A Survey of Evaluation Techniques Used in Augmented Reality Studies

Andreas Dünser, Raphaël Grasset, Mark Billinghurst

ABSTRACT

In this paper we report on an initial survey of user evaluation techniques used in Augmented Reality (AR) research. To identify all papers which include AR evaluations we reviewed research publications between the years 1993 and 2007 from online databases of selected scientific publishers. Starting with a total of 6071 publications we filtered the articles in several steps which resulted in 165 AR related publications with user evaluations. These publications were classified in two different ways: according to the evaluation type used following an earlier literature survey classification scheme; and according to the evaluation methods or approach used. We present the results of our literature survey, provide a comprehensive list of references of the selected publications, and discuss some possible research opportunities for future work.

Keywords: Augmented Reality, user evaluation, evaluation methods.
1 Introduction

Although Augmented Reality (AR) has been studied for over forty years it has only been recently that researchers have begun to formally evaluate AR applications. Most of the published AR research has been on enabling technologies (tracking or displays, etc), or on experimental prototype applications, but there has been little user evaluation of AR interfaces [43]. In 2004 Swann et al. [141] produced a literature survey reviewing a total of 1104 articles from the leading journals and conferences. Of the 266 AR-related publications identified, only 38 (~14%) addressed some aspect of human computer interaction, and only 21 (~8%) described a formal user evaluation. They conclude that user-based studies have been under utilized in AR research.

One reason for the lack of user evaluations in AR could be a lack of education on how to evaluate AR experiences, how to properly design experiments, choose the appropriate methods, apply empirical methods, and analyse the results. There also seems to be a lack of understanding of the need of doing studies or sometimes the incorrect motivation for doing them. If user evaluations are conducted out of incorrect motivation or if empirical methods are not properly applied, the reported results and findings are of limited value or can even be misleading.

It is worthwhile to collect knowledge on user evaluations gathered in other disciples and to bring it into AR settings. For example studying peoples’ behaviour with various methods is very common in general Human Computer Interaction (HCI) or Psychology. Various tools have been developed and tested that can be applied in AR research. Although the specifics of AR interfaces are different to more traditional interfaces, the basic tools to evaluate user behaviour or perception are quite similar.

This report aims to provide a resource that can be used by the AR research community to design user evaluations. We first describe the role of user evaluation studies, and then review previous usability surveys of AR interfaces.

2 Related work

Gabbard and Swan [52] argue that user-based experiments are critical for driving design activities, usability, and discovery early in an emerging technology’s development, such as in the case of Augmented Reality. They point out that lessons learned from user studies provide value to the field as a whole in terms of insight into the user interface design space.

There have been other previous related literature survey’s covering Virtual Reality (VR) and AR user evaluation. For example, Gabbard [51] presented a research summary with the goal of identifying AR design and evaluation guidelines which may be specifically applied to augmented reality systems. His approach was to collect and
synthesize information from many different sources, and create a list comprised of a structured collection of otherwise piecemeal findings. These sources included:

- Virtual Environment related journals and conferences (e.g., Presence, IEEE VR (formally VRAIS) conference proceedings, Human Factors in Computing Systems (CHI) conference proceedings, and SIGGRAPH conference proceedings)
- Human-computer interaction related literature
- Experiences using AR systems
- Comments made by users of specific observed AR systems
- World Wide Web searches for AR-related work.

Anastassova et al. [3] state that literature reviews in the areas of VR and Mixed Reality (MR) reveal that current research focuses on building ad-hoc systems and on evaluating them in artificial or informal settings. User needs analysis is seldom carried out and if included it is done by very few “task experts”, by quick field studies of future users’ activity, or by questionnaires. They conclude that the evaluation of high-fidelity prototypes and iterative prototyping are efficient ways to uncover problematic design issues and analyse user needs.

Bach and Scapin [7] studied issues with evaluating Mixed Reality systems. They found that there are no usability evaluation methods specifically designed for such systems. They also discuss if and how evaluation methods used in other domains can be adopted to evaluate MR systems. They identify three categories of methods that are general enough in their approach to be suitable for these systems:

1. questionnaires and interviews
2. inspection methods, and
3. user testing.

Questionnaires and interviews are useful for gathering subjective data, user preferences, missing functionalities and to compare against performance data. Inspection methods can be limited in that there is still limited knowledge about specific ergonomic issues and design guidelines for mixed reality systems. User testing has been the main method in other disciplines. However, according to Bach and Scapin, various methodological problems need to be tackled in order to apply user testing to MR systems. These challenges are similar to those of other emerging technologies and are related to the limited knowledge about different aspects of these new systems (e.g. task experts are needed to operate the systems, learning by trial and error is common, technical challenges, etc.).

