Evaluating Mixed and Augmented Reality: A Systematic Literature Review (2009–2019)

Leonel Merino* Magdalena Schwarzl† Matthias Kraus‡ Michael Sedlmair§ Dieter Schmalstieg¶
University of Stuttgart University of Stuttgart University of Konstanz University of Stuttgart Graz University of Technology
Daniel Weiskopf‖
University of Stuttgart

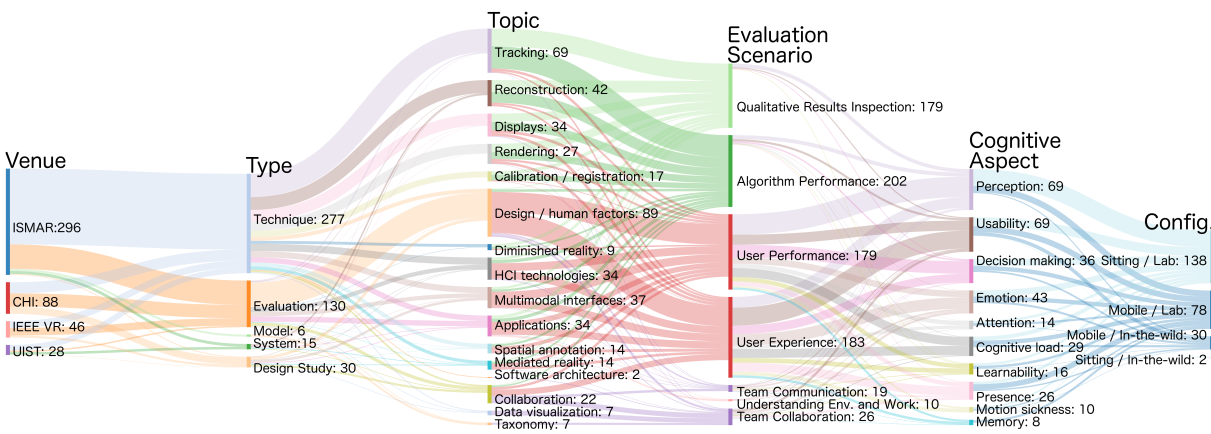

Figure 1: The 458 papers that report on evaluations in mixed and augmented reality (MR/AR) published in ISMAR, CHI, IEEE VR, and UIST between 2009 and 2019 classified by (left-to-right): venue, paper type, research topic, evaluation scenario, cognitive aspects involved in user studies, and study configuration (conducted in the lab or in-the-wild while participants were static or mobile).

Abstract

We present a systematic review of 458 papers that report on evaluations in mixed and augmented reality (MR/AR) published in ISMAR, CHI, IEEE VR, and UIST over a span of 11 years (2009–2019). Our goal is to provide guidance for future evaluations of MR/AR approaches. To this end, we characterize publications by paper type (*e.g.*, technique, design study), research topic (*e.g.*, tracking, rendering), evaluation scenario (*e.g.*, algorithm performance, user performance), cognitive aspects (*e.g.*, perception, emotion), and the context in which evaluations were conducted (*e.g.*, lab vs. in-the-wild). We found a strong coupling of types, topics, and scenarios. We observe two groups: (a) technology-centric performance evaluations of algorithms that focus on improving tracking, displays, reconstruction, rendering, and calibration, and (b) human-centric studies that analyze implications of applications and design, human factors on perception, usability, decision making, emotion, and attention. Amongst the 458 papers, we identified 248 user studies that involved 5,761 participants in total, of whom only 1,619 were identified as female. We identified 43 data collection methods used to analyze 10 cognitive aspects. We found nine objective methods, and eight methods that support qualitative analysis. A majority (216/248) of user studies are conducted in a laboratory setting. Often (138/248), such studies involve participants in a static way. However, we also found a fair number (30/248) of in-the-wild studies that involve participants in a mobile fashion. We consider this paper to be relevant to academia and industry alike in presenting the state-of-the-art and guiding the steps to designing, conducting, and analyzing results of evaluations in MR/AR.

Keywords: Mixed and Augmented Reality, Evaluation, Systematic Literature Review.

Index Terms: I.3.7 [Computing Methodologies]: Computer Graphics—Three-Dimensional Graphics and Realism; A.1 [General Literature]: Introductory and Survey—

1 Introduction

Across multiple domains, there is an increasing interest in investigating approaches that employ mixed reality (MR) and augmented reality (AR) technologies [100]. One example is data visualization, in which researchers of the emerging immersive analytics [76] domain study how adopting MR/AR technologies can boost the effectiveness of displaying information and interacting with visualizations. However, designing appropriate evaluations that examine MR/AR is challenging, and suitable guidance to design and conduct evaluations of MR/AR are largely missing. There are several strategies that can be adopted to evaluate MR/AR approaches. When the subject of the evaluation is an algorithm or a novel method, benchmarks can help analyze increases in performance. Sometimes, when a user interface is involved, user studies can provide rich data for the analysis not only of user performance but also of user experience. If the focus of the evaluation is on user environments and work places, surveys and case studies can provide important insights. However, generally it is difficult to identify a suitable evaluation strategy, variables to be examined, and adequate methods to collect relevant data in an evaluation for answering a particular research question.

*e-mail: leonel.merino@visus.uni-stuttgart.de
†e-mail: magdalena.schwarzl@visus.uni-stuttgart.de
‡e-mail: matthias.kraus@uni-konstanz.de
§e-mail: michael.sedlmair@visus.uni-stuttgart.de
¶e-mail: schmalstieg@tugraz.at
‖e-mail: daniel.weiskopf@visus.uni-stuttgart.de

Often, approaches are considered most effective when they boost users' performance in decision making. However, we observe that there are several other cognitive aspects (*e.g.*, perception, emotion, presence, cognitive load, attention, learnability, and memory) that can also play a fundamental role in the effectiveness of approaches in MR/AR. We conjecture that there is an interplay of cognitive aspects that require evaluations to be comprehensive, for instance, to understand the reasons that led to a high user performance. Our goal is to better understand MR/AR evaluation practices with an eye toward guidance on when to perform which type of evaluation.

To address our goal we conducted a systematic literature review. We concentrated on the analysis of papers published in ISMAR [8], CHI [2], IEEE VR [6], and UIST [9]. We consider these to be the leading venues in MR/AR research, offering a sound and representative body of literature for MR/AR research. We confirm our impressions based on the flagship A^* classification that ISMAR and CHI and the *A* classification that IEEE VR and UIST obtain in the CORE ranking [4] (which considers various indicators such as citation rate, paper submissions, and acceptance rate). We opted to select papers published in the recent past, and analyzed proceedings of these main MR/AR conferences published across.

To facilitate the analysis of the papers, we relied on our experience and adopted a popular classification from the visualization community [5], putting papers into one of five types based on their main contribution. We observe that MR/AR encompasses multiple topics. Consequently, we complemented the classification by paper type with sixteen research topics that emerged from our analysis. We also adapted the seven evaluation scenarios introduced by Lam *et al.* [66] (of which we excluded one), and the scenario extended by Isenberg *et al.* [50], to the context of MR/AR approaches. In the end, we classified the scenario of evaluations into one of seven types. For evaluations that involve users, we identified whether the evaluation was conducted in-the-wild (the targeted real-world usage environment) or in a laboratory, and whether participants of such studies used MR/AR while they were sitting or in a mobile way. To analyze the implications of approaches that use MR/AR in human cognition, we inferred ten cognitive aspects based on the data collection methods employed in user studies. An overview of the relationships of the analyzed dimensions is presented in Fig. 1. The main contribution of our paper is threefold: (a) a systematic analysis of paper types, research topics, evaluation scenarios, cognitive aspects that emerge from data collection methods, and configurations of evaluations in MR/AR, (b) a synthesis that describes implications for evaluating MR/AR approaches, and (c) a publicly available data set of the data collected in our systematic literature analysis [82].

2 Related Work

To elaborate on the related work, we discuss previous papers that cover various aspects of evaluations in MR/AR. Next, we leverage our experience in visualization research and extend our analysis to visualization studies that share our focus on evaluations. Finally, we elaborate on commonalities and differences of these related works to our investigation.

There are a number of survey articles in MR/AR. Swan and Gabbard [107] surveyed AR papers published in 1992–2004 and analyzed 21 papers that describe user evaluations. Similarly, a few years after, Duenser *et al.* [30] analyzed 161 AR papers published in 1993–2007. Both studies classified papers by evaluation type (*e.g.*, perception, performance) and involved methods (*e.g.*, objective/subjective measurements, qualitative analysis). They found that 47% of user evaluations measured user task, and 22% analyzed variables of perception or cognition. We consider these works complementary to our study. Zhou *et al.* [122] focused their literature review on tracking, interaction, and display technologies. They found an emerging trend of papers that focused primarily on evaluations, which accounted to 5.8% of the reviewed papers. Kruijff *et al.* [65] presented a classification of perceptual issues grouped into categories such as environment, capturing, augmentation, display, and individual user differences. Fite-Georgel [33] surveyed AR industrial applications, organized into categories that relate to the stages of the life-cycle of products. The applications were evaluated using the following criteria: workflow integration, scalability, cost/benefit, out of the lab, user tested, out of developers' hands, and involvement of the industry. Radu and MacIntyre [94] investigated how AR designs relate to children's skills, such as motor abilities, spatial cognition, attention, logic, and memory. Krichenbauer *et al.* [63] surveyed professionals who create 3D media content using AR user interfaces. A set of requirements were distilled and implemented in a prototype tool. Grubert *et al.* [42] presented a taxonomy for pervasive AR and context-aware AR based on context sources, context targets, and context controllers. Chen *et al.* [21] classified medical MR to identify areas with little research as well as to provide references to practitioners. Recently, Kim *et al.* [56] reviewed the literature in MR/AR published in 2008–2017, and found a sharp increase in AR evaluation to which they related 16.4% of the reviewed papers. Fonnet and Pri [35] surveyed 177 immersive analytics papers published in 1991–2018. They included in the analysis aspects of evaluations such as immersion, technologies, interaction, and visualization techniques. Dey *et al.* [26] reviewed the MR/AR research literature that reports on user studies published in 2004–2014. They found an increasing trend of involving handhelds in AR user studies. They also confirmed that most user studies are conducted in laboratory settings. In contrast, our study includes more recent papers and elaborates on a broader view that includes both human- and technology-based evaluations.

There are studies in other fields that reviewed their respective literature and analyzed evaluations. We name a few examples from the field of visualization: Carpendale [20] discussed characteristics of information visualization evaluation in terms of evaluation strategies, data collection methods, and analysis methods. She reflected on the need for conducting more evaluations and postulated that evaluation should be more diverse in terms of employed methodologies. Lam *et al.* [66] identified seven scenarios of information visualization evaluation. The scenarios encapsulate current evaluation practices, which can guide researchers to design more effective evaluations. Isenberg *et al.* [50] later expanded the scenarios to include evaluations based on qualitative results inspection. Elmqvist and Yi [31] proposed a set of general and reusable patterns to commonly occurring problems in evaluating visualization approaches. Merino *et al.* [81] found that 62% of software visualization evaluations involved usage examples and anecdotal evidence, 29% experiments, and 7% case studies. Our work is methodically inspired by these systematic analyses, which we apply to MR/AR for the first time. In particular, our coding scheme uses the paper types described by Munzner [84] and seven of the evaluation scenarios defined by Lam *et al.* [66] and Isenberg *et al.* [50].

