experiments. In earlier research (Gabbard and Hix, 2001), Gabbard and colleagues summarized usability design and evaluation research in Virtual Environments (VEs) and AR. This resulted in a set of guidelines covering: user and user tasks; the virtual model; user interface input mechanisms and user interface presentation components. Swan and Gabbard subsequently proposed an iterative usability engineering model (Gabbard and Swan, 2008) based on their earlier model for VE (Gabbard et al., 1999), which offered design feedback and guidelines in the absence of established AR principles or interaction metaphors.

Zhou et al. (2008) conducted a complementary analysis of technology trends in *ISMAR* from 1998 to 2007. Although their research focus was on AR technologies, they also pointed out the significance of usability evaluation for interaction and user interface design. Meanwhile, Dünser and his colleagues reported a survey (Dünser et al., 2008) of all AR evaluation techniques between the years 1993 and 2007, including a comprehensive reference list and a comparison to results in Swan and Gabbard (2005). However, due to the large sample size, this paper only focused on overall trend analysis in AR evaluation, rather than a detailed critique of evaluation strategies, or of problematic issues in evaluation.

In our paper we focus on three aspects: Firstly, we investigate the evaluation goal, measurement and method of *ISMAR* papers according to their usability research in four categories: performance, perception and cognition, collaboration and User Experience (UX). Secondly, we consider the balance of evaluation approaches in regard to empirical–analytical, quantitative–qualitative and participant composition. Thirdly we identify potential emphases for usability study of AR systems in the future under each usability research category. These analyses provide a reference point for current evaluation techniques and for future improvements of usability evaluation specific to the AR domain, such as error tolerance and representative task design; individual differences; distinguish usability issues between collaboration and non-collaboration performance; separate UX issues caused by technology limitations, pragmatic and holistic design.

2. Method

We conducted a detailed review of papers published in *ISMAR* proceedings from 2001 to 2010, analyzing every paper in this period that included a usability evaluation. The total size of this sample is 71 papers. Our review builds on that by Swan and Gabbard (2005). We used their threefold categorisation as our starting point, but found it necessary to add the further category of User Experience (UX).

We defined UX as addressing subjective user issues, such as technology preference, affect, perceptual and physical experiences. This is in accordance with a recent comprehensive survey of user experience research (Hassenzahl et al., 2009) that validates the ISO definition: "a person's perceptions and responses that result from the use or anticipated use of a product, system or service" (ISO DIS 9241-210, 2008). UX is, however, less likely to be investigated using controlled experimental methods, which may account for its omission from Swan and Gabbard's classification of experimental research.

In our review of usability evaluation at *ISMAR*, we found two groups of papers addressing UX. In the first, the primary evaluation focus was on UX itself, with the goal of assessing users' attitude to, and acceptance of, the system. In the second, the main evaluation focus was on perception, performance or collaboration, with UX evaluation as a supplementary measurement.

A broad definition of UX might include "perception" as one aspect of human experience, but we have chosen to separate the well-established tradition of *ISMAR* studies where human performance on a perceptual task is measured as an engineering factor within overall system performance, rather than as a component of subjective user experience. We adopt the following categories of evaluation focus in *ISMAR* papers:

- *Task performance*: study the user accomplishments of application tasks or interactions (Swan and Gabbard, 2005).
- *Perception and cognition*: study low-level tasks that assess human perception/cognition in an AR environment (Swan and Gabbard, 2005).
- *Collaboration*: study user interaction related to communication between multiple collaborating users (Swan and Gabbard, 2005).
- *UX*: study users' subjective feelings and experiences.

We assign each paper to a single category according to its primary focus. Fig. 2 shows the overall proportion of the sample assigned to each category.

3. Review of usability evaluation in *ISMAR*

In this section, we present detailed analysis of papers published in *ISMAR* proceedings from 2001 to 2010. We group the papers according to the four usability evaluation focuses discussed in Section 2: task performance, perception and cognition, collaboration and user experience.

3.1. Task performance

We found 19 papers that evaluated task performance in specific application domains (e.g. maintenance, manufacturing, medicine). Time (100% of papers) and accuracy (63%) are the most common measures for performance evaluation. In addition, 84% of these pa-

Table 1
Summary of performance-focused usability evaluation.

Application	Reference	Performance evaluation goal	Performance measures	UX factors	UX measures
Driving	Bubb et al., 2005	Compare two visualization schemes to guide driver's attention of danger	Response ***time***, ***error*** quotient, average mistake, average lane deviation	***Preference, ease of use, perceived performance***	Questionnaire
	Klinker and Tönnis, 2006	Compare four visualization schemes to guide driver's attention of danger	Speed, lane deviation, response ***time***, ***error*** quotient, average mistake	Preference, ease of use, perceived performance	Questionnaire
	Klinker et al., 2007	Compare two visualization schemes for longitudinal and lateral driver assistance	Speed deviation, average speed difference, lane deviation, lane departure ***time***, ***time*** of line crossing	Task load, ***perceived performance***, concentration	NASA TLX, Questionnaire
Maintenance	Feiner and Henderson, 2009	Compare maintenance performance between AR, HUD and LCD	***Time*** of completion, ***error***	***Ease of use***, ***intuitiveness***, musculoskeletal workload, strain, satisfaction	Questionnaire, head movement
Manufacturing	Böckelmann et al., 2009	Compare performance and stress between paper, 2D and AR technologies in order picking process	***Time*** of completion, ***error***	Strain	Physiological measures, Questionnaire
	Zäh and Vogl, 2006	Compare industrial robot programming between traditional teach-in and AR	***Time*** of completion, ***accuracy***	N/A	N/A
	Schwerdtfeger and Klinker, 2008	Compare three visualization schemes to guide the picking target	***Time, error***	N/A	N/A
Medicine	Bichlmeier et al., 2007	Compare performance with and without an AR medical navigational tool	***Time, error***	***Preference***, user attitude, task realization	Questionnaire
	Bichlmeier et al., 2010	Compare performance with and without an AR medical navigational tool	***Accuracy***: number of collisions, path length, depth motion, trial number, duration of collisions, completion ***time***	User attitude	Questionnaire
	Quarles et al., 2008	Understand an Anaesthesia Machine in an AR environment	Level of understanding	***Preference***, confidence, usefulness	Questionnaire
Tangible	Billinghurst et al., 2004	Compare performance between immersive and non-immersive tangible AR authoring tools	***Time, error***	***Preference***	Questionnaire
	Oh and Hua, 2006	Compare aspect ratio and size of the tangible Magic Lens in searching and path following in a map navigation scenario	***Time***, scale factor, eye-hand-table distance	***Preference***	Questionnaire
	Fjeld et al., 2002	Compare a tangible AR tool with 2D and 3D tools for spatial planning and layout	***Time***, number of operations, time per trial	***Ease of use***, clarity of task explanation, suitability	Questionnaire
	Feng et al., 2009	Compare seven presentation methods for visual hint of tangible Shake Menus	***Time, error***	***Preference, ease of use, intuitiveness,*** satisfaction	Questionnaire
Visualization	Feiner and Güven, 2006	Compare three visualization methods in recognizing occluded objects	***Time, accuracy***	***Preference, ease of use, intuitiveness***	Questionnaire
	Goto et al., 2010	Compare six visualization methods for an AR task support system	***Time***, number of repeated times to play the instructional video	***Ease of use***, suitability	Questionnaire
	Petersen and Stricker, 2009	Compare performance between AR, mouse and GUI in the Continuous Natural UI	Time	***Preference, ease of use, intuitiveness,*** naturalness	Questionnaire
Others	Thomas et al., 2006	Compare three input techniques: handheld, head cursor and image-plane vision-tracked device for selection and annotation	***Time, error***	***Preference, ease of use, perceived performance,*** fatigue	Questionnaire
	Tang et al., 2003	Compare accuracy of four variants of the SPAAM calibration method	***Time, error***	N/A	N/A

pers also included some UX evaluation, typically by questionnaire or physiological monitoring. Table 1 reports the application domain, performance evaluation goal, performance measurements, UX evaluation factors and UX evaluation measurements.