3 Methodology

Our literature survey method is characterized by iterative selection, filtering and classification processes. We started by defining and selecting appropriate sources for our literature survey (section 3.1) and filtering the initial collection of articles to meet our objectives (section 3.2). We removed articles that were incorrectly selected in the search process (false positives) and identified those articles that included user
evaluations. Finally we classified the AR user evaluation articles according to two different classification schemes.

3.1 Initial database and publisher search

For the data collection the ISI Web of Science is frequently used as a basis for literature studies. However, it is of limited use for AR related research and computer science research in general (see [108]). Therefore we chose to collect articles from publisher databases with a focus on the major publishers of computer science literature (see Figure 1 for a list of databases). We used the search engine of these publishers’ online databases to search for publications between the years 1992 and 2007 containing the term “Augmented Reality”. This resulted in a collection of 6071 publications (see Figure 1).

![Figure 1 Result of the initial “Augmented Reality” keyword search from various publisher databases](image)

We could identify 3 main publishers in our area: ACM Digital Library, IEEE Xplore and SpringerLink. Because over the years most AR research related conferences and journals were published through ACM or IEEE, we restricted our present literature review to these two publishers: ACM and IEEE with a total of 3309 articles.

3.2 Keyword selection

To select articles related to AR user evaluation we refined our database in a second step. For this we applied a search directly to the collected publication database (in pdf format), resulting in a list of 1203 articles. The search queries we used are listed in Table 1 (in the remainder of the document we will refer to this step as keyword
If a paper satisfied at least one of these selection criteria they were included in the list of documents to be searched further.

**Table 1** Keyword selection search queries

<table>
<thead>
<tr>
<th>Selection queries</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;augmented reality&quot; AND &quot;user evaluation(s)&quot;</td>
</tr>
<tr>
<td>&quot;augmented reality&quot; AND &quot;user study/-ies&quot;</td>
</tr>
<tr>
<td>&quot;augmented reality&quot; AND &quot;feedback&quot;</td>
</tr>
<tr>
<td>&quot;augmented reality&quot; AND &quot;experiment(s)&quot;</td>
</tr>
<tr>
<td>&quot;augmented reality&quot; AND &quot;pilot study&quot;</td>
</tr>
<tr>
<td>&quot;augmented reality&quot; AND participant AND study</td>
</tr>
<tr>
<td>&quot;augmented reality&quot; AND participant AND experiment</td>
</tr>
<tr>
<td>&quot;augmented reality&quot; AND subject AND study</td>
</tr>
<tr>
<td>&quot;augmented reality&quot; AND subject AND experiment</td>
</tr>
</tbody>
</table>

In a further step, we removed the articles that, although containing the term “Augmented Reality”, are not actual AR research papers. Such “false positives” may just mention AR in the related work, discussion, or future work section, or contain some AR literature references.

We chose articles which comply with Azuma’s [6] definition of Augmented Reality and focus on visual augmentation. Hence we did not include other kinds of augmentation such as audio or haptic in this survey. However, we included publications discussing AR specific enabling technologies such as see through head mounted displays (HMDs). After this filtering step we had a total of 557 AR related publications (database I).

After this we selected those articles that included a user evaluation, which resulted in a total of 161 publications (database II). This selection was rather broad including articles with very basic user evaluations or informal user observations.

### 3.3 Classification of AR user-evaluation publications

We classified all AR user-based-evaluation publications according to:

1. evaluation area / evaluation type (section 4.1)
2. evaluation methods / evaluation approach used (section 4.2)

The first classification scheme was based on Swan and Gabbard’s scheme used in [141]. The second classification scheme is based on the general research methods or evaluation approach.

### 4 Survey Results

Figure 2 shows the distribution of the total number of ACM and IEEE AR papers after the keyword selection (database I), and the number of AR papers with user evaluations (database II), between 1992 and 2007.

---

1 In this literature survey we do not include any judgement of the scientific and methodological quality of the applied research strategy, or evaluation design and method.
Table 2 summarizes the results of the iterative article selection. We determined the final percentage of AR user evaluation publications in two ways. First we calculated the actual percentage of publications containing user evaluations from the number of papers selected as AR research papers after our keyword selection (columns ‘a’)’. Second we calculated the percentage of papers containing AR user evaluations from the estimated number of AR research publications in ACM + IEEE (columns ‘b’). After the keyword selection we had 53.4% of “false positives” and 46.6% of actual AR research publications. We estimate a similar percentage of AR research papers (46.6%) in the total number of ACM and IEEE publications.

