A Systematic Review of Usability Studies in Augmented Reality between 2005 and 2014

Arindam Dey*
Empathic Computing Lab
University of South Australia

Mark Billinghurst†
Empathic Computing Lab
University of South Australia

Robert W. Lindeman‡
HIT Lab NZ
University of Canterbury

J. Edward Swan II§
Mississippi State University

ABSTRACT

Augmented Reality (AR) interfaces have been studied extensively over the last few decades, with a growing number of user-based experiments. In this paper, we systematically review most AR papers published between 2005 and 2014 that include user studies. A total of 291 papers have been reviewed and classified based on their application areas. The primary contribution of the review is to present the broad landscape of user-based AR research, and to provide a high-level view of how that landscape has changed. We also identify areas where there have been few user studies, and opportunities for future research. This poster describes the methodology of the review and the classifications of AR research that have emerged.

1 INTRODUCTION

Augmented Reality (AR) research and development has come a long way in the last few decades. Adoption of AR technology is growing rapidly as more advanced and portable hardware has become available, and registration accuracy, graphics quality, and device size have been largely addressed to a satisfactory level. However, to be widely accepted by end users, AR usability and user experience issues need to be improved.

To help the AR community, reviews of AR usability research have been conducted in the past. In 2005, Swan and Gabbard conducted a survey of four important publication venues, and 1,104 AR papers published between 1992 and 2004 were considered, but they found only 21 papers that reported formal user studies [3]. In 2007, 165 AR related publications reporting user studies were reviewed by Dünser et al. [2] and classified into different types. In a relatively recent literature survey in 2012, Bai and Blackwell reviewed 71 AR papers reporting user studies, but only papers published in the ISMAR Conference Proceedings between 2001 and 2010 [1].

In the last few years there has been a rapid increase in the use of handheld AR devices, and more advanced hardware and sensors have become available. These new wearable and mobile devices, along with the advancement of other technical aspects of AR, have created new research directions. Hence, there is a need for categorization of current AR user research, in order to capture the current state-of-the-art. Additionally, the community also needs a new methodology to perform a practical and reliable review of user research in AR as the number of them has increased significantly in the last few years. For the first time, we have considered impact of a paper before including them to the final review.

To capture the latest trends in usability research in AR, we have conducted a thorough, systematic literature review of AR papers published between 2005 and 2014 that contain a user study, classifying the papers based on their application areas, methodologies used, and type of displays. Our aims are to (1) identify the primary application areas for user research in AR, (2) describe the methodologies and environments that are commonly used, and (3) propose future research opportunities and guidelines for making AR more user friendly.

*e-mail:arindam.dey@unisa.edu.au
†e-mail:mark.billinghurst@unisa.edu.au
‡e-mail:gogo@hitlabnz.org
§e-mail:swan@acm.org

2 METHODOLOGY

One of our goals is to make this review as inclusive as practically possible. As such we decided to consider all papers published in conferences and journals between 2005 and 2014 which include the term Augmented Reality and involve user studies. Papers were found through Scopus and Google Scholar using the same keywords used by Dünser et al. [2]. A total of 1,147 unique papers were found to meet this criteria. We then identified, by looking at each one, whether or not each paper actually reported on AR research by reading each one, and excluding the papers not related to AR, reducing the number to 1,063. We then identified whether or not each paper actually reported on a user study, and removed the papers that did not, bringing our pool to 604 papers. We then looked at these 604 papers and excluded papers that failed to provide any of the following information: (i) participant demographics (number, age, and gender), (ii) design of the user study, and (iii) the experimental task. Only 396 papers satisfied all three of these criteria. Finally, unlike previous work, we considered how much impact each paper had, by measuring its Average Citation Count (ACC) using the following formula:

ACC = Total lifetime citation / lifetime (in years)

For example, if a paper was published in 2010 (a 5 year lifetime until 2014) and had a total of 10 citations in Google Scholar in April 2015, its ACC was 10/5 = 2.0. Based on this formula we included all papers that had an ACC of at least 1.5, showing that they had at least a moderate impact in the field. This resulted in a final set of 291 papers that we reviewed in detail.