We found two types of controlled performance comparison, according to the evaluation goal of the research. For papers aiming to demonstrate that an AR system would improve user task performance relative to a conventional system (Feiner and Henderson, 2009; Böckelmann et al., 2009; Zäh and Vogl, 2006; Billinghurst et al., 2004; Fjeld et al., 2002; Petersen and Stricker, 2009), the conventional system was used as an experimental benchmark. For papers that compare alternative AR design solutions (Bubb et al., 2005; Klinker and Tönnis, 2006; Klinker et al., 2007; Klinker and Schwerdtfeger, 2008; Schwerdtfeger and Klinker, 2008; Bichlmeier et al., 2007, 2010; Oh and Hua, 2006; Feng et al., 2009; Feiner and Güven, 2006; Goto et al., 2010; Thomas et al., 2006; Tang et al., 2003), all measurements are made in an AR environment, in order to understand the relative advantages and disadvantages of each solution. One exception was a pre/post comparison of user understanding after using an AR training system (Quarles et al., 2008), which made no comparison of that system to conventional training or alternative designs.

In most evaluations, time and error/accuracy were used as a measure of satisfactory task completion. Some studies also aimed to assess specific aspects of human factors in task performance such as: cognitive load (Klinker et al., 2007), judgment reliability (Bubb et al., 2005; Klinker and Tönnis, 2006), distraction (Bubb et al., 2005), cognitive support (Fjeld et al., 2002), learning effect (Fjeld et al., 2002), longitudinal/lateral driving behavior (Klinker et al., 2007) and navigation behavior (Oh and Hua, 2006).

Table 2
Summary of perception-focused usability evaluation.

Type	Reference	Perception measures
Audio	Lindeman et al., 2007	Sound localization ***accuracy***
	Higa et al., 2007	Sound localization ***accuracy***, sound quality
Depth and occlusion	Liu et al., 2008	Perceived depth, accommodative, dioptre response
	Zhang and Hua, 2010	Perceived distance
	Livingston et al., 2003	Perceived target position, response ***time***, ***error***
	Furmanski et al., 2002	Perceived depth
	Avery et al., 2008	***Time***, ***accuracy***, design feedback, physical strain (***UX***), user attitude (***UX***)
	Sandor et al., 2010	***Time***, perceived background and foreground difference
Display	Livingston et al., 2006	***Accuracy*** of wave orientation, response ***time***, perceived color
	Wither et al., 2007	***Time*** of completion (cursor movement, visual search)
	Kiyokawa et al., 2007	User attitude (***UX***)
	Grasset et al., 2007	Handheld picking behavior, ***time*** of completion (follow and count object)
	Blum et al., 2010	View quality
	Livingston et al., 2009	Binocular disparity
	Billinghurst et al., 2003	Perception of real world, virtual objects, occlusion effects and virtual object touch, comfort of HMD (***UX***), amusement (***UX***)
Layout	Tanaka et al., 2008	Number of reactions to displayed message
	Peterson et al., 2008	Response ***time***, ***error rate***, timeout rate, perceived overlay jitter, marker visibility, stereo fusion, depth segregation, magnitude of depth segregation, headache (***UX***), eye strain (***UX***), neck pain(***UX***)
	Azuma and Furmanski, 2003	Response ***times***, ***error***
Registration error	Robertson et al., 2009	***Time***, ***accuracy***, confidence (***UX***), work load (***UX***)
	Robertson et al., 2008	Placing attempt number, ***time*** of completion, ***error***, work load (***UX***)
	Livingston et al., 2008	Response ***time***, follow target performance, ***error***, subjective difficulty (***UX***)
Imperceptible marker	Grundhöfer et al., 2007	Perceptual discomfort for integrated code intensity
3D presentation	Belcher et al., 2003	Response ***time***, ***error***, ease of use (***UX***), perceived performance (***UX***)
Shadow	Sugano et al., 2003	Virtual object presence, shadow realism, light position and shadow shape, response ***time***, frequency of head movements
Stiffness	Knorlein et al., 2009	Perceived stiffness
Presence	Alvarex et al., 2010	Perceived presence, work load (***UX***)

Most experiments relied on a single representative task. However there were two exceptions. In Feiner and Henderson (2009), to evaluate user performance with a mechanical maintenance system, participants completed 18 typical maintenance tasks under three display conditions. In Thomas et al. (2006), three AR input techniques were compared using six selection tasks and six annotation tasks.

Among studies with a primary focus on task performance, the most frequent additional UX evaluation factors were: user preference, subjective ease of use, perceived performance and intuitiveness. The validity of intuitiveness as a guide for user interface design is controversial, since it depends on users' previous experience with similar systems, resulting in a bias against more innovative solutions (Raskin, 1994). A variety of other UX factors include perception of performance aspects such as attention and task load (in driving tasks), and more general experience measures such as user satisfaction, user attitude, task suitability and user fatigue.

Assessment of UX usually involves completion of attitudinal and/or affective questionnaire items. However some studies also use physiological measurements of task strain, including heart rate, electrocardiogram (ECG), galvanic skin response (GSR) and skin temperature (Böckelmann et al., 2009), or head movement as a measure of musculoskeletal strain (Feiner and Henderson, 2009).

3.2. Perception and cognition

We found 26 papers with an emphasis on evaluating human perception and cognition issues in Augmented Reality. This continues to be a significant research topic because the perceptual and cognitive demands of combining virtual information with perception of three-dimensional physical space have little similarity to traditional WIMP (window, icon, menu, and pointing device) interaction.

Table 2 summarizes the types of perception performance that have been evaluated in *ISMAR* proceedings.

From the table we can see that depth and occlusion perception, display technology, virtual information layout, audio modality, and registration error were the most frequent types of user perception in an AR environment. Other than the specific perception aspects being investigated, time (52%) and accuracy (44%) are the two most common measurements.

The display conditions and application scenarios in these studies are specific and heterogeneous, meaning that these evaluation results cannot be utilized directly to guide the design of AR applications for general purposes. It is typically left to the readers to attempt to interpret specific findings as design guidelines to quickly decide what technologies and configurations are favorable to support appropriate perception for specific tasks. AR application designers would benefit from future papers providing specific design guidance as part of the analysis and conclusion.

3.3. Collaboration

We found 9 papers that conducted usability evaluations related to collaboration. Table 3 gives a detailed summary of the evaluation factors measured for collaboration applications.

According to mechanics of collaboration proposed by Pinelle et al. (2003), measurements of particular interest in the evaluation of AR collaborative systems include information gathering (*IG*) (basic awareness, eye gaze/contact), explicit communication (*EC*) (spoken and gestural messages) and ease of collaboration (*EOC*). For most measurements, subjective answers were collected via questionnaire. For eye gaze and contact, objective results were extracted from direct observation. From a UX perspective, researchers also paid attention to signs of discomfort and enjoyment during collaboration.

Table 3
Summary of collaboration-focused usability evaluation.

References	Collaboration measures
Grasset et al., 2005	*TP*: Time of completion, error, length of path; *IG*: head movement, ***awareness***; *EC*: ***ease of use*** (communication)
Kiyokawa et al., 2002	*TP*: Time of completion; *IG*: number of trainees' looking away, head angular velocity, ***awareness***, fluency of focus; *EC*: number of extra pointing gestures, average number of phrases
Tateno et al., 2005	*IG*: Gaze ***awareness***, perceived ***gaze*** direction error
Takemura et al., 2006	Perceived allowable facial color range and allowable range of color differences; discomfort (***UX***)
Billingshurst et al., 2005	EOC: ***Ease of collaboration***; IG: ***awareness*** of partner; enjoyment (***UX***)
Nilsson et al., 2010	*EOC*: ***Ease of use*** (cooperation, information mediation, and situational picture access), suitability, visualization, trust, confidence (intrapersonal and interpersonal); design preference (***UX***), physical discomfort (***UX***) and learnability, enjoyment (***UX***)
Prytz et al., 2010	IG: Number of **eye contacts**
Benko et al., 2004	User/expert attitude (***UX***), feedback (***UX***)
Oda and Feiner, 2009	*EOC*: Game duration, distance between players, effectiveness, distractibility

Table 4
Summary of UX-focused formal usability evaluation.

Domain	References	UX factors	UX measures
Manufacturing	Tümler et al., 2008	Strain, discomfort, wellbeing	Physiological measurement, questionnaire
Remote surveillance	Kiyokawa et al., 2006	User attitude of operability, searchability	Questionnaire
Tangible	Feiner et al., 2007	Preference, ease of understanding	Questionnaire Observation
	Anabuki and Ishii, 2007	Ease of use, attitudes of reversibility and scalability	Questionnaire Observation
Handheld	Veas and Kruijff, 2008	Attitude to hardware (weight, grip, material, fatigue), functional design (effectiveness, ease of use)	Questionnaire Observation
Entertainment	Jones et al., 2010	Ease of use, perceived speed (constructing surface/mapping content) and input device controls	Questionnaire

As to task performance (*TP*) during the collaboration, only two papers (Grasset et al., 2005; Kiyokawa et al., 2002) measured task completion time. This suggests that while the design of usable application and technologies are still the main focus of the AR collaboration research, more effort could be exerted in demonstrating the effectiveness of collaborative AR systems.