Our survey shows that an estimated 10% of the AR papers published in ACM and IEEE included some user evaluation. Excluding informal user evaluations the percentage is close to 8, the amount reported by Swann et al. [141]. Evaluations were classified as informal if they just include informal user observations (e.g. at public demonstrations) or informal collection of user feedback. Formal evaluations were those that followed a rigorous user evaluation program.
Table 2 Breakdown of the results of our AR user evaluation reviewing process:

a) the actual percentage of AR user evaluations in AR publications after keyword selection
b) the percentage of AR user evaluations in an estimated number of AR publications from the total number of ACM + IEEE publications

<table>
<thead>
<tr>
<th>Year</th>
<th>Total AC M and IEEE</th>
<th>Total after keyword selection</th>
<th>AR papers after KW selection</th>
<th>Total AR user evaluations</th>
<th>Formal AR user evaluations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>a)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>b)</td>
</tr>
<tr>
<td>1992</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1993</td>
<td>20</td>
<td>5</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1994</td>
<td>12</td>
<td>4</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1995</td>
<td>14</td>
<td>8</td>
<td>5</td>
<td>1</td>
<td>20.0%</td>
</tr>
<tr>
<td>1996</td>
<td>24</td>
<td>11</td>
<td>4</td>
<td>1</td>
<td>25.0%</td>
</tr>
<tr>
<td>1997</td>
<td>47</td>
<td>12</td>
<td>4</td>
<td>1</td>
<td>25.0%</td>
</tr>
<tr>
<td>1998</td>
<td>71</td>
<td>29</td>
<td>13</td>
<td>2</td>
<td>15.4%</td>
</tr>
<tr>
<td>1999</td>
<td>143</td>
<td>49</td>
<td>23</td>
<td>5</td>
<td>21.7%</td>
</tr>
<tr>
<td>2000</td>
<td>163</td>
<td>64</td>
<td>36</td>
<td>10</td>
<td>27.8%</td>
</tr>
<tr>
<td>2001</td>
<td>217</td>
<td>75</td>
<td>25</td>
<td>6</td>
<td>24.0%</td>
</tr>
<tr>
<td>2002</td>
<td>357</td>
<td>92</td>
<td>42</td>
<td>16</td>
<td>38.1%</td>
</tr>
<tr>
<td>2003</td>
<td>403</td>
<td>113</td>
<td>52</td>
<td>18</td>
<td>34.6%</td>
</tr>
<tr>
<td>2004</td>
<td>502</td>
<td>179</td>
<td>90</td>
<td>21</td>
<td>23.3%</td>
</tr>
<tr>
<td>2005</td>
<td>497</td>
<td>209</td>
<td>96</td>
<td>25</td>
<td>26.0%</td>
</tr>
<tr>
<td>2006</td>
<td>472</td>
<td>206</td>
<td>89</td>
<td>27</td>
<td>30.3%</td>
</tr>
<tr>
<td>2007</td>
<td>366</td>
<td>147</td>
<td>71</td>
<td>28</td>
<td>39.4%</td>
</tr>
<tr>
<td>3309</td>
<td>1203</td>
<td>557</td>
<td>161</td>
<td>28.9%</td>
<td>10.4%</td>
</tr>
<tr>
<td>120</td>
<td>21.5%</td>
<td>7.8%</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

4.1 Classification of papers by user evaluation type

In their work, Swan and Gabbard [141] report that the first user-based experimentation in AR was published in 1995, and since then usability studies have been reported in three related areas:

1. **Perception**: experiments that study low-level tasks, with the goal of understanding how human perception and cognition operate in AR contexts,
2. **Performance**: experiments that examine user task performance within specific AR applications or application domains, in order to gain an understanding of how AR technology could impact underlying tasks
3. **Collaboration**: experiments that examine generic user interaction and communication between multiple collaborating users.