3 Methodology

We employed a systematic literature review approach. To mitigate potential biases in the results of the survey, we followed the comprehensive guidelines by Kitchenham [60]. The methodology offers robust and transferable evidence for evaluating and interpreting relevant research on a topic of interest. To this end, we defined a review protocol to ensure rigor and reproducibility, in which we determine (a) a data collection method, (b) selection criteria, (c) a coding scheme, and (d) a coding process.

3.1 Data Collection Method

We collected papers published in ISMAR, CHI, IEEE VR, and UIST. To find primary studies for our analysis, we collected all papers in the proceedings of ISMAR of the period 2009–2019. Next, we used the ACM Digital Library [1] to collect papers from CHI. We

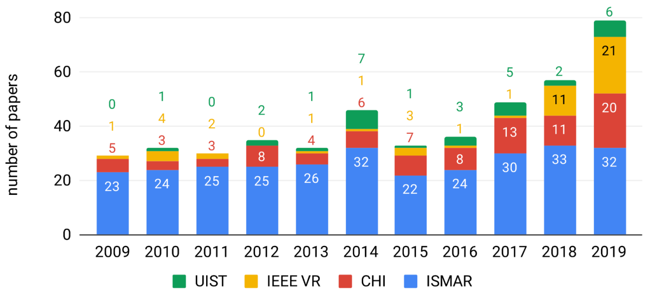

Figure 2: The 458 included papers by publication year and venue.

used IEEE Xplore [7] to collect papers from IEEE VR and UIST. In neither case, we included keywords such as "evaluation" in the search. That is, we first identified MR/AR papers, and then manually analyzed evaluations.

3.2 Selection Criteria

We analyzed the proceedings of 11 years (2009–2019) of ISMAR and included 296 papers. These papers correspond to full and short papers from 2009 until 2014, and T&S conference and TVCG journal papers from 2015 until 2019. We excluded other publication formats that, due to their brevity, are unlikely to contain enough details regarding an evaluation (*e.g.*, posters, demos, keynotes, extended abstracts). Also, we collected 88 papers from CHI, 46 papers from IEEE VR, and 28 papers from UIST. Since these venues not only focus on MR/AR, we excluded papers that either focus on a different topic (*e.g.*, virtual reality) or do not report on evaluation explicitly. Our set has 458 papers with 4 to 14 pages in length. A temporal histogram of the selected papers is shown in Fig. 2.

3.3 Coding Scheme

To analyze the evaluations reported in the MR/AR literature, we coded paper types, research topics, and evaluation scenarios. In evaluations where we identified users studies, we also coded data collection methods to infer cognitive aspects, number and gender of participants, and the adopted configuration of mobility.

Paper Types. We classified the paper type according to categories proposed by Munzner [84]. Although these categories aim at characterizing visualization papers, we observed that the categories can be generalized to MR/AR papers. As this classification has been widely adopted in multiple studies in visualization research, using it in our study of MR/AR evaluations provides a bridge between the two fields that enables comparison. In summary, a paper can be classified into one of five types from Table 1.

Research Topics. There are several research topics relevant to MR/AR. Although there are previous classifications of topics [56, 122], we opted to identify topics by ourselves. We think that comparing our resulting list to previous ones can help confirm the relevance of topics in common and identify as emergent topics the categories that are different. To this end, we analyzed topics listed in calls for papers and complemented the list with topics that emerge from the analysis of paper titles and keywords. For each paper included in our study, we identified one main research topic. In the end, we defined the 16 topics listed in Table 3.

Evaluation Scenarios. For each paper in our analysis, we looked for details of an evaluation, and, when we found some details, we classified the evaluation *scenario*. A scenario is the context in which an evaluation is carried out. We originally considered the eight scenarios that characterize the context in which evaluations are conducted in visualization [50, 66]. We observed that, while most scenarios fit well the context of MR/AR, "Visual Data Analysis and Reasoning" is too specific to the visualization domain, and therefore we excluded it. We adapted the remaining seven scenarios to the context of evaluation in MR/AR. We furthermore grouped these scenarios according to the level of user involvement in an evaluation: (a) *technique-centered* scenarios do not involve users, (b) *user-centered* scenarios involve users who individually interact with a technique, and (c) *team-centered* scenarios involve users who interact with each other with the support of a technique (often simultaneously). Table 2 presents the scenarios adapted to the evaluation of MR/AR approaches.

Cognitive Aspects. There are various aspects of human cognition that can be considered in user evaluations of MR/AR, which can allow researchers to obtain a more comprehensive understanding of the impact of their approaches. Commonly, studies focus only on a few cognitive aspects, so the scope of their analyses stays feasible. However, understanding multiple cognitive aspects together can be used to build theories that explain complex phenomena of human factors in MR/AR. We did not find cognitive aspects explicitly described in evaluations. Therefore, we adopted a bottom-up approach and inferred them from employed data collection meth-

Table 1: Paper Types.

Technique	Papers focusing on new algorithms that improve the performance of an approach.
Evaluation	Papers that elaborate on a judgment of the quality, importance, or value of an approach. They can describe careful examinations of a real-world case (*i.e.*, case study) or the behavior of users exposed to a tool (*i.e.*, user study).
System	Papers that elaborate on choices made in the design of the architecture of a proposed system or framework, and lessons learned from its use. These can be seen as meta-techniques that enable the generation of new techniques.
Model	Papers that include (a) commentaries of an expert in the field who argues to support a position, (b) formalisms of models, definitions, or terminology related to techniques, and (c) taxonomies and categories to help researchers analyze a domain.
Design study/ Application	Papers that describe how existing techniques can be useful to deal with a concrete problem in a domain.

Table 2: Evaluation Scenarios.

Technology	AP	Algorithm Performance: a quantitative evaluation of the technical performance, typically using benchmarks to compare rendering speed or memory performance.
	QRI	Qualitative Results Inspection: a qualitative discussion of results that encourages readers to agree on a quality statement.
Human	UP	User Performance: a quantitative or qualitative evaluation of the performance of the users of a system. Typically, user performance is measured in experiments using time and correctness of users to complete as set of tasks.
	UE	User Experience: an examination of how a user reacts to interacting with a tool. Likert scale questionnaires for subjective feedback, and interviews are commonly used to gather data of user experience.
	UWP	Understanding environment and Work Practices: an examination to understand the implications of adopting a technique into a working environment. Common examples of data collection methods are surveys and interviews with expert users.
Team	COM	Team communication in MR/AR: an assessment of the communicative value of a technique in regards to goals such as teaching or presentation.
	COL	Team collaboration in MR/AR: an evaluation of the level of support of a technique to facilitate collaboration in a team.

Table 3: Research Topics.

Tracking	Papers evolving around 3D tracking. It also contains most papers dealing with simultaneous localization and mapping, if the emphasis is on localization.
Reconstruction	Technical papers focusing on 3D reconstruction, either as a prerequisite for MR/AR applications (which will typically use the reconstructed models to derive some form of spatial annotation), or SLAM papers where the mapping part is most relevant.
Calibration/ registration	Papers focusing on spatial registration for real-time tracking. These papers have a thematic overlap with tracking and reconstruction.
Rendering	Papers dealing with coherent rendering, in particular, global illumination for MR, inverse rendering, and photometric registration.
Displays	Papers that deal with physical displays for MR/AR, mostly head-mounted displays and spatial AR.
HCI technologies	Papers discussing technical solutions to interaction problems.
Design/ human factors	Papers dealing with the design (and evaluation) of interaction techniques or with the study of human factors per se that occur in the context of MR/AR systems. One important group are perceptual issues, in particular, depth perception.
Applications	Papers exploring MR/AR interfaces in specific application use cases, covering both medical and non-medical applications.
Multimodal interfaces	Papers dealing with audio, haptics, and other non-visual modalities.
Collaboration	Papers describing collaborative MR/AR.
Mediated reality	Papers on changing the appearance of physical objects and scenes.
Spatial annotation	Papers that display semantic information registered to the real world, to instruct or guide the user. The main difference to mediated reality is that the real objects remain mostly visible and are "augmented", not "supplanted".
Data visualization	Papers that elaborate on the display of data registered to the real world in an MR/AR display. Difference to spatial annotation is that the data undergoes a noteworthy visual encoding, as opposed to annotations, which are visually trivial in most cases (such as a colored icon or text label).
Diminished reality	Papers on all kinds of techniques that make real things disappear or partially transparent.
Taxonomy	Papers describing theoretical discussions and taxonomies.
Software architecture	Papers describing software architectures.

ods. We did not collect general data collection methods such as questionnaires or interviews. Instead, we collected methods that are used in evaluations to analyze specific human aspects that deal with cognition. That is, these aspects deal with "the mental action or process of acquiring knowledge and understanding through thought, experience, and the senses" [3]. Therefore, we used these methods as a proxy to identify aspects of human cognition involved in evaluations. For example, in evaluations that describe the use of the Self-Assessment Manikin [17] method, we can infer that researchers investigate *emotions*. In the end, we identified 10 cognitive aspects, presented in Table 4.

Table 4: Cognitive Aspects.

Perception	Relates to the interpretation of sensory information (*e.g.*, visual, auditory, or haptic) to understand information of the environment.
Usability	The ease of use.
Emotion	A mental state that relates to thoughts, feelings, behavior, and affects.
Decision making	The process of identifying and choosing from several alternative possibilities.
Presence	The feeling of having no mediation between oneself and the (virtual) environment, which promotes the psychological sensation of "being there".
Cognitive load	Relates to the mental load imposed by instructional parameters, *e.g.*, task structure, the sequence of information given during an evaluation; and mental effort that refers to the capacity allocated by participants of a study to the instructional demands.
Attention	The process of selectively concentrating on an aspect while ignoring other information.
Learnability	Capability of a system to enable users to learn how to use it, usually considered as an aspect of usability.
Motion sickness	A disturbance of the senses due to a difference between actual and expected motion.
Memory	Relates to the ability of encoding, storing, and retrieving information when needed.

Study Configurations. For each user evaluation, we extract the number and gender of involved participants. As MR/AR devices often allow mobile use, we analyze whether this characteristic is present in user evaluations, or whether evaluations are conducted with users in a static way. Moreover, we code whether user evaluations are conducted in a laboratory setting or whether they correspond to field studies conducted in-the-wild. In-the-wild studies, a term commonly used in human-computer interaction (HCI), are conducted in a real-life scenario targeted by a MR/AR approach. Notice that in-the-wild does not necessarily imply outdoor usage, as multiple MR/AR approaches target indoor activities.