3.4. User Experience

We found 17 papers that only evaluated UX-related aspects of the AR system, without any objective task-related performance assessments. UX evaluations were mostly preliminary and included both formal and informal studies. Formal evaluations involved controlled experiments with a fixed sample of volunteer users and collected participants' experiences with structured surveys/questionnaires. Informal evaluations involved unstructured interviews or observations with a casual sample of potential users or domain experts.

3.5. Formal UX evaluation

A summary of UX factors and measurements are shown in Table 4.

By comparing those studies that only conducted UX evaluations and those that evaluated UX in addition to another main focus (performance, perception or collaboration) we can see that the two classes share one goal, which is to collect participants' opinions of the proposed system. However we might ask "why is it that only UX factors were measured in these formal experiments?" In Tümler et al. (2008), the goal of the study was specifically to evaluate the difference in stress between AR and non-AR shopping scenarios, so subjective stress level was measured, but not task performance. In Feiner et al. (2007), seven alternative gesture hints for tangible manipulation were compared. The main evaluation goal was comprehension of the hint, rather than time to complete the gesture, so this was assessed using a questionnaire. Jones et al. (2010) used a miniature golf game to study interactive scenes rendered on everyday surfaces. Since the objective of the game is entertainment, a questionnaire was carried out to evaluate subjective experience rather than task performance. Anabuki and Ishii (2007) evaluated a system for 3D free-form modeling in an AR environment, in which the creative modeling experience could be evaluated using a post-questionnaire and direct observation. In addition to the relevance of subjective criteria for evaluation, these systems were at a preliminary design stage, meaning that it was too early to assess users' performance in a defined task.

Interestingly in both Veas and Kruijff (2008) and Kiyokawa et al. (2006), the evaluation methods were task-oriented (operation and object search in Kiyokawa et al. (2006) and object selection and placing in Veas and Kruijff (2008)). However no quantitative performance data was collected, but only opinions from the participant.

3.6. Informal UX evaluation

The summary of papers with only informal evaluations is shown in Table 5.

Informal UX evaluations intend to gather rapid feedback on users' attitudes to proposed AR applications. Both interviews and observation were involved, with domain expert opinions highlighted in five out of eleven papers. Relevant domain experts included civil engineers (Schall et al., 2008), automobile maintenance staff (Platonov et al., 2006), product designers (Klinker et al., 2002), clinicians (Kotranza et al., 2009) and viticulturists (King et al., 2006).

Comparing the evaluation goal between performance-focused and UX-focused evaluations we can observe that if the develop-

Table 5
Summary of UX-focused informal usability evaluation.

Domain	References	UX factors
Manufacturing	Klinker et al., 2002	User/expert attitude, feedback
	Wang et al., 2005	User attitude
Medicine	Kotranza et al., 2009	User/expert attitude, feedback
Painting	Bandyopadhyay et al., 2001	User attitude
Civil engineering	Schall et al., 2008	User/expert attitude, feedback
Maintenance	Platonov et al., 2006	User/expert attitude
Oil/gas field exploration	Gordon et al., 2002	User attitude
Museum guide	Miyashita et al., 2008	User attitude
Viticulture	King et al., 2006	Expert attitude
Entertainment	Gandy et al., 2005	User attitude
	MacWilliams et al., 2003	User attitude

ment stage of the AR system was earlier concept prototyping, researchers focused more on rapid feedback from users and domain experts about general design and experience. Performance issues, meanwhile, start to be considered and measured when the prototypes are closer to applications for real scenarios. The combination of performance and UX evaluations give both objective and subjective results of participants using the system, which provide holistic information to demonstrate its advantages and disadvantages in both pragmatic and hedonic aspects.

4. Evaluation approaches and issues

A recent review of evaluation approaches in the CHI literature (Barkhuus and Rode, 2007) divided these along two dimensions: empirical vs. analytical and quantitative vs. qualitative. Empirical methods typically require a group of potential users to participate in evaluation while analytical methods require only a smaller group of expert analysts, as in Cognitive Walkthrough, Heuristic Evaluation and GOMS. Quantitative evaluations analyze numeric data with statistical approaches to characterize a sample that reflects the usability needs of the entire potential user group (e.g. experimental performance measures and questionnaires). Qualitative evaluation aims at gathering narrative data about users' subjective experiences or their behavior while using the system (e.g. interview, observation, open-ended questions). Following this classification, we divided the papers in our sample into the following groups: quantitative empirical, qualitative empirical, quantitative and qualitative empirical, analytical, and informal (evaluations conducted in a non-controlled experiment without a pre-defined structure, e.g. randomly selected users were asked how they like the system (Barkhuus and Rode, 2007). Fig. 3 represents the trend of the application of evaluation methods among *ISMAR* papers over our sample period.

4.1. Empirical vs. analytical

From Fig. 3 we can observe a significant imbalance between empirical and analytical evaluation. Empirical evaluation is by far the most popular evaluation approach. Both objective and subjective factors of users were measured during empirical experiments. Objective factors were primarily measured in a task-driven manner. The tasks are either designed expressly for perception evaluation, or functional tasks related to a specific application. Time and error rate were commonly used in both performance and perception focused research. Physiological evaluations were also used to indicate participants' subjective feelings such as stress and anxiety.

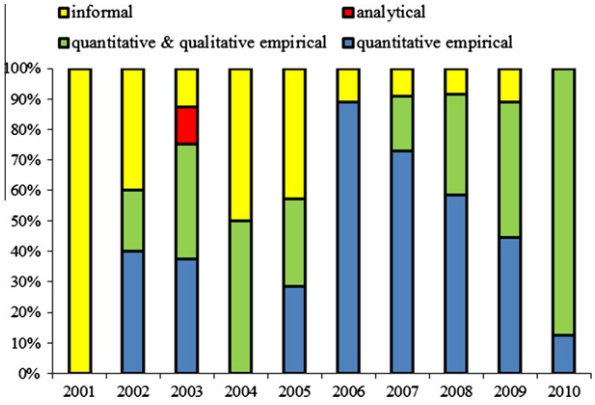

Fig. 3. Proportion of *ISMAR* publications according to their usability evaluation approaches.

As discussed in Section 3.1, preference, ease of use, perceived performance and intuitiveness were commonly recorded as representations of users' subjective experiences.

For analytical evaluation, we found only two related papers in the *ISMAR* proceedings. Furmanski et al. (2002) reviewed cognitive principles of visual perception and proposed design heuristics for Obscured Information Visualization. Livingston et al. in their research on occlusion visualization (Livingston et al., 2003) later conducted an expert heuristic evaluation although the heuristics were not reported. Given the generally exploratory nature of academic research, and its focus on novel outcomes, in some cases informal analytical evaluation, particularly relating to prototypes, may have been performed but not discussed. Further researchers beyond these two groups may have engaged in some informal analytical analysis, but only these two chose to report them.

Barkhuus and Rode's study of CHI (Barkhuus and Rode, 2007) similarly found that analytical evaluations were not widely applied between 1982 and 2006. Their view was that as practitioners played a less influential role in the CHI community, the analytical evaluation techniques that were popular in industry became regarded as somewhat un-scientific. However, we believe that the overlooking of analytical evaluation in AR has more complicated causes. While usability evaluation methods have evolved side-by-side with technologies in traditional HCI, hardware to support AR has only become widespread in the past fifteen years, after analytic HCI research was already waning. AR is expected to deliver a novel user experience, in which "learning by trying" (Bach and Scapin, 2004) is assumed necessary, making task-based usability evaluations particularly difficult.

If we follow this logic, then it might be suggested that analytical evaluations are not widely used because there is little expert knowledge of AR application domains. Or to be specific, as (Gabbard and Swan, 2008) suggested, emerging technologies like Augmented Reality "have no established design guidelines or interaction metaphors, or introduce completely new ways for users to perceive and interact with technology and the world around them". They therefore suggested that experts should first evaluate candidate user interfaces based on basic user interface or interaction designs, then supply both design feedback and potential user-based experimental factors. They emphasized the importance of user-based studies in driving design activities because in the long term those design suggestions examined during user-based studies will gradually contribute to the adopted design principle pool.