In this report we provide an update to their work by considering both more recent publications and older papers. During our classification we found that we had to extend the scheme to add one more category for papers that did not fit into the existing categories. The additional category is “Interface or system usability studies”. Although papers from this category can be quite similar to articles in the second category (Performance), they don’t necessarily involve measurement of user task performances but other ways of identifying issues with system usability. Figure 3 shows the final classification of AR user evaluation papers into these four categories.
Perception and cognition studies
35 publications were classified in this category: [5, 16-18, 40, 50, 53, 54, 70, 71, 83, 84, 88, 94, 97, 99, 105, 106, 127, 129, 131, 132, 139, 142, 143, 146, 147, 153, 157, 161-164, 166, 169]

User task performance studies
75 publications were classified in this category; 64 formal user evaluations: [1, 4, 9, 12, 24, 26, 29-33, 38, 39, 41, 44, 45, 48, 49, 58, 62, 63, 65, 66, 73-77, 81, 87, 89-93, 96, 98, 101-103, 107, 109, 113, 116, 117, 121, 122, 125, 126, 128, 133, 137, 144, 145, 150-152, 154-156, 158, 167, 168] and 11 informal user evaluations: [57, 61, 67, 72, 86, 100, 114, 134-136, 165]

Collaboration studies
10 publications were classified in this category; 8 formal user evaluations: [10, 22, 28, 60, 64, 82, 159, 160] and 2 informal user evaluations: [20, 21]

System usability studies
41 publications were classified in this category; 13 formal user evaluations: [15, 25, 34, 46, 47, 59, 68, 80, 95, 119, 130, 140, 148] and 28 informal user evaluations: [2, 8, 11, 13, 14, 19, 23, 27, 35-37, 42, 55, 56, 69, 78, 79, 85, 104, 110-112, 115, 118, 120, 123, 138, 149]

4.2 Classification of user study approaches and methods
This classification scheme was inspired by the authors’ experience with past AR related user evaluation publications. The goal was to get an overview over the...
research approaches and methods that have been applied in the field so far. Therefore this scheme does not necessarily follow other methodology classifications.

We classified AR user evaluation papers into five types:
1. Objective measurements
2. Subjective measurements
3. Qualitative analysis
4. Usability evaluation techniques
5. Informal evaluations

Some authors used a large variety of methods, but to avoid double classification the publications were classified just in one category according to the main evaluation approach or method used. Figure 4 shows the classification of the AR user evaluation papers into these types.

![Classification of publications by evaluation method / approach](image)

**Figure 4** Classification of publications by evaluation method / approach

1. **Objective measurements**

These are studies that include objective measurements. The most common measurements are task completion times and accuracy / error rates; other examples are scores, position, movement, number of actions, etc. In general these studies employ a statistical analysis of the measured variables, however, some only include a descriptive analysis of the results.

75 publications were classified in this category: [1, 5, 9, 12, 17, 24, 26, 29, 39, 40, 45, 46, 48, 49, 53, 54, 58, 60, 62-66, 70, 71, 73-75, 77, 81-84, 88-93, 97, 99, 101-103, 105-107, 113, 116, 117, 121, 122, 125, 126, 128, 131-133, 137, 143-146, 150-152, 154, 156-158, 161, 163, 166]
2. Subjective measurements
Here we selected publications which studied users using questionnaires, subjective user ratings, or judgements. With respect to analysis some of these studies also employ statistical analysis of the results, others only include a descriptive analysis. 29 publications were classified in this category: [4, 15, 16, 18, 31-33, 50, 59, 76, 80, 94-96, 109, 119, 124, 127, 130, 139, 142, 147, 153, 159, 162, 164, 167-169]

3. Qualitative analysis
This category comprises studies with formal user observations, formal interviews, or classification or coding of user behaviour (e.g. speech or gesture coding).

9 publications were classified in this category: [10, 28, 30, 38, 41, 44, 87, 140, 160]

4. Usability evaluation techniques
These are publications that employ evaluation techniques that are often used in interface usability evaluations such as heuristic evaluation, expert based evaluation, task analysis, think aloud method, or Wizard of OZ method.

7 publications were classified in this category: [25, 34, 47, 68, 98, 148, 155]

5. Informal evaluations
In this category we included informal user evaluations such as informal user observations or informal collection of user feedback.

41 publications were classified in this category: [2, 8, 11, 13, 19-21, 23, 27, 35-37, 42, 55-57, 61, 67, 69, 72, 78, 79, 85, 86, 100, 104, 110-112, 114, 115, 118, 120, 123, 134-136, 138, 149, 165]

5 Discussion
In our survey we have followed a different method to Swan and Gabbard [2]. Although both approaches are not directly comparable in terms of absolute numbers or percentage of AR evaluation publications, we found some similar results. Swan and Gabbard found 21 user based experiments in a total of 266 AR related publications, which represents around 8%. Our paper selection was somewhat broader, e.g. including other peer reviewed publications such as posters. Despite this we also found 8% of formal user evaluations in the estimated number of ACM and IEEE AR research publications. This percentage is 10% if informal user evaluations are included.