3.4 Coding Process

The coding process was carried out by the first three co-authors of this paper. Each of them analyzed a similar number of papers. Each paper was reviewed at least by two coders. Coders trained themselves by classifying the paper types of 72 publications of ISMAR in 2014–2018 and reached a "substantial" [67] 0.7353 Krippendorff's alpha [64] intercoder reliability. For all papers, we crosschecked the results and discussed conflicting results to reach a consensus. We built on our experience on visualization research to code paper types and evaluation scenarios using a defined set of categories. Research topics, cognitive aspects, and study configurations emerged from the analysis of papers and were iteratively refined. Categories of paper types, research topics, and study configuration are mutually exclusive, whereas multiple evaluation scenarios and cognitive aspects could be associated to individual papers.

To classify the papers, we followed an incremental reading approach. We started with the title, keywords, abstracts, and skimming figures, which in many cases already clarified paper types and main research topic. If unclear, we continued reading the introduction, and in some cases the entire paper. For coding evaluation scenario, cognitive aspects, and study configuration, we additionally identified the respective evaluation sections in the paper and closely read those.

4 Results

We now report on the results coding the 458 identified MR/AR papers. The results are organized according to the coding scheme introduced in Sect. 3.3 and Fig. 1. A summary of the results is

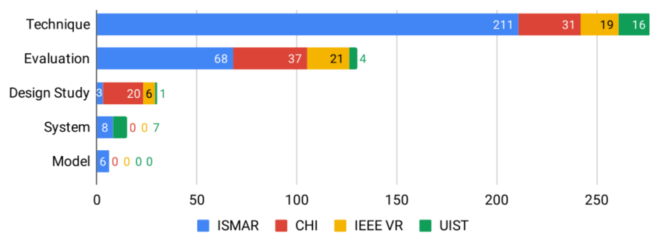

Figure 3: A classification of the 458 papers by type and publication venue: technique, evaluation, design study, system, and model.

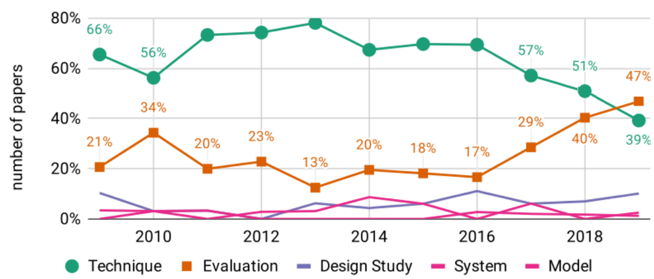

Figure 4: Trends of the types of papers per year.

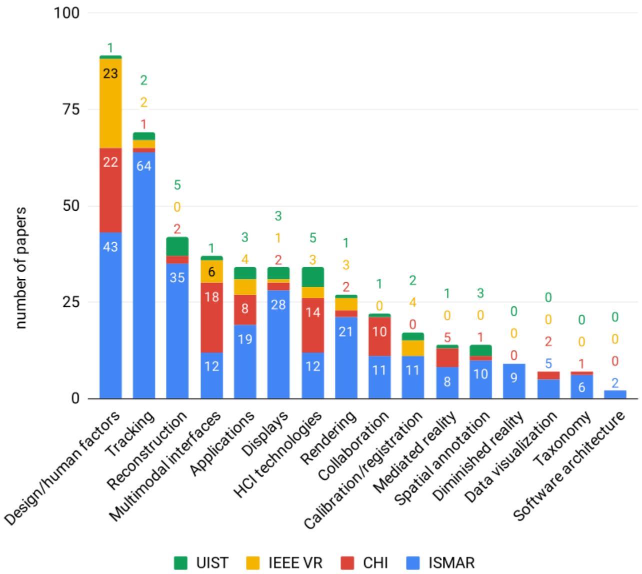

Figure 5: Research topics of interest to MR/AR.

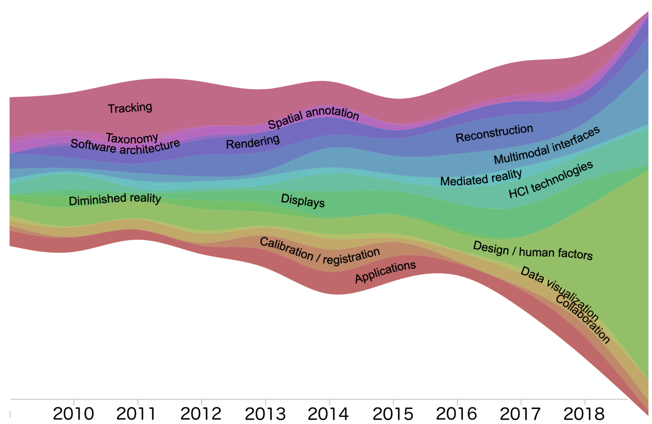

Figure 6: Trends of research topics per year.

presented in Table 6 with the number of papers in each research topic classified by paper types, evaluation scenarios, and cognitive aspects.

4.1 Paper Types

A summary of the results of our classification by paper type is presented in Fig. 3. We observe that the types of papers vary across venues. In ISMAR papers (296), technique papers (211) outnumber evaluation papers (68) by a factor of three. In CHI papers (88), evaluation papers (37), technique papers (31), and to a lesser extent design study papers (20) are almost balanced. In IEEE VR (46), papers are mostly of two types: evaluation (21) and technique (19). UIST papers (28), mostly consist of technique papers (16) and system papers (7).

We present the percentage of paper types over the total number of published papers during 2009–2019 in Fig. 4. Only technique and evaluation papers have non-marginal frequencies. Since there are small differences in the percentage of papers over time, the sum of the percentages of technique and evaluation papers is frequently close to 100%, making them the core pillars of the MR/AR literature. We observe that the percentage of evaluation papers is stable but low until 2016, in which a noticeable steady increase appears. In 2019, for the first time the percentage of evaluation papers exceeds the percentage of technique papers. We observe that this increase is triggered by the increased number of evaluation papers published in IEEE VR. We think that the increased focus on human-centric papers is a positive symptom of MR/AR becoming a more mature field, in which robust techniques and hardware are increasingly more available.

4.2 Research Topics

We identified 16 different research topics, as summarized in Fig. 5. When comparing the results to a previous study [56], we identified some new topics: design/human factors, mediated reality, spatial annotation, diminished reality, taxonomy, and software architecture. Although there might be an overlap of topics that could explain many differences, some of them could identify emergent topics in MR/AR. The frequency of topics varies amongst venues. In ISMAR papers (296), the main topics are tracking (64), design/human factors (43), reconstruction (35), displays (28), and rendering (21). These topics are coherent with the proportion of technique versus evaluation papers that we found. In contrast, the main topics of interest in CHI papers (88) are design/human factors (22), multimodal interfaces (18), HCI technologies (14), collaboration (10), and applications (8), which are in line with the balanced number of technique, evaluation, and design study papers. In IEEE VR papers (46), main topics are design/human factors (23), multimodal interfaces (6), applications (4), calibration/registration (4), and rendering (3). In UIST papers (28), the main topics are reconstruction (5), HCI technologies (5), applications (3), displays (3), and spatial annotation (3). IEEE VR and UIST seem to blend the topics of interest of ISMAR and CHI.

Fig. 6 shows a chart with the trends of the number of papers by research topic over time that help us analyze emergent topics. Indeed, since 2016, the number of papers dedicated to design/human factors have been greatly increasing and become predominant, exceeding the number of papers that focus on tracking, which exhibit a fairly decreasing trend. In turn, a fair number of papers dedicated to spatial annotation are found between 2009 and 2012, but completely absent after 2015; however, this topic of interest reappeared in 2018–2019. Other topics, such as diminished reality, software architecture, and taxonomy, are intermittent. A steady number of

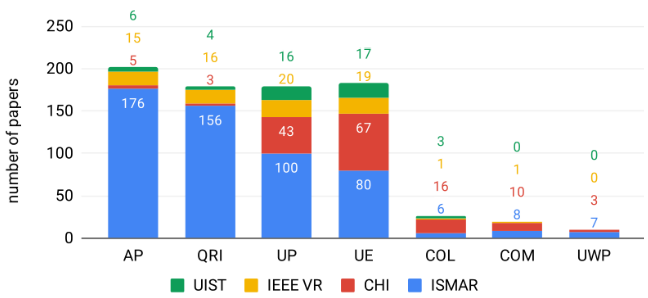

Figure 7: The 801 evaluation scenarios identified among the 458 analyzed papers: Algorithm Performance (AP), Qualitative Results Inspection (QRI), User Performance (UP), User Experience (UE), Understanding environment and Work Practices (UWP), Team Communication (COM), and Team Collaboration (COL).

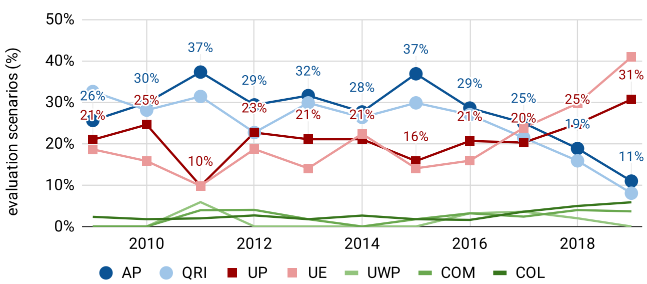

Figure 8: Trends of MR/AR evaluation scenarios (*notice that an evaluation can involve multiple scenarios*).

applications, mediated reality, collaboration, rendering papers are found in MR/AR.

4.3 Evaluation Scenarios

We summarize the number of papers per evaluation scenario in Fig. 7. As the evaluations in a paper can involve multiple scenarios, we found 801 evaluation scenarios in total, and we confirmed that: (a) most papers (382) involve evaluations of technique-centered scenarios (*e.g.*, benchmarks), (b) many papers (363) describe evaluations of user-centered scenarios (*e.g.*, user studies), and (c) a few papers (56) elaborate on evaluations of team-centered scenarios (*e.g.*, surveys). We observe that evaluations in ISMAR frequently focus on validating techniques involving AP+QRI scenarios (176/296) and design/human factors involving UP+UE scenarios (100/296), and much less frequently involving COL+COM+UWP (8/296). In contrast, evaluations in CHI and UIST mostly involve UP+UE scenarios (84/116), less frequently involve COL+COM+UWP (19/116), and rarely involve AP+QRI (11/116).

Fig. 8 shows a line chart with the trends of the seven evaluation scenarios: technique-centered scenarios (AP+QRI) in blue tones at the top, user-centered scenarios (UP+UE) in red tones in the middle, and team-centered scenarios (UWP+COM+COL) in green tones at the bottom of the chart.