Some AR design principles can already be applied to avoid fundamental design flaws before researchers conduct rather expensive user-based experiments. Besides the earlier mentioned VE/AR usability design guidelines surveyed by Gabbard and Hix (2001), Dünser et al. (2007) proposed the application of generic

HCI principles to AR system design, and suggested a non-exhaustive list of guidelines as: affordance, reducing cognitive overhead, low physical effort, learnability, user satisfaction, flexibility in use, responsiveness and feedback, and error tolerance. In addition, Kruijff et al. (2010) reviewed major AR perception issues in environment, capturing, augmentation, display and users, and summarized the corresponding mitigation approaches proposed by current research.

Gray and Salzman (1998) suggested that the fundamental differences between analytical and empirical evaluation methods are in the way that analytical evaluation methods "examine intrinsic features and attempt to make predictions concerning payoff performance" while empirical evaluation methods "attempt to measure payoff performance directly (e.g. speed number, error, learning time)." Analytical evaluation helps the designers to fully understand the intrinsic features of their systems, after which they can design proper experiments to examine such features. Intrinsic features need not be holistic but should be explicit. Thus instead of automatically measuring common payoffs such as time and error, researchers and designers are able to choose better validated measurements for each aspect of interest.

4.2. Quantitative vs. qualitative

There are two intriguing results found from the comparison between quantitative and qualitative empirical evaluation methods. First, there was no purely qualitative empirical evaluation. In other words, among papers conducting formal evaluations, all qualitative analysis was accompanied by quantitative evaluations. Qualitative evaluations often include observation of users' behavior and interviews about system design and user experience. The former is an effective way to get access to users' mental models and identify potential ergonomic issues. It is commonly used in AR collaboration (Billingshurst et al., 2005; Nilsson et al., 2010; Grasset et al., 2005; Kiyokawa et al., 2002). Many other publications reported observations (Billinghurst et al., 2003; Anabuki and Ishii, 2007; Schwerdtfeger and Klinker, 2008; Feng et al., 2009; Jones et al., 2010), and some with informal evaluations (Miyashita et al., 2008; Gandy et al., 2005). Interviews were either conducted alone or as a supplementary method with questionnaires.

Second, the proportion of papers including both qualitative and quantitative evaluations has risen significantly between 2007 and 2010. With regard to the nature of qualitative evaluation, there are both advantageous and disadvantageous aspects of this trend. If the purpose of qualitative evaluations is to obtain user-centered information as a supplement to task performance or assessment of perception, then it is definitely beneficial since such evaluations provide more comprehensive insight of users' external (from behavior) and internal (from interview) reaction in an AR environment. However, if the purpose of the qualitative evaluations is to explore potential design spaces and understand the mental model of a particular task context, then quick prototype mockup with in-depth observation might be more productive than the extra development effort required to conduct a full quantitative evaluation.

4.3. User sample validity

We looked into the gender balance and professional background of user samples for usability evaluation in *ISMAR*, as shown in Fig. 4a.

With regard to gender, we can see that the proportion of evaluations not reporting the gender balance has dropped steadily. We also observe the rough trend that the proportion of male participants remains high while the proportion of females does not expand over time. There is gender imbalance in some specific domains, such as police and military (Nilsson et al., 2010), however

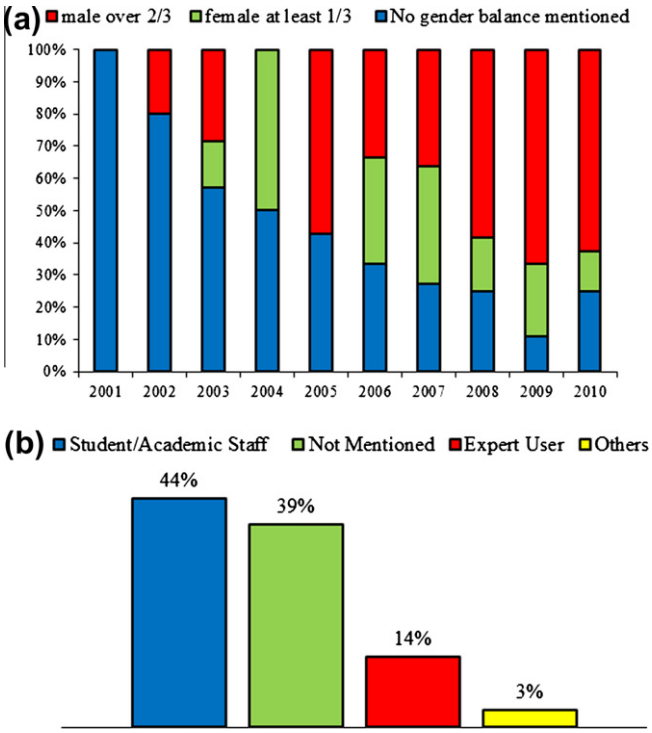

Fig. 4. (a) Gender balance. (b) Participants source of formal empirical experiments.

the majority of AR scenarios are designed for the general population with an equal gender balance. Thus a constant trend of gender imbalance in usability studies will affect the external validity in the long term.

External validity might also be affected by reliance on participant recruitment from student and academia as observed in Barkhuus and Rode's study (Barkhuus and Rode, 2007). From Fig. 4b we can see that among papers conducting formal evaluations, 44% of the experiments were based on university students and academic staff and another 39% did not mention the source of the subjects. Heavy computer experience, better learning ability and academic background bring bias to the evaluation results. We recommend that more care be taken in recruiting participants who are representative of the intended user population. Participants from an exclusively academic background may not affect experiment results for investigations of low-level perception, but it is likely to influence results for task-oriented and UX related studies. In any case, AR researchers should describe the participant recruiting process so that readers of their work can be aware of any potential bias.

5. Evaluation challenges

In this section we identify observed challenges in performance, perception, collaboration and UX evaluations and discuss potential improvements in usability evaluations of AR systems. Although these challenges are only a subset of issues that may affect the way AR researchers interpret their usability evaluation results, we would like to draw their attention towards potential factors that explain the unique complexities and challenges in evaluating AR systems.

5.1. Performance evaluation challenges

We identified two major challenges in task performance evaluations: tolerance of AR systems to user error, and design of representative tasks for experiments.

Error tolerance – "how well the design prevents errors, or helps with recovery from those that do occur" (Quesenbery, 2003) is an important usability dimension. However, our review found that tolerance to errors has rarely been examined. Meanwhile, the likelihood of human errors (Dünser et al., 2007) in AR environments is high, because users are generally unfamiliar with such environments, and metaphors for interaction with objects in AR environments are still under development. With increased likelihood of error, the ability to recover from mistakes is essential for building a usable system and a pleasant user experience.

Designing representative tasks is another challenge. As discussed in Section 3.1, only two studies (Feiner and Henderson, 2009; Thomas et al., 2006) evaluated performance in more than two tasks. The majority only focused on a single task. For research primarily concerned with AR technology, it is understandable that a single task was used. However researchers might also have benefitted from considering other tasks associated with the same application, to explore the impact of the technology. For example, if a novel interaction method is demonstrated to outperform others in a map navigation task, it is helpful to also consider performance in related scenarios such as search, zoom and labeling. Furthermore, for evaluations of performance differences between AR-based and traditional systems in an application domain, evaluating only a single task lacks external validity.

5.2. Perception evaluation challenges

Both individual differences and less controlled outdoor experiments raise challenges for evaluating perception in AR.

Perception is affected by both physical and psychological states of individuals (Davidoff, 1975). However, we found that few visual perception evaluations (e.g. depth and occlusion) measured visual acuity of participants, while no auditory acuity was reported in auditory perception evaluation. Furthermore, only a few experiments verified participants' stereo perception (Peterson et al., 2008; Blum et al., 2010; Livingston et al., 2009) and color vision capacities (Sandor et al., 2010). Researchers need to be very careful making perception suggestions out of the experiment results, without considering the subject's perception ability of the physical world.

Another issue of visual perception occurs in outdoor AR evaluations (Livingston and Ai, 2008; Tanaka et al., 2008; Peterson et al., 2008; Avery et al., 2008; Sandor et al., 2010). Outdoor illuminance is the foremost variable in such evaluations and it significantly affects participants' visual perception of both virtual objects and the physical environment according to different display conditions. Some experiments (Livingston et al., 2008; Sandor et al., 2010) reported the weather condition and its influence on subjects, but more detailed information that reflects precise illuminance status (e.g. lux) was not collected. Other factors in an outdoor environment such as background noise, shadow and wind are easier to control by carefully choosing the experiment site. As research interest grows in outdoor AR, these factors will need to be measured and better understood to control their impacts on the experiment results.