Our actual count of user evaluation publications is 161, which is 29% of the keyword selected publications. This percentage cannot be seen as a completely accurate estimate of AR user evaluations. It is somewhat biased by the order in which we applied the different filtering steps on the initial list of publications (i.e. the pre-selection using “user evaluation / study / experiment” keywords). We expect the
remaining bulk of ACM and IEEE publications, that were not included in our survey after the keyword selection, to contain many AR related publications (we estimated the percentage of AR related publications to be about 46%). However, as the omitted articles do not contain common user evaluation related keywords, we do not expect them to include AR user evaluations. In future work we plan to extend our survey to those publications that were not included after this keyword selection to get a better estimate of the amount of AR user evaluations.

Swan and Gabbard identified 12 perception, 6 user task performance, and 3 collaboration publications. In our survey the share of formal user task performance evaluation papers (46.6 %) was higher than the perception / cognition evaluation publications (21.7%). To some extent this can be explained by the observation that the number of user task performance evaluation publications increased significantly from 2005 onwards. Furthermore we considered different publication sources for our literature survey. Similar to their survey, we also found that studies that evaluate collaboration between users in AR are quite underrepresented. From a total of 161 AR user evaluation publications we found only 10 evaluations of collaborative AR interfaces.

During our work we found that we had to extended Swan and Gabbard’s classification with one additional category – interface / system usability studies. These are user studies with a more general approach to uncover usability issues with the tested augmented reality system or prototype. There were 41 publications in this category; 13 formal user studies and 28 informal studies.

In general not just the number but also the percentage of AR user evaluations seems to be slightly increasing over time. Especially the amount of user performance evaluations rose. Perception and system usability evaluations each seem to have stayed on roughly similar levels in the last few years, although the total number of evaluations increased. With just 8%, the overall ratio of user evaluations in AR research publications still is rather low and the trend of increasing numbers of user evaluations shows that there is more potential for evaluating AR systems. Our survey shows that there is also more potential for broadening the scope of evaluations employed in AR. For example studies evaluating collaboration are hardly represented in AR research.

In our second classification we categorized the papers according to the evaluation method used. Some of these methods are similar to the methods discussed by Bach and Scapin [7], including: questionnaires and interviews, inspection methods, and user testing. Our final categories are: objective measurements, subjective measurements, qualitative analysis, usability evaluation techniques, and informal user evaluations. The main focus here is on using objective measurements. Formal qualitative analysis and the use of more general usability evaluation techniques are not as common. However, the ratio of formal user evaluations compared to informal evaluations increased over the years. Between 1995 and 2001 there is an average of 57% formal evaluations, whereas between 2002 and 2007 this percentage is 76%.
Thus there seems to be a growing understanding for the need to formalize the evaluation process and conduct properly designed user studies.

Finally we should note that our current literature review has several limitations. Some of them are obvious and easy to overcome such as extending the survey to publications not included here. However, others are more difficult to resolve such as limitations to the search process in online databases in general. The later include the use of different terminology (e.g. mixed reality), published articles that are not available in the online databases, foreign language articles, etc. Furthermore the process of iteratively refining the literature selection and classification is a highly repetitive task with a certain likelihood of human error.

6 Conclusions and Future work
In this report we presented a literature survey of user evaluation techniques in selected Augmented Reality research publications between 1993 and 2007. First we aimed at extending previous work by considering more recent publications and extending the classification of types of user evaluations with a more task oriented focus. Second we categorized the publications by looking that the evaluation methods used. A goal of this literature survey was to provide a resource for the AR community. It may be used to get an overview of the use of user evaluation techniques in augmented reality and help to promote the use and further increase the quality of user evaluation in AR research.

Our survey is a first step of a more comprehensive survey of the field. In further work we plan to include the remaining 2106 IEEE and ACM publications that were omitted after our user evaluation keyword selection to give us a better picture of the amount of user evaluations carried out in the field of augmented reality. Furthermore we plan to extend our survey to other publisher databases such as Springer Link (LNCS), MIT press Journals, SPIE Digital Library, and ScienceDirect. In addition to publishers with a strong computer science focus we also plan to extend our study to publishers and databases from other research domains such as art, social science, psychology or medicine.
7 References