4.4 Cognitive Aspects

We present the list of the 43 methods in Table 5, and a summary of the 10 inferred cognitive aspects in Fig. 9. Notice that cognitive aspects (and data collection methods) are not mutually exclusive. That is, a user evaluation can involve multiple of them. In particular, 28% of user evaluations (69/248) involve various aspects of perception (*e.g.*, visual, haptic, auditory), and a similar number of evaluations examined the usability of MR/AR approaches. We found that 17% of evaluations (43/248) involve the analysis of emotions,

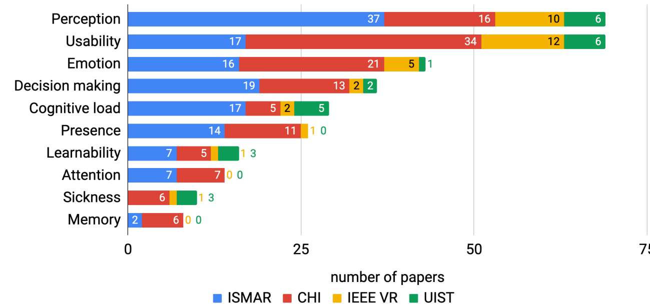

Figure 9: The number of papers that involve various cognitive aspects in MR/AR evaluation.

e.g., intuitiveness, usefulness, or joyfulness. We rarely found decision making explicitly mentioned as an aspect of evaluations in MR/AR papers. However, we identified 15% of user evaluations (36/248) that implicitly focus on it. Usually, these evaluations target MR/AR approaches that support users to make better decisions, for instance, in a short time and with high accuracy. Cognitive load is a frequent aspect involved in 12% of user evaluations (29/248). Presence is included in 10% of evaluations (26/248), which sometimes are combined with the analysis of awareness, embodiment, discernability, immersion, influence, and privacy. Less frequently, we found evaluations that examined (a) learnability (16/248) by means of pre- and post-tests of performance, (b) attention (14/248), typically by means of eye-tracking technology, (c) motion sickness (10/248), usually to assess fatigue amongst participants, and (d) memory (8/248), regarding learning and spatial memory. We also found a few other papers that reflect on cognitive aspects (but not in the context of an evaluation). We found two surveys: one [85] that reports on emotions of end-users who adopted an MR/AR tool, and another [87] that describes collected data for the analysis of user perception in the context of MR/AR. We found two model papers. One of them [94] discusses the impact of MR/AR in cognitive aspects (*i.e.*, perception, attention, memory). The other model paper [65] presents a taxonomy to characterize human perception. We found three highly comprehensive MR/AR evaluations [92, 95, 118] that involved five cognitive aspects and three evaluations [58, 108, 120] that involve four cognitive aspects each. We also found that 14 papers described evaluations that involve 3 cognitive aspects, and the remaining 228 user studies involved up to 2 cognitive aspects.

We now describe examples of data collection methods of each inferred cognitive aspect.

Perception. We found that 28% of user studies (69/248) analyze perception to examine topics such as design/human factors (23/69) and multimodal interfaces (14/69), which employ various methods depending on the type of perception.

Visual. When conducting perception studies, researchers selected either objective data collection methods, such as tracking head, eye, and body movements, or subjective data collection methods, such as (a) Absolute Category Rating (ACR11-HR) [68] (also called Single Stimulus Method). In it, participants are asked to evaluate the quality of a sequence of images that are presented one at a time and rated independently on a category scale. (b) Two-Alternative Forced-Choice (2-AFC) [91] has been used to measure various types of perception and attention. In it, participants are required to perform a central task and a peripheral task (*e.g.*, based on the visual angle or spatial location) simultaneously. For instance, researcher asked participants to scan a display panel, while at the same time, participants had to respond to light stimuli perceived in the periphery of their visual field. Head and eye movements are sometimes restricted depending on the focus of the evaluation. *Haptic.* Studies that focus on haptic perception (*e.g.*, softness or stiffness) have used the Two-Interval

Table 5: Data Collection Methods.

Aspect	Method	Description	Ref.	Approach	Type	Sense
Perception	**2-AFC**	Two-Alternative Forced-Choice method	[32]	Obj.	Quant.	A/H/V
	2-IFC	Two-Interval Forced-Choice method	[119]	Obj.	Quant.	A/H/V
	ACR11-HR	Absolute Category Rating	[22, 110]	Subj.	Both.	V
	SAQI	Spatial Audio Quality Inventory	[71]	Subj.	Quali.	A
Usability	**AD3**	Ad-hoc Usability Questionnaire (Awareness)	[51]	Subj.	Quant.	
	MREQ	Mixed Reality Experience Questionnaire	[98]	Subj.	Quant.	
	PEQ	Post Experience Questionnaire	[72]	Subj.	Quant.	
	SUS	The Slater-Usoh-Steed Questionnaire	[105]	Subj.	Quant.	
Emotion	**GEQ**	Game Experience Questionnaire	[49]	Subj.	Quant.	
	IMI	Intrinsic Motivation Inventory	[77]	Subj.	Quant.	
	PANAS	Positive and Negative Affect Schedule	[114]	Subj.	Quant.	
	SAM	Self-Assessment Manikin	[17]	Subj.	Quant.	
	USQ	IBM's Usability Satisfaction Questionnaire	[70]	Subj.	Quant.	
Presence	**AD1**	Ad-hoc Post Experimental Questionnaire	[106]	Subj.	Quant.	
	AD2	Ad-hoc Co-Presence Questionnaire	[43, 59]	Subj.	Quant.	
	BRQ	Body Representation Questionnaire (Embodiment)	[12]	Subj.	Quant.	
	IOS	Inclusion of Other in the Self Scale	[11]	Subj.	Quant.	
	IPQ	The Igroup Presence questionnaire	[101]	Subj.	Quant.	
	MEC	Spatial Presence Questionnaire	[113]	Subj.	Quant.	
	SPQ	Social Presence Questionnaire	[44]	Subj.	Quant.	
	MTQ	McKnight Trust Questionnaire (Trust)	[79]	Subj.	Quant.	
	TPI	The Temple Presence Inventory	[73]	Subj.	Quant.	
Cognitive load	**NASA-TLX**	NASA-TLX (Task Load Index)	[45]	Subj.	Quant.	
	SMEQ	Subjective Mental Effort Questionnaire	[123]	Subj.	Quant.	
	Paas	Paas Mental-Effort Rating Scale	[89]	Subj.	Quant.	
	ECG	Electrocardiogram	[38]	Obj.	Quant.	
	GSR	Galvanic Skin Response	[38]	Obj.	Quant.	
	ST	Skin Temperature	[38]	Obj.	Quant.	
Attention	**2-AFC**	Two-Alternative Forced-Choice method	[10, 32]	Subj.	Quant.	V/A
	AD4	Ad-hoc Self-Report Questionnaire	[40]	Subj.	Quant.	A
	VisEng.	User Engagement Self-Report Questionnaire	[74]	Subj.	Quant.	A
	ET	Eye-tracking	[48, 88, 115]	Obj.	Quant.	V
	HT	Head-tracking	[48, 115]	Obj.	Quant.	V
Learnability	**PRE**	Pre-tests	[15]	Subj.	Both	
	POS	Post-tests	[121]	Subj.	Both	
	PAS	Pattern-of-Search	[62]	Obj.	Both	
	LOP	Level of Pressure	[62]	Obj.	Quant.	
	VAK	Learning Styles Self-Report Questionnaire	[34]	Subj.	Both	
Motion sickness	**SSQ**	Simulator Sickness Questionnaire (Fatigue)	[55]	Subj.	Quant.	
Memory	**ASC**	Awareness State score	[25]	Subj.	Quant.	
	ANAM	Automated Neuropsychological Assessment Metrics	[53]	Subj.	Quant.	
	CUED	Cue recall	[99]	Subj.	Both	
	FREE	Free recall	[99]	Subj.	Both	

Table 6: A summary of research topics in MR/AR by paper type, evaluation scenario, cognitive aspect, and configuration.

Topic	Paper Type						Evaluation Scenario							Cognitive Aspect										Configuration			
	Technique	Evaluation	Design Study	System	Model	Total	AP	QRI	UE	UP	COL	COM	UWP	Perception	Usability	Emotion	Decision making	Cognitive load	Presence	Learning	Attention	Motion sickness	Memory	Static\Lab	Mobile\Lab	Mobile\In-the-Wild	Static\In-the-Wild
Design/human factors	11	69	8	0	1	89	7	2	57	57	6	4	1	23	14	13	7	15	11	7	4	3	6	43	32	9	1
Tracking	65	3	0	1	0	69	62	49	4	7	1	0	0	3	2	0	3	0	0	0	0	0	0	2	4	0	0
Reconstruction	36	1	1	4	0	42	30	33	5	5	0	0	0	2	1	1	0	2	0	2	0	1	0	4	4	0	0
Multimodal interfaces	17	12	7	1	0	37	6	7	21	23	0	1	1	14	7	5	7	3	3	0	2	1	0	24	6	2	0
Applications	15	15	2	2	0	34	9	9	20	15	0	1	4	1	7	5	8	3	0	4	1	0	1	8	4	8	1
Displays	32	1	0	1	0	34	23	23	8	8	1	1	0	6	1	2	1	0	1	0	1	0	0	8	2	0	0
HCI technologies	20	12	1	1	0	34	9	7	21	22	2	1	0	3	6	3	5	1	3	0	1	1	0	16	7	3	0
Rendering	26	1	0	0	0	27	19	23	7	6	0	1	0	4	2	0	1	0	0	0	0	0	0	6	2	0	0
Collaboration	5	9	6	2	0	22	4	2	17	7	14	7	0	1	5	4	0	1	7	0	1	1	1	9	9	2	0
Calibration/registration	17	0	0	0	0	17	15	10	1	2	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0
Mediated reality	9	1	2	2	0	14	5	4	7	8	1	0	0	4	1	0	1	0	1	1	1	2	0	7	2	0	0
Spatial annotation	11	1	1	1	0	14	6	2	8	10	0	1	0	2	4	2	1	4	0	2	1	1	0	5	3	3	0
Diminished reality	8	1	0	0	0	9	5	8	0	4	0	0	0	3	0	0	1	0	0	0	0	0	0	4	0	0	0
Data visualization	3	2	2	0	0	7	0	0	5	5	1	2	0	2	1	0	1	0	0	0	2	0	0	2	2	3	0
Taxonomy	0	2	0	0	5	7	0	0	2	0	0	0	4	1	1	1	0	0	0	0	0	0	0	0	0	0	0
Software architecture	2	0	0	0	0	2	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Total	277	130	30	15	6	458	202	179	183	179	26	19	10	69	53	36	36	29	26	16	14	10	8	138	78	30	2

Forced-Choice (2-IFC) method [61], which is similar to 2-AFC, but in which options are presented sequentially in two intervals. The studies analyzed two variables: (a) Just Noticeable Difference (JND), which is the point at which participants do not perceive differences between two similar options, and (b) Point of Subjective Equality (PSE), which is the point at which participants perceive options of different nature as equal. *Auditory.* Studies that analyzed, in particular, the relationship between auditory perception and spatial perception employed the Spatial Audio Quality Inventory (SAQI) vocabulary [40], which is intended for a qualitatively comparative auditory assessment of acoustic scenes.