5.3. Collaboration evaluation challenges

An immediate challenge when evaluating a collaborative system is to separate those usability issues that are specifically related to the collaboration of multiple users, from general usability issues that will also affect a single user. Collaborative systems suffer from both kinds of problem, meaning that it can be hard to isolate the factors that are specifically relevant to collaboration. Furthermore, measures of collaborative performance often embody theories of social interaction that may problematise the straightforward application of findings beyond the experimental context, or even carry implicit critique of the research enterprise. Finally, existing AR collaboration systems are still relatively simple, and are yet to uncover the evaluation challenges that realistically complex scenarios will face.

As Lindgaard and Parush (2008) pointed out, a collaborative system should first support single user interaction, and then multiple users' interaction. In a shared workspace, individual's confidence about how explicitly communication is delivered largely depends on his or her own perception of the quality of the communication method. For example, in the simplest whiteboard brainstorm scenario, the user subconsciously inspects whether letters on the whiteboard are readable to others and the markers of important concepts stand out in a clear way. Likewise, users' communication experience towards others is directly affected by individual perception of the AR collaboration environment. Latency, display distortion, registration error, just to name a few, are system quality issues that while independent of the collaboration interaction design, are very likely to affect a user's performance in the evaluated collaborative scenario.

AR collaboration systems are still in their infancy, with current research focused on exploring the potential of AR for multi-user tasks in a shared workspace. Current evaluations emphasize display affordance, collaboration behavior in AR environment, spatial relationship of users, domain applications, and gaming. As deeper insight is obtained into the affordances of AR collaboration, more complex activities should be supported, such as virtual object manipulation, real-time annotation and remote collaboration. Consequent usability issues will include the movement of virtual objects and physical tools between people, manipulation coordination and protection of individual work in a collaboration workspace (Pinelle et al., 2003).

5.4. UX evaluation challenges

Fig. 5 illustrates three tiers of factors that influence users' subjective response to an AR system. Many users are initially impressed by the unique user experience of AR. However they may also be surprised by its technological limitations such as rendering latency, intermittent tracking failure, registration error and perception issues that they would not normally experience in a conventional interaction environment. Such limitations, then, constitute the initial challenge of UX evaluation.

A pragmatic design focus for UX evaluation can be divided into two levels: general attitude to an early concept prototype, and ease of use for an application with core functions fulfilled. The challenge of the former is to gather just enough information about users' impressions of the suitability and meaningfulness of the AR concept. Feedback from such an evaluation is of course only a preliminary indication of potential user acceptance. The latter, ease of use, covers experiences of functionality, such as perceived effectiveness and efficiency, compatibility with users' mental models and ergonomics of physical apparatus.

Our survey did not find any unified strategy for function-related UX evaluation since applications are built for various purposes.

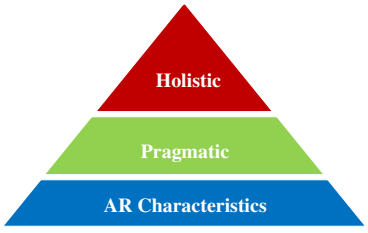

Fig. 5. Dependence relationship of AR user experience.

Generally, though, there are two factors that affect users' impressions at this pragmatic level: the quality of the underlying technologies and the functional design of the application. Unfortunately there is a lack of established metrics indicating the acceptance range of fundamental AR technologies. It is therefore difficult to determine whether the user's experiences, either positive or negative, are caused by the immaturity of specific technologies, or by ill-designed application concepts or processes. As already noted, AR system performance issues such as tracking failure rate, jittering, rendering latency, registration offset and so on are rarely reported as factors in user studies (although relevant performance parameters may have been discussed in a technical system description). System performance also varies across applications due to display condition (optical vs. video), physical environment (indoor or outdoor) and interaction context (stationary or dynamic), so acceptance metrics vary accordingly. This is a challenge for interpretation of UX evaluations, and a potential bottleneck for both technology development and design.

Learnability is another challenge of UX evaluation at the pragmatic design level. The novelty of AR requires users to interact with a system that is neither like virtual or real interaction, based on limited prior experience. Consequently, learnability becomes an indivisible part of the ease of use of an AR system. In conventional systems, it is assumed that users are familiar with WIMP interaction methods. The same cannot be assumed of AR interaction methods, and so learnability begins with the ease with which a user realizes how to interact with virtual and real artifacts in an unfamiliar AR environment. In most experiments or observations, such self-exploratory learning processes were substituted with careful supervision and letting the participant try out the system multiple times to build confidence prior to the actual experiment. This is an efficient way to carry out evaluation in a controlled environment, but it is unlikely to be the case in a natural environment without the presence of the researcher. The balance between nurturing users' interaction behavior with specially designed learning cues, and catering for users' natural mental models is critical and mostly relies on the feedback from UX evaluations on learnability.

Finally, the holistic UX derives from both the subjective feelings of the pragmatic design and a combination of non-pragmatic factors including affect, physical comfort, aesthetic appreciation, enjoyment and perceived value of the system. Most holistic UX evaluations focus on momentary or periodical experiences while the long-term influences of AR systems on an individual's physical and psychological state remain unexplored. Questions like '*will day-to-day working experience in an AR environment affect users' vision capacity*', '*will long-term engagement in an immersive AR game affect teenagers' perception of real world and social behavior'* and *'will students become more proactive or passive with learning under an AR education program'* all require long-term UX study, thus this becomes another challenge that requires constant effort from multiple research disciplines.

6. Conclusions

We have conducted a comprehensive review of the leading AR conference *ISMAR* over the past ten years, using this sample to reflect current concerns at the core of the AR research community. Our review identified 71 papers that reported usability evaluations. We found four evaluation focuses: task performance, perception, collaboration and user experience. In each of these categories, we have provided a comprehensive comparison of the evaluation goals, measurements and methods that have been applied.

Second, we have characterized the evaluation approaches according to the nature of the study (empirical or analytical) and the data collected (quantitative or qualitative). We observed that analytical methods are rarely applied, largely because of the lack of established AR design guidelines and principles. We did find, however, some available guidelines that could profitably be applied during early design and inspection phases. We have also drawn attention to the benefits of analysis – avoidance of fundamental design flaws and guiding the design of user-based experiments.

Thirdly, we have reported trends in empirical studies. The proportion of papers including both quantitative and qualitative evaluations has risen within the last 4 years – we note both advantages and disadvantages of that result. We have also drawn attention to potential external validity concerns that are raised by gender imbalance and reliance on student participants in *ISMAR* evaluation studies.

Finally we have identified a number of special challenges for usability evaluation of AR systems, as revealed by our methodological review within each of the four evaluation focuses. We outlined several emphases in usability evaluations with AR systems including: tolerance of error and representative task design in the performance evaluation; identifying individual differences and outdoor condition analysis in the perception evaluation; distinguishing usability issues between the design of multiple users collaboration versus the system performance with a single user in the collaboration evaluation; recognizing UX issues caused by technology limitations (rendering latency, intermittent tracking failure, registration error), pragmatic design and holistic design.

Our methodological decision to focus on a single conference does, of course, mean that these findings should be treated with some caution. AR research is published in an extremely diverse range of venues, including some specialist HCI contexts (which will naturally be more sophisticated in their consideration of user issues), and also publications in general computing or popular science contexts (which are likely to be more speculative with regard to the implications of AR technology for users). Given this diversity, a fully comprehensive survey of recent AR research publications would be unlikely to provide such clear results as our study, even if it were practically feasible. Nevertheless, we believe that the advantage of focusing on a single venue has been that this meeting can be taken as a proxy for the community of practice at the center of AR research. Those who publish regularly at *ISMAR* are the current and future leaders of the field, and the review criteria applied in selecting the 20% of submissions to be published reflect the current concerns of the community with regard to its research priorities. On this basis, we argue that a comprehensive review of a single publication venue can provide higher value than a less tightly controlled sample of a broader range of research venues. Nevertheless, our results should be interpreted in this light, and should not be taken, for example, to represent specialist HCI research into AR technologies.

Overall, these findings will both provide researchers with a timely overview of the present situation for usability evaluation within the AR community, and also a reference point for future development of strategies specific to the AR domain.

Acknowledgements

Zhen Bai would like to thank Raymond and Helen Kwok and the Cambridge Overseas Trust for sponsoring her research.

References

Alvarex, C., Catrambone, R., Davidson, B., Eiriksdottir, E., Gandy, M., Hillimire, M., MacIntyre, B., McLaughlin, A.C., 2010. Experiences with an AR evaluation test bed: presence, performance, and physiological measurement. In: Proceedings of the Ninth IEEE International Symposium on Mixed and Augmented Reality, 13–16 October, 2010, Seoul, Korean, pp. 127–136.