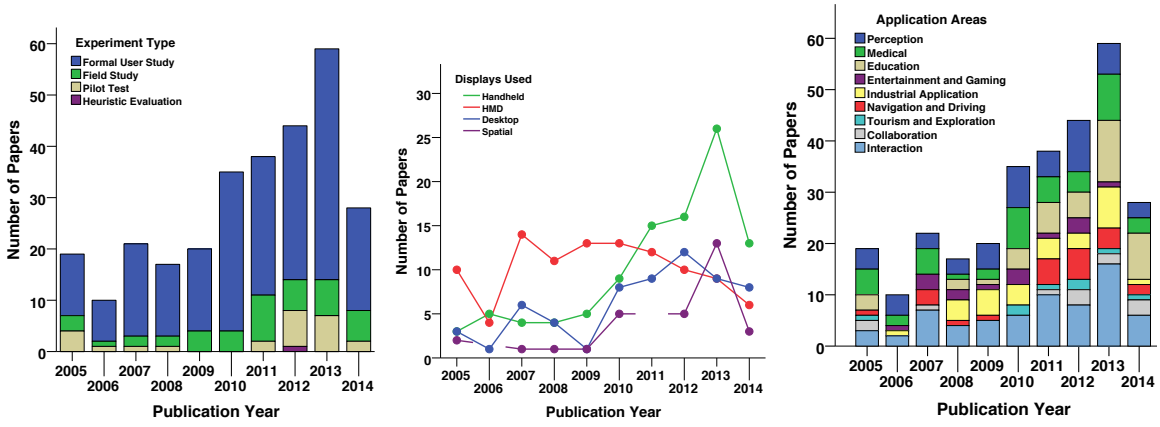
To review these 291 papers, we focused on the following attributes: (i) Application areas and keywords, (ii) Experimental design (e.g., within-subjects, between-subjects, or mixed-factorial) and data collected (e.g., qualitative or quantitative), (iii) Participant demographics (age, gender, number, etc.), (iv) Experimental tasks and environments, (v) Types of experiment (e.g., pilot, formal, field, heuristic, or case study), (vi) Types of senses augmented (e.g., visual, haptic, olfactory, etc.) and displays used

We divided the papers to be reviewed between all of the co-authors. However, before beginning the individual reviews, we performed a norming process where we randomly selected five papers and all of us reviewed those five papers. We then discussed the five papers as a group and reached a consensus about how we were going to review the rest of the papers. We regularly discussed the papers to maintain consistency in the collective reviews.[1]

3 MAJOR FINDINGS

In this poster, we provide a very high-level analysis of the data, while a more detailed analysis is still being completed. Overall, in the 291 (113 journal and 178 conference) papers, there were 353 studies reported. We see an increasing number of papers that reported user studies since 2008 with three times the number of in 2013 than in 2008. The drop noticed in 2014 is due to the selection criteria of papers having at least 1.5 average citations per year, as these were too recent to be cited often.

[1]Full list of 291 papers is available here.



(a) Experiment Type (b) Displays Used (c) Application Areas

Figure 1: Out of the 291 reviewed papers, most of the experiments were in controlled laboratory environments (a). Interestingly, since 2011 more papers used handheld displays than HMDs (b). We categorized the papers into 9 application areas, most in Perception and Interaction (c).

Most of the papers (213, 73%) used a within-subjects design, 43 papers (15%) used a between-subjects design, and 12 papers (4%) used a mixed-factorial design. There were 23 papers (8%) which used different study designs than the ones mentioned above.

Only 55 papers (19%) reported conducting at least one pilot study in their experimentation process and 25 of them reported the pilot studies with adequate details, which shows that the importance of pilot studies is not well recognized. The majority of the papers (221, 76%) conducted the experiments in controlled laboratory environments, while 44 papers (15%) conducted the experiments in a natural environment or field studies (Figure 1(a)). However there were almost no heuristic studies.