Usability. We found that 21% of user studies (53/248) assessed the usability of MR/AR approaches. These studies often focused on design/human factors (14/53) and involved methods such as: (a) The Slater-Usoh-Steed Questionnaire (SUS) [16,23,99] , (b) Mixed Reality Experience Questionnaire (MREQ) [97], and (c) Post Experience Questionnaire (PEQ) [97] .

Emotions. We found that 15% of user studies (36/248) examined the emotions of participants. We did not identify studies that applied objective methods to analyze emotions. Instead, we observed that studies often collect data of the emotions perceived by participants to examine various emotions using methods such as: (a) Game Experience Questionnaire (GEQ) [23] to measure game experience based on user engagement *e.g.*, competence, sensory and imaginative immersion, flow, challenge, positive affect, negative affect, tension, and annoyance; (b) Intrinsic Motivation Inventory (IMI) [46] to measure the overall user experience with regard to general experimental tasks. Other methods used to assess emotions in general were: (a) Positive and Negative Affect Schedule (PANAS) [27,28] to assess positive and negative affect and (b) Self-Assessment Manikin (SAM) [27, 28], which is a pictorial assessment technique to measure pleasure, arousal, and dominance associated with a person's affective impressions.

Decision making. We found that 15% of user studies (36/248) analyzed decision making. Such studies complemented an analysis of the time and the correctness of participants to complete tasks by collecting data of head movements (*e.g.*, rotation, exertion, or velocity) to examine effort, efficiency, and effectiveness [47, 75].

Cognitive load. We found that 12% of user studies (29/248) analyzed the cognitive load of participants exposed to MR/AR approaches. Cognitive load involves two concepts: (a) mental load imposed by instructional parameters, *e.g.*, task structure, the sequence of information given during an evaluation, and (b) mental effort that refers to the capacity allocated by participants of a study to the instructional demands. Therefore, when evaluating cognitive load in laboratory settings, the mental load can be fixed and kept the same across the evaluated conditions. Thus, measures of mental effort can be considered an index of cognitive load. Studies that examined cognitive load used objective data collection methods, for instance, to examine the anxiety of participants. There exist several physiological measures that have been examined to analyze mental effort, such as pupil dilation [52], heart rate variability [83], event-related brain potentials [29], muscle tension [111], adrenaline level [36], skin temperature and galvanic skin response (GSR) [38]. Studies also used subjective methods such as: (a) Paas 9-step mental-effort Likert scale (Paas) [13, 16, 112], in which score indexes of metal effort go from "very, very low effort" (1) to "very, very high effort" (9); (b) Subjective Mental Effort Question (SMEQ) [69], in which participants indicate their mental effort using a scale that goes from "not at all hard to do" (0) to "tremendously hard to do" (150); and (c) NASA-TLX [51, 90, 102] that is the assessment of total workload divided into six subscales: mental demand, physical demand, temporal demand, performance effort, and frustration.

Presence. We found that 10% of user studies (26/248) analyzed presence. Studies [38] that examine the feeling of presence use several objective methods to measure presence-based physiological responses: (a) Electrocardiogram (ECG), (b) Galvanic Skin Response (GSR), (c) skin temperature, (d) brain activity (*e.g.*, EEG), (e) heart rate, and (f) respiration rate.

Studies can also use several subjective data collection methods. The methods are based on questionnaires that participants in a study are asked to fill to cover various aspects of presence *e.g.*, co-presence, spatial presence, social presence, social richness, closeness, or connectedness. Other methods that can be used to complement the analysis of presence are: (a) Ad-Hoc Usability Questionnaire (AD3) [51] to analyze the *awareness* of participants in an immersive environment; (b) Body Representation Questionnaire (BRQ) [97] for the analysis of embodiment; (c) McKnight Trust Questionnaire (MTQ) [57] to assess whether participants trust in technology, *e.g.*, reliability, helpfulness, functionality, and situational normality.

Simulator Sickness Questionnaire (SSQ) [18, 51, 91]. SSQ is a well-known test to check symptoms of nausea, fatigue, and disorientation, which could affect the integrity of participants, and in consequence, the results of the evaluation. SSQ can be a suitable complement to the analysis of multiple cognitive aspects. However, we often observed its application in the assessment of presence.

Learnability. We found that 7% of user studies (16/248) analyzed learnability promoted by MR/AR approaches. Studies focused on learnability used general methods such as pre-tests to assess the prior knowledge of participants of a subject. The results offered researchers a baseline for comparing to the results of a post-test. A significant increase in the measured knowledge suggested that the approach under analysis promoted learnability. Other specific objective methods employed in the analysis of learnability of particular topics are (a) Level of Pressure (LOP) [62] to assess learners use of correct pressure or (b) Pattern-of-Search (PAS) [62] that is applicable, for instance, to medical training of breast exams. The study of learnability was complemented employing a classification the learning styles self-assessment questionnaire (VAK) [121]. Learning styles are characterized based on (a) the use of seen (visual), (b) the transfer of information through listening (auditory), and (c) physical experience (kinesthetic). However, we notice that the method has some detractors [103].

Attention. We found that 6% of user studies (14/248) analyzed the attention of participants of an MR/AR evaluation. Studies that involved perception used various methods depending on the type of perception.

Visual. Studies selected objective methods to collect data from eye- and head-tracking. Using eye-tracking, experimenters analyzed when participants were distracted. Eye-tracking complemented with an analysis of head movements (*e.g.*, orientation angle of participants heads) indicated when participants were distracted as well. Visual attention was also analyzed based on subjective methods that consider, for instance, the assessment of engagement, through (a) Ad-hoc Self-Report Questionnaire of Engagement (AD4) [40] or (b) User Engagement Self-Report Questionnaire (VisEngage) [74]. *Auditory.* Studies examined, in particular, the connection between listener sensitivity in audio localization and experienced attention for immersion in MR. Subjective auditory attention was measured using an ad-hoc self-report questionnaire (AD4). *Eyesight tests*. Snellen eye charts [88] were used to ensure the visual aptitude of participants of studies that involved visual attention.

Memory. We found that 3% of user studies (8/248) analyzed the recollection of participants of an MR/AR evaluation. Studies that examined memory ensured first that participants in an evaluation had a normal memory ability. To this end, they used the Automated Neuropsychological Assessment Metrics (ANAM) survey [99]. Other general methods for collecting data on the recollection of participants in evaluations of MR/AR were: (a) Free Recall, in which participants were asked to tell a narrative from their recollection; (b) Cued Recall, in which participants were asked questions to drive

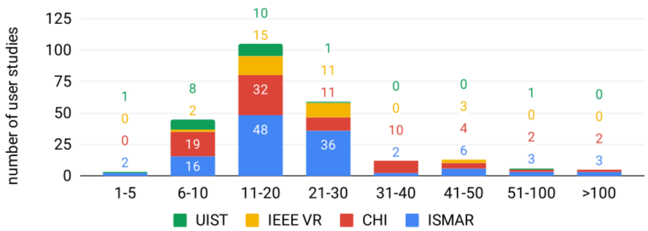

Figure 10: Histogram of the number of participants in user studies in MR/AR.

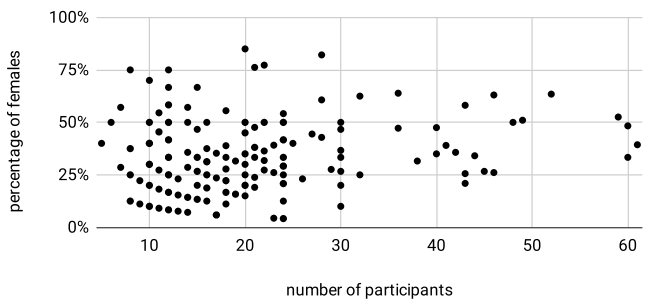

Figure 11: The sample sizes by gender of 204 of the 248 user evaluations in MR/AR (44 studies did not report on gender).

their recollection in order to recognize relevant points that were missing in their narrative. Since the confidence of participants in their recollection varied sometimes, researchers complemented the analysis with measures of participants' level of confidence using the Awareness State score [25, 99].

4.5 Study Configuration

We found that 54% of all papers reported on evaluations with users (248/458). One important choice when designing a user evaluation is to define the number of participants that are going to be involved. We found that user evaluations involved 5,761 participants in total, of which only 1,619 were identified as female and 3,087 as male. Fig. 10 shows a histogram of the distribution of the samples sizes that we found among the reported user studies. We found that 93% of user evaluations (231/248) include between 3 and 30 participants (median of 15) of which a median of 4 were females. The remaining 7% of user evaluations include a median of 44 participants overall and a median of 16 female participants. The sample sizes in MR/AR seem larger than in HCI in general. A previous study [19] found that the most common sample size is 12 in evaluations published in CHI. Similarly, in visualization, most user studies involve ten or fewer participants [50]. We found that 42% of the users studies (105/248) include 11–20 participants, and user studies that involve a smaller number of participants (*i.e.*, 1–5) are not frequent in MR/AR. Such smaller numbers are more common in qualitative studies, which do not seem to be as frequent in MR/AR as in other domains. The distribution of the number of participants involved in MR/AR user studies is similar across the four analyzed venues.

We further examine the number of participants split by gender. We present in Fig. 11 a scatterplot with the the percentage of females versus the total number of participants. We observe that only a few studies involve a majority of female participants. We excluded from the chart 44 studies that do not specify the gender of participants.

We also analyzed whether user studies were conducted in a laboratory setting or in-the-wild, as well as, whether participants in the study where static (*i.e.*, sitting or standing) or mobile. The results are shown in Fig. 12. We split the results by venue. We observe that the majority of user studies (*i.e.*, 82% CHI–100% IEEE VR) are conducted in a laboratory. Amongst such studies, in UIST, 50% of user studies involve participants using MR/AR in a mobile way, whereas in IEEE VR, only 8% do so. In contrast, user studies occasionally are conducted in-the-wild. User studies conducted in-the-wild usually involve participants in a mobile fashion, for example, to support needs in the automotive industry [37, 80, 80, 115]. We only found two in-the-wild studies in which participants were static [86, 121]. We did not identify in-the-wild studies published in IEEE VR.

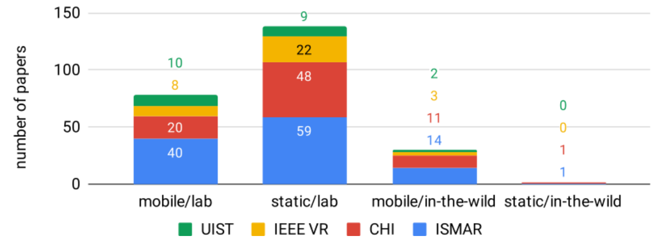

Figure 12: A classification of the 248 user studies in MR/AR per static/mobile and lab/in-the-wild configuration.

5 Discussion and Guidance

We now discuss considerations of cross-cutting concerns that emerged from the analyzed dimensions that can guide researchers to design evaluations in MR/AR.