Anabuki, M., Ishii, H., 2007. AR-Jig: a handheld tangible user interface for modification of 3D digital form via 2D physical curve. In: Proceedings of the Sixth IEEE International Symposium on Mixed and Augmented Reality, 13–16 November, 2007, Nara, Japan, pp. 55–66.

Avery, B., Thomas, B.H., Piekarski, W., 2008. User evaluation of see-through vision for mobile outdoor augmented reality. In: Proceedings of the Seventh IEEE International Symposium on Mixed and Augmented Reality, 15–18 September, 2008, Cambridge, UK, pp. 69–72.

Azuma, R.T., 1997. A survey of augmented reality. Presence: Teleoperators and Virtual Environments 6 (4), 355–385.

Azuma, R., Furmanski, C., 2003. Evaluating label placement for augmented reality view management. In: Proceedings of the Second IEEE International Symposium on Mixed and Augmented Reality, 7–10 October, 2003, Tokyo, Japan, pp. 66–75.

Bach, C., Scapin, D.L., 2004. Obstacles and perspectives for evaluating mixed reality systems usability. In: Proceedings of the IUI-CADUI Workshop on Exploring the Design and Engineering of Mixed Reality Systems, 13 January, 2004, Funchal, Island of Madeira.

Bandyopadhyay, D., Raskar, R., Fuchs, H., 2001. Dynamic Shader lamps: painting on movable objects. In: Proceedings of the IEEE and ACM International Symposium on Augmented Reality, 29–30 October, 2001, New York, USA, pp. 207–216.

Barkhuus, L., Rode, J.A., 2007. From mice to men – 24 years of evaluation in from mice to men – 24 years of evaluation in CHI. Proceedings of CHI 2007, vol. 28. San Jose, USA.

Belcher, D., Billinghurst, M., Hayes, S.E., Stiles, R., 2003. Using augmented reality for visualizing complex graphs in three dimensions. In: Proceedings of the Second IEEE International Symposium on Mixed and Augmented Reality, 7–10 October, 2003, Tokyo, Japan, pp. 84–93.

Benko, H., Ishak, E.W., Feiner, S., 2004. Collaborative mixed reality visualization of an archaeological excavation. In: Proceedings of the Third IEEE and ACM International Symposium on Mixed and Augmented Reality, 2–5 November, 2004, Arlington, VA, USA, pp.132–140.

Bichlmeier, C., Heining, S.M., Rustaee, M., Navab, N., 2007. Laparoscopic virtual mirror for understanding vessel structure evaluation study by 12 surgeons. In: Proceedings of the Sixth IEEE International Symposium on Mixed and Augmented Reality, 13–16 November, 2007, Nara, Japan, pp. 125–128.

Bichlmeier, C., Blum, T., Euler, E., Navab, N., 2010. Evaluation of the virtual mirror as a navigational aid for augmented reality driven minimally invasive procedures. In: Proceedings of the Ninth IEEE International Symposium on Mixed and Augmented Reality, 13–16 October, 2010, Seoul, Korean, pp. 91–97.

Billinghurst, M., Campbell, B., Kiyokawa, K., Woods, E., 2003. An occlusion-capable optical see-through head mount display for supporting co-located collaboration. In: Proceedings of the Second IEEE International Symposium on Mixed and Augmented Reality, 7–10 October, 2003, Tokyo, Japan, pp. 133–141.

Billinghurst, M., Nelles, C., Lee, G.A., Kim, G.J., 2004. Immersive authoring of tangible augmented reality applications. In: Proceedings of the Third IEEE and ACM International Symposium on Mixed and Augmented Reality, 2–5 November, 2004, Arlington, VA, USA, pp.172–181.

Billingshurst, M., Henrysson, A., Ollila, M., 2005. Face to face collaborative AR on mobile phones. In: Proceedings of the Fourth IEEE/ACM International Symposium on Mixed and Augmented Reality, 5–8 October, 2005, Vienna, Austria, pp. 80–89.

Blum, T., Wieczorek, M., Aichert, A., Tibrewal, R., Navab, N., 2010. The effect of out-of-focus blur on visual discomfort when using Stereo displays. In: Proceedings of the Ninth IEEE International Symposium on Mixed and Augmented Reality, 13–16 October, 2010, Seoul, Korean, pp. 3–12.

Böckelmann, I., Doil, F., Günthner, W.A., Hamacher, D., Klinker, G., Reif, R., Schwerdtfeger, B., Tümler, J., 2009. Pick-by-vision: a first stress test. In: Proceedings of the Eight IEEE International Symposium on Mixed and Augmented Reality, 19–22 October, 2009, Orlando, Florida, USA, pp.115–124.

Bubb, H., Lange, C., Sandor, C., Tönnis, M., 2005. Experimental evaluation of an augmented reality visualization for directing a car driver's attention. In: Proceedings of the Fourth IEEE/ACM International Symposium on Mixed and Augmented Reality, 5–8 October, 2005, Vienna, Austria, pp. 56–59.

Carroll, John M. (Ed.), 2003. HCI Models, Theories and Frameworks: Toward a Multidisciplinary Science. Morgan Kaufman Publishers, San Francisco, CA, USA.

Davidoff, J., 1975. Differences in Visual Perception: The Individual Eye. Crosby Lockwood, London, UK.

Dünser, A., Grasset, R., Seichter, H., Billinghurst, M., 2007. Applying HCI Principles to AR Systems Design. HIT Lab NZ, University of Canterbury, New Zealand. <http://ir.canterbury.ac.nz/bitstream/10092/2340/1/12604890_2007-MRUI-Applying_HCI_principles.pdf>.

Dünser, A., Grasset, R., Billinghurst, M., 2008. A survey of evaluation techniques used in augmented reality studies. In: Proceedings of the First ACM SIGGRAPH Conference and Exhibition in Asia, 10–13 December, 2008, Singapore, pp. 1–27.

Feiner, S., Güven, S. 2006. Visualizing and navigating complex situated hypermedia in augmented and virtual reality, In: Proceedings of the 4th IEEE and ACM International Symposium on Mixed and Augmented Reality, 22–25 October, 2006, Santa Barbara, USA, pp. 155–158.

Feiner, S., Henderson, S.J. 2009. Evaluating the Benefits of Augmented Reality for Task Localization in Maintenance of an Armored Personnel Carrier Turret, In: Proceedings of the Eight IEEE International Symposium on Mixed and Augmented Reality, 13–16 November, 2007, Orlando, Florida, USA, pp. 135–144.

Feiner, S., Lister, L., White, S., 2007. Visual hints for tangible gestures in augmented reality. In: Proceedings of the Sixth IEEE International Symposium on Mixed and Augmented Reality, 13–16 November, 2007, Nara, Japan, pp. 47–50.

Feng, D., Feiner, S., White, S., 2009. Interaction and presentation techniques for shake menus in tangible augmented reality. In: Proceedings of the Eight IEEE International Symposium on Mixed and Augmented Reality, 19–22 October, 2009, Orlando, Florida, USA, pp. 39–48.

Fjeld, M., Schar, S.G., Signorello, D., Krüger, H., 2002. Alternative tools for tangible interaction: a usability evaluation. In: Proceedings of the First IEEE and ACM International Symposium on Mixed and Augmented Reality, 30 September–1 October, 2002, Darmstadt, Germany, pp. 157–318.

Furmanski, C., Azuma, R., Daily, M., 2002. Augmented-reality visualizations guided by cognition: perceptual heuristics for combining visible and obscured information. In: Proceedings of the First IEEE and ACM International Symposium on Mixed and Augmented Reality, 30 September–1 October, 2002, Darmstadt, Germany, pp. 215–224.

Gabbard, J.L., Hix, D., 2001. Researching Usability Design and Evaluation Guidelines for Augmented Reality (AR) Systems. Laboratory for Scientific Visual Analysis, Virginia Tech, USA. <http://www.sv.vt.edu/classes/ESM4714/Student_Proj/class00/gabbard/index.html>.

Gabbard, J.L., Swan, J.E., 2008. Usability engineering for augmented reality: employing user-based studies to inform design. IEEE Transactions on Visualization and Computer Graphics 14 (3), 513–525.

Gabbard, J.L., Hix, D., Swan, J.E., 1999. User-centered design and evaluation of virtual environments. IEEE Computer Graphics and Applications 19 (6), 51–59.