In terms of data collection, a total of 139 papers (48%) collected both quantitative and qualitative data, 78 (27%) papers only qualitative, and 74 (25%) only quantitative. For the experimental task, we found that the most popular task involved performance (178, 61%), followed by filling out questionnaires (146, 50%), perceptual (53, 18%), interviews (41, 14%) and collaborative tasks (21, 7%). In terms of dependent measures subjective ratings were the most popular with 167 papers (57%), followed by error/accuracy measures (130, 45%), and task completion time (123, 42%). Many experiments used more than one experimental task or dependent measure, so the percentages are more than 100%. Finally, the bulk of the user studies were conducted in an indoor environment (246, 83%), not outdoor (43, 15%), or a combination of both (6, 2%).

As expected, an overwhelming majority of papers (281, or 96%) augmented the visual sense. Haptic and Auditory senses were augmented in 27 (9%) and 21 (7%) papers respectively. Only five papers (2%) reported augmenting only the haptic sense and six papers reported augmenting only the auditory sense.

The demographics of the participants showed that most of the studies were run with young participants. A total of 182 papers (62%) used participants with an approximate mean age of less than 30 years. A total of 227 papers (78%) reported involving female participants in their experiments, but the ratio of female participants to male participants was low (43% of total participants in those 227 papers). However, when all 291 papers are considered only 36% of participants were females. Several papers (117, 40%) did not explicitly mention the source of participant recruitment. From those that did, a large majority of the papers (102, 35%) sourced their participants from Universities, whereas only 36 papers (12%) mentioned sourcing participants from the general public.

We recorded the displays used in these experiments. Most of the papers used either head-mounted displays (HMDs, 102 papers, or 34.9%) or handhelds (100 papers, or 34.2%), including six papers that used both. Between 2010 and 2014 (204 papers in our review), 50 papers used HMDs and 79 used handhelds, including one paper that used both. However, since 2009, the number of papers using HMDs started to decrease (Figure 1(b)); and since 2011 papers using handheld displays consistently outnumbered papers using HMDs. This shows that handheld mobile AR is beginning to dominate overall AR research efforts, at least in terms of publications with user studies.

We categorized the papers into nine different application areas: (i) Perception (51 papers), (ii) Medical (44), (iii) Education (42), (iv) Entertainment and Gaming (14), (v) Industrial (30), (vi) Navigation and Driving (23), (vii) Tourism and exploration (8), (viii) Collaboration (12), and (ix) Interaction (67). The Perception and Interaction categories are rather general areas of AR research, and contain work that reports on more low-level experiments, possibly across multiple application areas (Figure 1(c)). Our analysis shows that there are a low number of AR user studies published in Collaboration, Tourism, and Entertainment, identifying future application areas for user studies. There is a noticeable increase in user studies in educational applications.

4 Conclusion

In this paper we report on ten years of user studies published in AR papers. This initial exploration shows that there has been an increase in the number of usability studies performed in AR research over the last decade and a shift towards more studies on handheld displays. Most of these studies are formal user studies, with little field testing and almost no heuristic evaluations. Over the years there is an increase in AR user studies from Educational applications, but there are few collaborative user studies or use of pilot studies. The most popular experimental task involve filling out questionnaires, which lead to subjective ratings being the most widely used dependent measure. This suggests opportunities for increased research in collaboration, and use of field studies and a wider range of evaluation methods. We also notice that improvements in the choice of subjects might be needed, since the sample populations used are dominated by mostly young, educated, and male participants, so it will be good incorporate more diversity. We are currently performing a more-detailed analysis of the data, and hope to identify the limitations and challenges of user-based experiments in AR, while providing some guidelines for conducting future user studies in AR.

References

[1] Z. Bai and A. F. Blackwell. Analytic review of usability evaluation in ISMAR. *Interacting with Computers*, 24(6):450–460, Nov. 2012.

[2] A. Dünser, R. Grasset, and M. Billinghurst. A survey of evaluation techniques used in augmented reality studies. Technical report, 2008.

[3] J. E. Swan II and J. L. Gabbard. Survey of user-based experimentation in augmented reality. In *Proceedings of 1st International Conference on Virtual Reality, HCI International 2005*, pages 1–9, July 2005.