Bridging technology-centric and human-centric evaluations. We found two main types of MR/AR papers: (a) technology-centric papers (293/458) that focus on topics such as tracking, displays, reconstruction, rendering, or calibration, which are frequently evaluated through AP+QRI scenarios, and (b) human-centric papers (160/458) that deal with topics such as applications, design/human factors, which are frequently evaluated through UP+UE scenarios. There seem to be a few small crossover topics, *i.e.*, spatial annotation, multimodal interfaces, or collaboration. However, technology-centric and human-centric approaches use methodologies that are mostly disjoint. Consequently, we ask how we can bridge technology-centric and human-centric evaluations when this combination helps address a research question. In fact, we have observed some papers that followed such pattern [14, 39, 41, 54, 68, 93, 96, 104, 109, 116, 117]. We call researchers in the field to complement the results of benchmarks with user studies when techniques involve a user interface and the combination of evaluations contributes to the research question at hand.

Toward comprehensive cognitive methods. We identified that 236 out of 297 cognitive aspects are involved in evaluations of design/human factors, multimodal interfaces, applications, HCI technologies, collaboration, and spatial annotation (see Table 5). Based on their frequency (see Table 6), major concerns for (a) design/human factors are perception, presence, and cognitive load, (b) applications are decision making, emotions, and cognitive load/perception, (c) multimodal interfaces are perception and emotions, and (d) spatial annotations are cognitive load, presence, and emotions. We observe that the wide range of inherent topics to MR/AR can pose a challenge for newcomers to the field. We think our paper can help newcomers with such directions.

We found 34 subjective and nine objective data collection methods (see Table 5). We observe that evaluations that included objective data collection methods usually reported difficulties encountered when analyzing and interpreting the collected data (*e.g.*, noisy data) [38] and represent a valuable source of knowledge. Although subjective questionnaires can be helpful to collect impressions of the perception of participants, we observe a need for objective data that avoids biases of subjective data collection methods. For instance, participants in a study might be biased to share positive rather than negative emotions. Objective data are not limited to physiological sources, *e.g.*, ECG, EEG, GSR, but can include behavioral measures as well, *e.g.*, eye-, head-, and body-tracking. We call researcher in the field to involve multiple cognitive aspects that are suitable to comprehensively examine a given research question.

Gender bias limitations due lack of female participants. Although our results show that participants populations in MR/AR are larger than populations in studies in HCI [19] and visualization [50], we found that often such populations exhibit unbalanced genders (see Fig. 11). Certainly, the unbalanced gender of participants in studies is related to the recruiting strategies, which often rely on the (predominantly male) student population in computer science programs. However, gender bias can be an important concern in multiple fields [24] and should not be taken lightly. Recruiting an adequate number of participants is a step to more credible quantitative evaluation, but it should not be done at the expense of introducing gender bias. We call researchers in the field to involve an adequate participant population with a more inclusive and gender-balanced distribution.

Increasing the ecological validity of evaluations. The papers that we classified in the design study type (30), the applications topic (34), the UWP scenario (10), or the in-the-wild configuration (32) report on MR/AR approaches that were successfully used to deal with concrete real-world cases. Studying a real-world use case in-situ is clearly challenging, but, with more mature enabling technology at our disposal, successful real-world deployment has become realistic. We call to researchers in the field to conduct case studies that can help investigate in-depth phenomena in a concrete real-world case. We found that 87% of user evaluations (216/248) are conducted in a laboratory setting, of which 78 involve participants in a mobile fashion. We found that 13% of studies (32/248) are conducted in-the-wild, which often involve participants in a mobile setting. We call researchers in the field to increase ecological validity by conducting in-the-wild evaluations with mobile or static participants depending on the targeted user behavior.

Depth-first (re)search: qualitative vs. quantitative analysis. We observe that thorough evaluations that entail qualitative approaches (*e.g.*, case studies) can facilitate a deep understanding of a phenomenon [20]. However, we found that only 8 out of 43 different data collection methods used in MR/AR can support qualitative analysis (see Table 5). Due to the nature of qualitative data, evaluations that adopt a qualitative approach usually involve a limited number of participants, and thus, the results of these evaluations are hard to generalize to a large population. In contrast, evaluations that use quantitative methods can involve a higher number of participants, and, through the use of statistical analysis, can generalize results. However, such studies need to be highly controlled, hampering the application of results under real-world circumstances (ecological validity). Picking the right method is a well-known trade-off process with different methods offering different benefits and drawbacks [78]. We think that to formulate appropriate hypotheses that can be thoroughly tested and generalized using quantitative methods, first, researchers need to obtain a deep understanding of the examined phenomenon and identify which qualitative methods are suited best. An approach of depth-first and breadth-second (re)search is clearly underrepresented in the surveyed MR/AR work, and we speculate that it may offer interesting findings in several cases.

Where does my paper fit best: ISMAR, CHI, IEEE VR, UIST? We now discuss the types of work that are mostly accepted at these venues. Certainly, ISMAR is the main venue for MR/AR. Whereas in the past most papers published in ISMAR centered on techniques, today papers mostly focus on human-centered evaluations. Usually, ISMAR papers describe thorough evaluations that involve the analysis of multiple cognitive aspects to help researchers address the analysis of complex phenomena. CHI exhibits an increasing trend of MR/AR papers with a balanced interest in techniques, evaluations, and, to a lesser degree, design studies. Although CHI papers mostly focus on design, human factors, and HCI technologies, there is also interest in research that focuses on multimodal interfaces and collaboration. Recently, IEEE VR and CHI doubled the number of MR/AR papers with a balanced number of techniques and evaluations. IEEE VR papers mostly focus on design and human factors. In UIST, there is a small and consistent number of MR/AR papers published every year, which mostly correspond to techniques. Often, technique papers published in UIST focus on reconstruction whereas evaluation papers focus on HCI technologies.

The future of MR/AR. In the future, we expect that the number of MR/AR papers will keep increasing and remain balanced amongst ISMAR, CHI, and IEEE VR, and, to a lesser extent, UIST. As MR/AR technologies become more mature, questions that involve human aspects will gain focus in MR/AR research. Consequently, we expect that future MR/AR papers will elaborate on human-centered evaluations that involve not only the analysis of user performance and user experience, but also the analysis of other scenarios, like understanding the role of MR/AR in working places and in communication and collaboration. Hence, we envision that there will be an increasing need for developing methods that support researchers to deal with such scenarios, which might involve in-the-wild configurations. Our results confirm that MR/AR is a very complex technology. We observe that, even in laboratory settings, it is difficult to conduct a user study in which higher-level cognition is tested without being confounded by imperfections of MR/AR technology. For example, it is very hard to perform long-term studies if mobile devices run out of batteries. Tasks that require much time will likely not be carried out completely. That could explain why authors tend to report on "low-level" user performance, which is often complemented with questionnaires like NASA TLX to document cognition aspects. The rise of commercial-grade devices like the Microsoft HoloLens lowers the barrier of having standardized conditions for evaluations, but we have yet to see this has an effect in publications.

6 Conclusion

We analyzed evaluations reported in 485 papers of the research literature of MR/AR. We confirmed that (a) technology-centric evaluations (through benchmarks of tracking, displays, reconstruction, rendering, and calibration) and (b) human-centric evaluations (of applications, and design/human factors) are the core pillars of MR/AR evaluation. We found a marginal number of team-centric evaluations that involve collaboration, communication, and understanding environments and work practices. We call researchers in the field to conduct thorough evaluations by: (a) conducting user studies that complement the results of benchmarks when techniques involve a user interface and the combination is coherent with a research question; (b) involving multiple cognitive aspects that can help comprehensive examination of a research question; (c) choosing appropriate methods for assessing the impact of an approach in human cognition, for which they can consult our selected examples; (d) involving an adequate participant population with a more inclusive gender-balanced distribution; (e) increasing the ecological validity of evaluations through in-the-wild and mobile or static configurations depending on the intended user behavior.

Acknowledgments

Merino, Kraus, and Weiskopf acknowledge funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 251654672 – TRR 161. Schwarzl and Sedlmair were supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2120/1 – 390831618.

References

[1] ACM DL. https://dl.acm.org. Accessed on 2020-05-01.

[2] CHI. http://chi.acm.org/. Accessed on 2020-05-01.

[3] "Cognition". Lexico. https://www.lexico.com. Accessed on 2020-05-06.

[4] CORE. http://portal.core.edu.au/conf-ranks/. Accessed on 2020-05-01.

[5] IEEE VIS. http://ieeevis.org/. Accessed on 2020-05-01.

[6] IEEE VR. hhttp://ieeevr.org/. Accessed on 2020-05-01.

[7] IEEE Xplore. https://ieeexplore.ieee.org. Accessed on 2020-05-01.

[8] ISMAR. http://ismar.net/. Accessed on 2020-05-01.

[9] UIST. http://uist.acm.org. Accessed on 2020-05-01.

[10] R. W. Adams. *Peripheral vision and visual attention*. PhD thesis, Iowa State University, 1971.

[11] A. Aron, E. Aron, and D. Smollan. Inclusion of other in the self scale and the structure of interpersonal closeness. *Journal of Personality and Social Psychology*, 63:596–612, Oct. 1992.

[12] D. Banakou, R. Groten, and M. Slater. Illusory ownership of a virtual child body causes overestimation of object sizes and implicit attitude changes. *National Academy of Sciences*, 110(31):12846–12851, 2013.

[13] J. Baumeister, S. Y. Ssin, N. A. M. ElSayed, J. Dorrian, D. P. Webb, J. A. Walsh, T. M. Simon, A. Irlitti, R. T. Smith, M. Kohler, and B. H. Thomas. Cognitive cost of using augmented reality displays. *Transactions on Visualization and Computer Graphics*, 23(11):2378–2388, 2017.

[14] D. Baur, S. Boring, and S. Feiner. Virtual projection: Exploring optical projection as a metaphor for multi-device interaction. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, pp. 1693–1702, 2012.

[15] F. Bork, R. Barmaki, U. Eck, K. Yu, C. Sandor, and N. Navab. Empirical study of non-reversing magic mirrors for augmented reality anatomy learning. In *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 169–176, 2017.

[16] F. Bork, C. Schnelzer, U. Eck, and N. Navab. Towards efficient visual guidance in limited field-of-view head-mounted displays. *Transactions on Visualization and Computer Graphics*, 24(11):2983–2992, 2018.

[17] M. M. Bradley and P. J. Lang. Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, 25(1):49–59, 1994.

[18] G. Bruder, P. Wieland, B. Bolte, M. Lappe, and F. Steinicke. Going with the flow: Modifying self-motion perception with computer-mediated optic flow. In *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 67–74, 2013.

[19] K. Caine. Local standards for sample size at CHI. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, pp. 981–992, 2016.