Gandy, M., MacIntyre, B., Presti, P., Dow, S., Bolter, J., Yarbrough, B., O'Rear, N., 2005. AR karaoke: acting in your favorite scenes. In: Proceedings of the Fourth IEEE/ACM International Symposium on Mixed and Augmented Reality, 5–8 October, 2005, Vienna, Austria, pp. 114–117.

Gordon, G., Billinghurst, M., Bell, M., Woodfill, J., Kowalik, B., Erendi, A., Tilander, J., 2002. The use of dense stereo range data in augmented reality. In: Proceedings of the First IEEE and ACM International Symposium on Mixed and Augmented Reality, 30 September–1 October, 2002, Darmstadt, Germany, pp. 14–23.

Goto, M., Uematsu, Y., Saito, H., Senda, S., Iketani, A., 2010. Task support system by displaying instructional video onto AR workspace. In: Proceedings of the Ninth IEEE International Symposium on Mixed and Augmented Reality, 13–16 October, 2010, Seoul, Korean, pp. 83–90.

Grasset, R., Lamb, P., Billinghurst, M., 2005. Evaluation of mixed-space collaboration. In: Proceedings of the Fourth IEEE/ACM International Symposium on Mixed and Augmented Reality, 5–8 October, 2005, Vienna, Austria, pp. 90–99.

Grasset, R., Dünser, A., Billinghurst, M., 2007. Human-centered development of an AR handheld display. In: Proceedings of the Sixth IEEE International Symposium on Mixed and Augmented Reality, 13–16 November, 2007, Nara, Japan, pp. 177–180.

Gray, W.D., Salzman, M.C., 1998. Damaged merchandise? A review of experiments that compare usability evaluation methods. Human–Computer Interaction 13, 203–261.

Grundhöfer, A., Seeger, M., Hantsch, F., Bimber, O., 2007. Dynamic adaptation of projected imperceptible codes. In: Proceedings of the Sixth IEEE International Symposium on Mixed and Augmented Reality, 13–16 November, 2007, Nara, Japan, pp. 181–190.

Hassenzahl, M., Kort, J., Law, E.L.C., Roto, V., Vermeeren, A.P.O.S., 2009. Understanding, scoping and defining user experience: a survey approach. In: Proceedings of the 27th International Conference on Human Factors in Computing Systems, 04–09 April, 2009, Boston, MA, USA, pp. 719–728.

Higa, K., Nishiura, T., Kimura, A., Shibata, F., Tamura, H., 2007. A two-by-two mixed reality system that merges real and virtual worlds in both audio and visual senses. In: Proceedings of the Sixth IEEE International Symposium on Mixed and Augmented Reality, 13–16 November, 2007, Nara, Japan, pp. 203–206.

ISO DIS 9241-210:2008. Ergonomics of Human System Interaction – Part 210: Human-Centred Design for Interactive Systems (Formerly Known as 1340). International Organization for Standardization (ISO), Switzerland, 2008.

Jones, B.R., Sodhi, R., Campbell, R.H., Garnett, G., Bailey, B.P., 2010. Build your world and play in it: interacting with surface particles on complex objects. In: Proceedings of the Ninth IEEE International Symposium on Mixed and Augmented Reality, 13–16 October, 2010, Seoul, Korean pp. 165–174.

King, G.R., Piekarski, W., Thomas, B.H., 2006. ARVino – outdoor augmented reality visualisation of viticulture GIS data. In: Proceedings of the Fifth IEEE and ACM International Symposium on Mixed and Augmented Reality, 22–25 October, 2006, Santa Barbara, pp. 52–55.

Kiyokawa, K., 2007. A wide field-of-view head mounted projective display using hyperbolic half-silvered mirrors. In: Proceedings of the Sixth IEEE International Symposium on Mixed and Augmented Reality, 13–16 November, 2007, Nara, Japan, pp. 207–210.

Kiyokawa, K., Billinghurst, M., Hayes, S.E., Gupta, A., Sannohe, Y., Kato, H., 2002. Communication behaviors of co-located users in collaborative AR interfaces. In: Proceedings of the First IEEE and ACM International Symposium on Mixed and Augmented Reality, 30 September–1 October, 2002, Darmstadt, Germany, pp. 139–148.

Kiyokawa, K., Machida, T., Saitoh, K., Takemura, H., 2006. A 2D–3D integrated interface for mobile robot control using omnidirectional images and 3D geometric models. In: Proceedings of the Fifth IEEE and ACM International Symposium on Mixed and Augmented Reality, 22–25 October, 2006, Santa Barbara, pp. 173–176.

Klinker, G., Tönnis, M., 2006. Effective control of a car driver's attention for visual and acoustic guidance towards the direction of imminent dangers. In: Proceedings of the Fifth IEEE and ACM International Symposium on Mixed and Augmented Reality, 22–25 October, 2006, Santa Barbara, pp. 13–22.

Klinker, G., Dutoit, A.H., Bauer, M., 2002. Fata Morgana – a presentation system for product design. In: Proceedings of the First IEEE and ACM International Symposium on Mixed and Augmented Reality, 30 September–1 October, 2002, Darmstadt, Germany, pp. 1–10.

Klinker, G., Lange, C., Tönnis, M., 2007. Visual longitudinal and lateral driving assistance in the head-up display of cars. In: Proceedings of the Sixth IEEE International Symposium on Mixed and Augmented Reality, 13–16 November, 2007, Nara, Japan, pp. 91–94.

Knorlein, B., Di Luca, M., Harders, M., 2009. Influence of visual and haptic delays on stiffness perception in augmented reality. In: Proceedings of the Eight IEEE International Symposium on Mixed and Augmented Reality, 19–22 October, 2009, Orlando, Florida, USA, pp. 49–52.

Kotranza, A., Lind, D.S., Pugh, C.M., Lok, B., 2009. Real-time in situ visual feedback of task performance in mixed environments for learning joint psychomotor-cognitive tasks. In: Proceedings of the Eight IEEE International Symposium on Mixed and Augmented Reality, 19–22 October, 2009, Orlando, Florida, USA, pp. 125–134.

Kruijff, E., Swan, J.E., Feiner, S., 2010. Perceptual issues in augmented reality revisited. In: Proceedings of the Ninth IEEE International Symposium on Mixed and Augmented Reality, 13–16 October, 2010, Seoul, Korean, pp. 3–12.

Lindeman, R.W., Noma, H., de Barros, P.G., 2007. Hear-through and mic-through augmented reality: using bone conduction to display spatialized audio. In: Proceedings of the Sixth IEEE International Symposium on Mixed and Augmented Reality, 13–16 November, 2007, Nara, Japan, pp. 173–176.

Lindgaard, G., Parush, A., 2008. Utility and Experience in the Evolution of Usability. Maturing Usability: Quality in Software, Interaction and Value. Springer-Verlag, London, pp. 222–249.

Liu, S., Cheng, D., Hua, H., 2008. An optical see-through head mounted display with addressable focal planes. In: Proceedings of the Seventh IEEE International Symposium on Mixed and Augmented Reality, 15–18 September, 2008, Cambridge, UK, pp. 33–42.

Livingston, M.A., 2006. Quantification of visual capabilities using augmented reality displays. In: Proceedings of the Fifth IEEE and ACM International Symposium on Mixed and Augmented Reality, 22–25 October, 2006, Santa Barbara, pp. 3–12.

Livingston, M.A., Ai, Z., 2008. The effect of registration error on tracking distant augmented objects. In: Proceedings of the Seventh IEEE International Symposium on Mixed and Augmented Reality, 15–18 September, 2008, Cambridge, UK, pp. 77–86.

Livingston, M., Swan, J., Gabbard, J., Höllerer, T., Hix, D., Baillot, Y., Brown, D., 2003. Resolving multiple occluded layers in augmented reality. In: Proceedings of the Second IEEE International Symposium on Mixed and Augmented Reality, 7–10 October, 2003, Tokyo, Japan, pp. 56–65.

Livingston, M.A., Ai, Z., Decker, J.W., 2009. A user study towards understanding stereo perception in head-worn augmented reality displays. In: Proceedings of the Eight IEEE International Symposium on Mixed and Augmented Reality, 19–22 October, 2009, Orlando, Florida, USA, pp. 53–56.

MacWilliams, A., Sandor, C., Wagner, M., Bauer, M., Klinker, G., Brügge, B., 2003. Herding sheep: live system development for distributed augmented reality. In: Proceedings of the Second IEEE International Symposium on Mixed and Augmented Reality, 7–10 October, 2003, Tokyo, Japan, pp. 123–132.