[20] S. Carpendale. Evaluating information visualizations. In A. Kerren, J. T. Stasko, J.-D. Fekete, and C. North, eds., *Information Visualization: Human-Centered Issues and Perspectives*, pp. 19–45. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[21] L. Chen, T. W. Day, W. Tang, and N. W. John. Recent developments and future challenges in medical mixed reality. In *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 123–135, 2017.

[22] J.-H. Choe, T.-U. Jeong, H. Choi, E.-J. Lee, S.-W. Lee, and C.-H. Lee. Subjective video quality assessment methods for multimedia applications. *Journal of Broadcast Engineering*, 18(3):416–424, 1999.

[23] M. A. Cidota, P. J. M. Bank, P. W. Ouwehand, and S. G. Lukosch. Assessing upper extremity motor dysfunction using an augmented reality game. In *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 144–154, 2017.

[24] P. Coiro and D. D. Pollak. Sex and gender bias in the experimental neurosciences: the case of the maternal immune activation model. *Translational Psychiatry*, 9(1):90, 2019.

[25] M. Coxon and K. Mania. Measuring memories for objects and their locations in immersive virtual environments: The subjective component of memorial experience. In *Handbook of Human Centric Visualization*, pp. 453–471. Springer, New York, NY, 2014.

[26] A. Dey, M. Billinghurst, R. W. Lindeman, and J. Swan. A systematic review of 10 years of augmented reality usability studies: 2005 to 2014. *Frontiers in Robotics and AI*, 5:37, 2018.

[27] A. Dey, H. Chen, A. Hayati, M. Billinghurst, and R. W. Lindeman. Sharing manipulated heart rate feedback in collaborative virtual environments. In *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 248–257, 2019.

[28] A. Dey, H. Chen, C. Zhuang, M. Billinghurst, and R. W. Lindeman. Effects of sharing real-time multi-sensory heart rate feedback in different immersive collaborative virtual environments. In *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 165–173, 2018.

[29] E. Donchin. Surprise!... surprise? *Psychophysiology*, 18(5):493–513, 1981.

[30] A. Duenser, R. Grasset, and M. Billinghurst. A survey of evaluation techniques used in augmented reality studies. Technical report, Human Interface Technology Laboratory New Zealand, 2008.

[31] N. Elmqvist and J. S. Yi. Patterns for visualization evaluation. *Information Visualization*, 14(3):250–269, 2015.

[32] J. A. Ferwerda. Psychophysics 101: how to run perception experiments in computer graphics. In *ACM SIGGRAPH 2008 classes*, pp. 1–60. 2008.

[33] P. Fite-Georgel. Is there a reality in industrial augmented reality? In *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 201–210, 2011.

[34] N. Fleming and D. Baume. Learning styles again: Varking up the right tree! *Educational developments*, 7(4):4, 2006.

[35] A. Fonnet and Y. Prié. Survey of immersive analytics. *Transactions on Visualization and Computer Graphics*, pp. 1–1, 2019.

[36] M. Frankenhaeuser. Experimental approaches to the study of catecholamines and emotion. In *Proceedings of the Symposium on Parameters of Emotion*, p. 684685, 1975.

[37] W.-T. Fu, J. Gasper, and S.-W. Kim. Effects of an in-car augmented reality system on improving safety of younger and older drivers. In *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 59–66, 2013.

[38] M. Gandy, R. Catrambone, B. MacIntyre, C. Alvarez, E. Eiriksdottir, M. Hilimire, B. Davidson, and A. C. McLaughlin. Experiences with an AR evaluation test bed: Presence, performance, and physiological measurement. In *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 127–136, 2010.

[39] C. Gebhardt, B. Hecox, B. van Opheusden, D. Wigdor, J. Hillis, O. Hilliges, and H. Benko. Learning cooperative personalized policies from gaze data. In *Proceedings of the ACM Symposium on User Interface Software and Technology (UIST)*, pp. 197–208, 2019.

[40] M. Geronazzo, E. Sikstrm, J. Kleimola, F. Avanzini, A. de Gtzen, and S. Serafin. The impact of an accurate vertical localization with HRTFs on short explorations of immersive virtual reality scenarios. In *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 90–97, 2018.

[41] L. Gruber, S. Gauglitz, J. Ventura, S. Zollmann, M. Huber, M. Schlegel, G. Klinker, D. Schmalstieg, and T. Hllerer. The city of sights: Design, construction, and measurement of an augmented reality stage set. In *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 157–163, 2010.

[42] J. Grubert, T. Langlotz, S. Zollmann, and H. Regenbrecht. Towards pervasive augmented reality: Context-awareness in augmented reality. *Transactions on Visualization and Computer Graphics*, 23(6):1706–1724, 2017.

[43] K. Gupta, G. A. Lee, and M. Billinghurst. Do you see what I see? The effect of gaze tracking on task space remote collaboration. *Transactions on Visualization and Computer Graphics*, 22(11):2413–2422, 2016.

[44] C. Harms and F. Biocca. Internal consistency and reliability of the networked minds measure of social presence. In M. Alcaniz and B. Rey, eds., *Proceedings of the International Workshop on Presence*, 2004.

[45] S. G. Hart and L. E. Stavenland. Development of NASA-TLX (task load index): Results of empirical and theoretical research. In P. A.

Hancock and N. Meshkati, eds., *Human Mental Workload*, chap. 7, pp. 139–183. Elsevier, North-Holland, 1988.

[46] A. Hartl, J. Grubert, D. Schmalstieg, and G. Reitmayr. Mobile interactive hologram verification. In *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 75–82, 2013.

[47] S. J. Henderson and S. Feiner. Evaluating the benefits of augmented reality for task localization in maintenance of an armored personnel carrier turret. In *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 135–144, 2009.

[48] A. Ibrahim, B. Huynh, J. Downey, T. Höllerer, D. Chun, and J. O'Donovan. ARbis pictus: A study of vocabulary learning with augmented reality. *Transactions on Visualization and Computer Graphics*, 24(11):2867–2874, 2018.

[49] W. Ijsselsteijn, W. Van Den Hoogen, C. Klimmt, Y. De Kort, C. Lindley, K. Mathiak, K. Poels, N. Ravaja, M. Turpeinen, and P. Vorderer. Measuring the experience of digital game enjoyment. In *Proceedings of the International Conference on Methods and Techniques in Behavioral Research*, pp. 88–89, 2008.

[50] T. Isenberg, P. Isenberg, J. Chen, M. Sedlmair, and T. Möller. A systematic review on the practice of evaluating visualization. *Transactions on Visualization and Computer Graphics*, 19(12):2818–2827, 2013.

[51] J. Jung, H. Lee, J. Choi, A. Nanda, U. Gruenefeld, T. Stratmann, and W. Heuten. Ensuring safety in augmented reality from trade-off between immersion and situation awareness. In *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 70–79, 2018.

[52] D. Kahneman. *Attention and Effort*. Prentice-Hall, Englewood Cliffs, New Jersey, 1973.

[53] R. L. Kane, T. Roebuck-Spencer, P. Short, M. Kabat, and J. Wilken. Identifying and monitoring cognitive deficits in clinical populations using automated neuropsychological assessment metrics (ANAM) tests. *Archives of Clinical Neuropsychology*, 22:115–126, 2007.

[54] A. Karnik, W. Mayol-Cuevas, and S. Subramanian. MUSTARD: A multi user see through ar display. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, pp. 2541–2550, 2012.

[55] R. S. Kennedy, J. M. Drexler, D. E. Compton, K. M. Stanney, D. S. Lanham, and D. L. Harm. Configural scoring of simulator sickness, cybersickness and space adaptation syndrome: Similarities and differences. *Virtual and adaptive environments: Applications, implications, and human performance issues*, p. 247, 2003.

[56] K. Kim, M. Billinghurst, G. Bruder, H. B.-L. Duh, and G. F. Welch. Revisiting trends in augmented reality research: A review of the 2nd decade of ISMAR (2008–2017). *Transactions on Visualization and Computer Graphics*, 24(11):2947–2962, 2018.

[57] K. Kim, L. Boelling, S. Haesler, J. Bailenson, G. Bruder, and G. F. Welch. Does a digital assistant need a body? The influence of visual embodiment and social behavior on the perception of intelligent virtual agents in AR. In *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 105–114, 2018.

[58] S. Kim, G. Lee, W. Huang, H. Kim, W. Woo, and M. Billinghurst. Evaluating the combination of visual communication cues for HMD-based mixed reality remote collaboration. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, 2019. Paper 173.

[59] S. Kim, G. Lee, N. Sakata, and M. Billinghurst. Improving copresence with augmented visual communication cues for sharing experience through video conference. In *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 83–92, 2014.

[60] B. A. Kitchenham, S. L. Pfleeger, L. M. Pickard, P. W. Jones, D. C. Hoaglin, K. E. Emam, and J. Rosenberg. Preliminary guidelines for empirical research in software engineering. *Transactions on Software Engineering*, 28(8):721–734, 2002.

[61] B. Knorlein, M. Di Luca, and M. Harders. Influence of visual and haptic delays on stiffness perception in augmented reality. In *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 49–52, 2009.

[62] A. Kotranza, D. Scott Lind, C. M. Pugh, and B. Lok. Real-time in-situ visual feedback of task performance in mixed environments for learning joint psychomotor-cognitive tasks. In *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 125–134, 2009.

[63] M. Krichenbauer, G. Yamamoto, T. Taketomi, C. Sandor, and H. Kato. Towards augmented reality user interfaces in 3D media production. In *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 23–28, 2014.

[64] K. Krippendorff. *Content Analysis: An Introduction to its Methodology*. Sage Publications, California, 2018.

[65] E. Kruijff, J. E. Swan, and S. Feiner. Perceptual issues in augmented reality revisited. In *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 3–12, 2010.

[66] H. Lam, E. Bertini, P. Isenberg, C. Plaisant, and S. Carpendale. Empirical studies in information visualization: Seven scenarios. *Transactions on Visualization and Computer Graphics*, 18(9):1520–1536, 2012.

[67] J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, pp. 159–174, 1977.

[68] T. Langlotz, M. Cook, and H. Regenbrecht. Real-time radiometric compensation for optical see-through head-mounted displays. *Transactions on Visualization and Computer Graphics*, 22(11):2385–2394, 2016.

[69] G. A. Lee, T. Teo, S. Kim, and M. Billinghurst. A user study on MR remote collaboration using live 360 video. In *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 153–164, 2018.

[70] J. R. Lewis. IBM computer usability satisfaction questionnaires: psychometric evaluation and instructions for use. *International Journal of Human-Computer Interaction*, 7(1):57–78, 1995.

[71] A. Lindau, V. Erbes, S. Lepa, H.-J. Maempel, F. Brinkmann, and S. Weinzierl. A spatial audio quality inventory (SAQI). *Acta Acustica united with Acustica*, 100(5):984–994, 2014.

[72] J. Llobera, M. V. Sanchez-Vives, and M. Slater. The relationship between virtual body ownership and temperature sensitivity. *Journal of the Royal Society*, 10(85):1–11, 2013.

[73] M. Lombard, T. B. Ditton, and L. Weinstein. Measuring presence: the temple presence inventory. In *Proceedings of the International Workshop on Presence*, pp. 1–15, 2009.

[74] F. Lu, D. Yu, H. Liang, W. Chen, K. Papangelis, and N. M. Ali. Evaluating engagement level and analytical support of interactive visualizations in virtual reality environments. In *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 143–152, 2018.

[75] M. R. Marner, A. Irlitti, and B. H. Thomas. Improving procedural task performance with augmented reality annotations. In *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 39–48, 2013.

[76] K. Marriott, F. Schreiber, T. Dwyer, K. Klein, N. H. Riche, T. Itoh, W. Stuerzlinger, and B. H. Thomas. *Immersive Analytics*. Springer, Cham, 2018.

[77] E. McAuley, T. Duncan, and V. V. Tammen. Psychometric properties of the intrinsic motivation inventory in a competitive sport setting: A confirmatory factor analysis. *Research Quarterly for Exercise and Sport*, 60(1):48–58, 1989.

[78] J. E. McGrath. Methodology matters: Doing research in the behavioral and social sciences. In R. M. Baecker, J. Grudin, W. A. Buxton, and S. Greenberg, eds., *Readings in Human–Computer Interaction*, Interactive Technologies, pp. 152–169. Morgan Kaufmann, Burlington, 1995.

[79] D. H. Mcknight, M. Carter, J. B. Thatcher, and P. F. Clay. Trust in a specific technology: An investigation of its components and measures. *Transactions on Management Information Systems*, 2(2):12:1–12:25, 2011.

[80] C. Merenda, H. Kim, K. Tanous, J. L. Gabbard, B. Feichtl, T. Misu, and C. Suga. Augmented reality interface design approaches for goal-directed and stimulus-driven driving tasks. *Transactions on Visualization and Computer Graphics*, 24(11):2875–2885, 2018.

[81] L. Merino, M. Ghafari, C. Anslow, and O. Nierstrasz. A systematic literature review of software visualization evaluation. *Journal of Systems and Software*, 144:165–180, 2018.

[82] L. Merino, M. Schwarzl, M. Kraus, M. Sedlmair, D. Schmalstieg, and D. Weiskopf. Dataset: Evaluating Mixed and Augmented Reality: A Systematic Literature Review (2009–2019), Mar. 2020. doi: 10.5281/zenodo.3832114

[83] G. Mulder. *The Heart of Mental Effort*. PhD thesis, University of Groningen, The Netherlands, 1980.

[84] T. Munzner. Process and pitfalls in writing information visualization research papers. In A. Kerren, J. T. Stasko, J.-D. Fekete, and C. North, eds., *Information Visualization: Human-Centered Issues and Perspectives*, pp. 134–153. Springer, Berlin, Heidelberg, 2008.

[85] T. Olsson and M. Salo. Online user survey on current mobile augmented reality applications. In *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 75–84, 2011.

[86] L. Oppermann, C. Putschli, C. Brosda, O. Lobunets, and F. Prioville. The smartphone project: An augmented dance performance. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, pp. 2569–2572, 2015.

[87] J. Orlosky, P. Kim, K. Kiyokawa, T. Mashita, P. Ratsamee, Y. Uranishi, and H. Takemura. VisMerge: Light adaptive vision augmentation via spectral and temporal fusion of non-visible light. In *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 22–31, 2017.

[88] J. Orlosky, T. Toyama, K. Kiyokawa, and D. Sonntag. ModulAR: Eye-controlled vision augmentations for head mounted displays. *Transactions on Visualization and Computer Graphics*, 21(11):1259–1268, 2015.

[89] F. Paas. Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach. *Journal of Educational Psychology*, 84:429–434, Dec. 1992.

[90] T. Piumsomboon, D. Altimira, H. Kim, A. Clark, G. Lee, and M. Billinghurst. Grasp-shell vs gesture-speech: A comparison of direct and indirect natural interaction techniques in augmented reality. In *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 73–82, 2014.

[91] T. Piumsomboon, G. A. Lee, B. Ens, B. H. Thomas, and M. Billinghurst. Superman vs giant: A study on spatial perception for a multi-scale mixed reality flying telepresence interface. *Transactions on Visualization and Computer Graphics*, 24(11):2974–2982, 2018.

[92] T. Piumsomboon, G. A. Lee, A. Irlitti, B. Ens, B. H. Thomas, and M. Billinghurst. On the shoulder of the giant: A multi-scale mixed reality collaboration with 360 video sharing and tangible interaction. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, 2019. Paper 228.

[93] L. Qian, A. Plopski, N. Navab, and P. Kazanzides. Restoring the awareness in the occluded visual field for optical see-through head-mounted displays. *Transactions on Visualization and Computer Graphics*, 24(11):2936–2946, 2018.

[94] I. Radu and B. MacIntyre. Using children's developmental psychology to guide augmented-reality design and usability. In *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 227–236, 2012.

[95] I. Radu and B. Schneider. What can we learn from augmented reality (AR)? In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, 2019. Paper 544.

[96] F. Rameau, H. Ha, K. Joo, J. Choi, K. Park, and I. S. Kweon. A real-time augmented reality system to see-through cars. *Transactions on Visualization and Computer Graphics*, 22(11):2395–2404, 2016.

[97] H. Regenbrecht, K. Meng, A. Reepen, S. Beck, and T. Langlotz. Mixed voxel reality: Presence and embodiment in low fidelity, visually coherent, mixed reality environments. In *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 90–99, 2017.

[98] H. Regenbrecht, T. Schubert, C. Botella, and R. Baños. Mixed reality experience questionnaire (MREQ)-reference. Technical report, University of Otago, 2017.

[99] C. Reichherzer, A. Cunningham, J. Walsh, M. Kohler, M. Billinghurst, and B. H. Thomas. Narrative and spatial memory for jury viewings in a reconstructed virtual environment. *Transactions on Visualization and Computer Graphics*, 24(11):2917–2926, 2018.

[100] D. Schmalstieg and T. Höllerer. *Augmented Reality: Principles and Practice*. Addison-Wesley Professional, Boston, 2016.

[101] T. Schubert, F. Friedmann, and H. Regenbrecht. The experience of presence: Factor analytic insights. *Presence: Teleoperators and Virtual Environments*, 10:266–281, June 2001.

[102] B. Schwerdtfeger, R. Reif, W. A. Gunthner, G. Klinker, D. Hamacher, L. Schega, I. Bockelmann, F. Doil, and J. Tumler. Pick-by-vision: A first stress test. In *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 115–124, 2009.

[103] J. G. Sharp, R. Bowker, and J. Byrne. VAK or VAK-uous? Towards the trivialisation of learning and the death of scholarship. *Research Papers in Education*, 23(3):293–314, 2008.

[104] X. Shi, J. Pan, Z. Hu, J. Lin, S. Guo, M. Liao, Y. Pan, and L. Liu. Accurate and fast classification of foot gestures for virtual locomotion. In *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 178–189, 2019.

[105] M. Slater, M. Usoh, and A. Steed. Depth of presence in virtual environments. *Presence: Teleoperators and Virtual Environments*, 3(2):130–144, 1994.

[106] W. Steptoe, S. Julier, and A. Steed. Presence and discernability in conventional and non-photorealistic immersive augmented reality. In *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 213–218, 2014.

[107] J. E. Swan and J. L. Gabbard. Survey of user-based experimentation in augmented reality. In *Proceedings of International Conference on Virtual Reality*, vol. 22, pp. 1–9, 2005.

[108] T. Teo, L. Lawrence, G. A. Lee, M. Billinghurst, and M. Adcock. Mixed reality remote collaboration combining 360 video and 3D reconstruction. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, 2019. Paper 201.

[109] S. Thompson, A. Chalmers, and T. Rhee. Real-time mixed reality rendering for underwater 360 videos. In *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 74–82, 2019.

[110] T. Tominaga, T. Hayashi, J. Okamoto, and A. Takahashi. Performance comparisons of subjective quality assessment methods for mobile video. In *Proceedings of the International Workshop on Quality of Multimedia Experience*, pp. 82–87, 2010.

[111] A. Van Boxtel and M. Jessurun. Amplitude and bilateral coherency of facial and jaw-elevator EMG activity as an index of effort during a two-choice serial reaction task. *Psychophysiology*, 30(6):589–604, 1993.

[112] B. Volmer, J. Baumeister, S. Von Itzstein, I. Bornkessel-Schlesewsky, M. Schlesewsky, M. Billinghurst, and B. H. Thomas. A comparison of predictive spatial augmented reality cues for procedural tasks. *Transactions on Visualization and Computer Graphics*, 24(11):2846–2856, 2018.

[113] P. Vorderer, W. Wirth, F. Gouveia, F. Biocca, T. Saari, L. Jäncke, S. Böcking, H. Schramm, A. Gysbers, T. Hartmann, C. Klimmt, J. Laarni, N. Ravaja, A. Sacau, T. Baumgartner, and P. Jöncke. MEC spatial presence questionnaire (MEC-SPQ): Short documentation and instructions for application. *Report to the European Community, Project Presence: MEC (IST-2001-37661)*, 2004.

[114] D. Watson, L. Anna Clark, and A. Tellegen. Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, 54:1063–1070, June 1988.

[115] C. A. Wiesner, M. Ruf, D. Sirim, and G. Klinker. 3D-FRC: Depiction of the future road course in the head-up-display. In *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 136–143, 2017.

[116] S. Willi and A. Grundhfer. Spatio-temporal point path analysis and optimization of a galvanoscopic scanning laser projector. *Transactions on Visualization and Computer Graphics*, 22(11):2377–2384, 2016.

[117] Y. Wu, L. Chan, and W. Lin. Tangible and visible 3D object reconstruction in augmented reality. In *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 26–36, 2019.

[118] W. Xu, H.-N. Liang, Y. Zhao, D. Yu, and D. Monteiro. DMove: Directional motion-based interaction for augmented reality head-mounted displays. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, 2019. Paper 444.

[119] Y. Yeshurun, M. Carrasco, and L. T. Maloney. Bias and sensitivity in two-interval forced choice procedures: Tests of the difference model. *Vision Research*, 48(17):1837–1851, 2008.
[120] D. Yu, K. Fan, H. Zhang, D. Monteiro, W. Xu, and H. Liang. PizzaText: Text entry for virtual reality systems using dual thumbsticks. *Transactions on Visualization and Computer Graphics*, 24(11):2927–2935, 2018.
[121] J. Zhang, A. Ogan, T. Liu, Y. Sung, and K. Chang. The influence of using augmented reality on textbook support for learners of different learning styles. In *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 107–114, 2016.
[122] F. Zhou, H. B.-L. Duh, and M. Billinghurst. Trends in augmented reality tracking, interaction and display: A review of ten years of ISMAR. In *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 193–202, 2008.
[123] F. Zijlstra. *Efficiency in Work Behavior: A Design Approach for Modern Tools*. PhD thesis, Technical University of Delft, 1993.