Miyashita, T., Meier, P., Tachikawa, T., Orlic, S., Eble, T., Scholz, V., Gapel, A., Gerl, O., Arnaudov, S., Lieberknecht, S., 2008. An augmented reality museum guide. In: Proceedings of the Seventh IEEE International Symposium on Mixed and Augmented Reality, 15–18 September, 2008, Cambridge, UK, pp. 103–106.

Nilsson, S., Johansson, B., Jönsson, A., 2010. Using AR to support cross-organisational collaboration in dynamic tasks. In: Proceedings of the Ninth IEEE International Symposium on Mixed and Augmented Reality, 13–16 October, 2010, Seoul, Korean, pp. 3–12.

Feng, D., Feiner, S., White, S., 2009. Interaction and presentation techniques for shake menus in tangible augmented reality. In: Proceedings of the Eight IEEE International Symposium on Mixed and Augmented Reality, 19–22 October, 2009, Orlando, Florida, USA, pp. 39–48.

Oh, J., Hua, H., 2006. User evaluations on form factors of tangible magic lenses. In: Proceedings of the Fifth IEEE and ACM International Symposium on Mixed and Augmented Reality, 22–25 October, 2006, Santa Barbara, pp. 23–32.

Petersen, N., Stricker, D., 2009. Continuous natural user interface: reducing the gap between real and digital world. In: Proceedings of the Eight IEEE International Symposium on Mixed and Augmented Reality, 19–22 October, 2009, Orlando, Florida, USA, pp. 23–26.

Peterson, S.D., Axholt, M., Ellis, S.R., 2008. Label segregation by remapping stereoscopic depth in far-field augmented reality. In: Proceedings of the Seventh IEEE International Symposium on Mixed and Augmented Reality, 15–18 September, 2008, Cambridge, UK, pp. 143–152.

Pinelle, D., Gutwin, C., Greenberg, S., 2003. Task analysis for groupware usability evaluation: modeling shared-workspace tasks with the mechanics of collaboration. ACM Transactions on Computer–Human Interaction 10 (4), 281–311.

Platonov, J., Heibel, H., Meier, P., Grollmann, B., 2006. A mobile markerless AR system for maintenance and repair. In: Proceedings of the Fifth IEEE and ACM International Symposium on Mixed and Augmented Reality, 22–25 October, 2006, Santa Barbara, pp. 105–108.

Prytz, E., Nilsson, S., Jönsson, A., 2010. The importance of eye-contact for collaboration in AR systems. In: Proceedings of the Ninth IEEE International Symposium on Mixed and Augmented Reality, 13–16 October, 2010, Seoul, Korean, pp. 119–126.

Quarles, J., Lampotang, S., Fischler, I., Fishwick, P., Lok, B., 2008. Collocated AAR: augmenting after action review with mixed reality. In: Proceedings of the Seventh IEEE International Symposium on Mixed and Augmented Reality, 15–18 September, 2008, Cambridge, UK, pp. 107–116.

Quesenbery, W., 2003. The Five Dimensions of Usability. Content and Complexity: Information Design in Technical Communication. Routledge, USA, pp. 75–94.

Raskin, J., 1994. Viewpoint: Intuitive Equals Familiar. Communications of the ACM 37 (9), 17–18.

Robertson, C.M., MacIntyre, B., Walker, B.N., 2008. An evaluation of graphical context when the graphics are outside of the task area. In: Proceedings of the Seventh IEEE International Symposium on Mixed and Augmented Reality, 15–18 September, 2008, Cambridge, UK, pp. 73–76.

Robertson, C.M., MacIntyre, B., Walker, B.N., 2009. An evaluation of graphical context as a means for ameliorating the effects of registration error. IEEE Transactions on Visualization and Computer Graphics 15 (2), 179–192.

Sandor, C., Cunningham, A., Dey, A., Mattila, V.V., 2010. An augmented reality X-ray system based on visual saliency. In: Proceedings of the Ninth IEEE International Symposium on Mixed and Augmented Reality, 13–16 October, 2010, Seoul, Korean, pp. 27–36.

Schall, G., Mendez, E., Schmalstieg, D., 2008. Virtual redlining for civil engineering in real environments. In: Proceedings of the Seventh IEEE International Symposium on Mixed and Augmented Reality, 15–18 September, 2008, Cambridge, UK, pp. 95–98.

Schwerdtfeger, B., Klinker, G., 2008. Supporting order picking with augmented reality. In: Proceedings of the Seventh IEEE International Symposium on Mixed and Augmented Reality, 15–18 September, 2008, Cambridge, UK, pp. 91–94.

Sugano, N., Kato, H., Tachibana, K., 2003. The effects of shadow representation of virtual objects in augmented reality. In: Proceedings of the Second IEEE International Symposium on Mixed and Augmented Reality, 7–10 October, 2003, Tokyo, Japan, pp. 76–83.

Swan, J.E., Gabbard, J.L., 2005. Survey of user-based experimentation in augmented reality. In: Proceedings of 1st International Conference on Virtual Reality, Las Vegas, Nevada, pp. 1–9.

Takemura, M., Kitahara, I., Ohta, Y., 2006. Photometric inconsistency on a mixed-reality face. In: Proceedings of the Fifth IEEE and ACM International Symposium on Mixed and Augmented Reality, 22–25 October, 2006, Santa Barbara, pp. 129–138.

Tanaka, K., Kishino, Y., Miyamae, M., Terada, T., Nishio, S., 2008. An information layout method for an optical see-through head mounted display focusing on the viewability. In: Proceedings of the Seventh IEEE International Symposium on Mixed and Augmented Reality, 15–18 September, 2008, Cambridge, UK, pp. 139–142.

Tang, A., Zhou, J., Owen, C., 2003. Evaluation of calibration procedures for optical see-through head-mounted displays. In: Proceedings of the Second IEEE International Symposium on Mixed and Augmented Reality, 7–10 October, 2003, Tokyo, Japan, pp. 161–168.

Tateno, K., Takemura, M., Ohta, Y., 2005. Enhanced eyes for better gaze-awareness in collaborative mixed reality. In: Proceedings of the Fourth IEEE / ACM International Symposium on Mixed and Augmented Reality, 5–8 October, 2005, Vienna, Austria, pp. 100–103.

Thomas, B., 2006. Evaluation of three input techniques for selection and annotation of physical objects through an augmented reality view. In: Proceedings of the Fifth IEEE and ACM International Symposium on Mixed and Augmented Reality, 22–25 October, 2006, Santa Barbara, pp. 33–36.

Tümler, J., Mecke, R., Schenk, M., Huckauf, A., Doil, F., Paul, G., Pfister, E.A., Böckelmann, I., Roggentin, A., 2008. Mobile augmented reality in industrial applications: approaches for solution of user-related issues. In: Proceedings of the Seventh IEEE International Symposium on Mixed and Augmented Reality, 15–18 September, 2008, Cambridge, UK, pp. 87–90.

Veas, E., Kruijff, E., 2008. Vesp'R: design and evaluation of a handheld AR device. In: Proceedings of the Seventh IEEE International Symposium on Mixed and Augmented Reality, 15–18 September, 2008, Cambridge, UK, pp. 43–52.

Wang, X., Kotranza, A., Quarles, J., Lok, B., 2005. A pipeline for rapidly incorporating real objects into a mixed environment. In: Proceedings of the Fourth IEEE/ACM International Symposium on Mixed and Augmented Reality, 5–8 October, 2005, Vienna, Austria, pp. 170–173.

Wither, J., DiVerdi, S., Höllerer, T., 2007. Evaluating display types for AR selection and annotation. In: Proceedings of the Sixth IEEE International Symposium on Mixed and Augmented Reality, 13–16 November, 2007, Nara, Japan, pp. 95–98.

Zäh, M., Vogl, W., 2006. Interactive laser-projection for programming industrial robots. In: Proceedings of the Fifth IEEE and ACM International Symposium on Mixed and Augmented Reality, 22–25 October, 2006, Santa Barbara, pp. 125–128.

Zhang, R., Hua, H., 2010. Effects of a retroreflective screen on depth perception in a head-mounted projection display. In: Proceedings of the Ninth IEEE International Symposium on Mixed and Augmented Reality, 13–16 October, 2010, Seoul, Korean, pp. 137–145.

Zhou, F., Duh, H.B.-L., Billinghurst, M., 2008. Trends in augmented reality tracking, interaction and display: a review of 10 years in *ISMAR*. In: Proceedings of the Seventh IEEE International Symposium on Mixed and Augmented Reality, 15–18 September, 2008, Cambridge, UK, pp. 193–202.