NAIST-IS-DD1261207

Doctoral Dissertation

Usability Evaluation Framework for Handheld Augmented Reality Applied to Learning Support

Marc Ericson C. Santos

September 14, 2015

Department of Information Science
Graduate School of Information Science
Nara Institute of Science and Technology

A Doctoral Dissertation
submitted to Graduate School of Information Science,
Nara Institute of Science and Technology
in partial fulfillment of the requirements for the degree of
Doctor of ENGINEERING

Marc Ericson C. Santos

Thesis Committee:

Professor Hirokazu Kato	(Supervisor)
Professor Naokazu Yokoya	(Co-Supervisor)
Associate Professor Christian Sandor	(Co-Supervisor)
Assistant Professor Goshiro Yamamoto	(Co-Supervisor)
Professor Ma. Mercedes T. Rodrigo	(Ateneo de Manila University)

Usability Evaluation Framework for Handheld Augmented Reality Applied to Learning Support*

Marc Ericson C. Santos

Abstract

Augmented reality (AR) is an emerging technology for various fields of application. Researchers and teachers perceive AR running on handheld devices to be useful in educational settings, despite the few evaluations conducted on handheld AR (HAR) systems for learning support. Conducting usability evaluations of HAR is difficult in most application areas because it relies on the experience of AR experts. In response, I present a usability evaluation framework to guide evaluations of HAR systems in general. In my framework, I defined two usability constructs, namely, manipulability – the ease of handling the device, and comprehensibility – the ease of understanding the presented information. Based on this framework, I developed a valid and reliable usability questionnaire called the HAR Usability Scale (HARUS). I applied HARUS to conduct a summative usability evaluation of a situated vocabulary learning support system called FlipPin and I used my framework to design a series of formative usability evaluations of AR x-ray for learning support.

Keywords:

augmented reality, handheld devices, learning support, usability evaluation

*Doctoral Dissertation, Department of Information Science, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-DD1261207, September 14, 2015.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

Current handheld devices, such as smartphones and tablet computers, have powerful processors, large screens, and built-in location sensors and cameras. These features make handheld devices convenient platforms for augmented reality (AR) – the seamless integration of virtual objects with real environments [10]. Handheld AR (HAR) affords many new ways of interacting with digital content, with applications in several industries, including entertainment, marketing and sales, education and training, navigation, tourism, and social networking. Although some applications have already been adopted by general consumers, interfaces using HAR systems remain limited and researchers are continuously developing more reliable and more intuitive interfaces.

The evaluation of HAR systems is challenging because it relies on the years of experience of AR experts. Researchers who have spent years studying AR are familiar with the common problems with AR interfaces. Moreover, AR experts know the best practices to systematically conduct assessments of these problems. As more and more researchers, designers, and developers conduct interdisciplinary research around AR, the need for evaluation frameworks and standard evaluation tools becomes increasingly important. In particular, there is growing interest in using AR for learning support [11]. Researchers from the fields of

human-computer interaction, augmented reality, usability engineering, instructional design, and education technology, among others, need to work together to create effective learning support systems with AR. In Chapter 2, I discuss my systematic review of the design and evaluation of AR learning support systems.

In response to the need for an evaluation framework to guide HAR usability evaluations, I present my framework for conducting usability evaluations of HAR systems. In my evaluation framework, I define two usability constructs, namely, manipulability – the ease of handling the device, and comprehensibility – the ease of understanding the presented information. I then list the possible issues that designers and developers should investigate in usability evaluations. Usability refers to how well target users can use a systems functionality to accomplish a specific task [135]. For summative usability evaluations, the Handheld Augmented Reality Usability Scale (HARUS) that is designed from my evaluation framework can be used. The HARUS gives a HAR system three scores, namely, the manipulability score, comprehensibility score, and usability score. In Chapter 3, I discuss my evaluation framework, the HARUS, and my experiments which explore the validity and reliability of the HARUS.

I investigate learning support because it is an application area wherein non-AR experts are interested to develop their own HAR applications. Researchers perceive that using AR is beneficial to the learning process, despite the limited studies focusing on AR's design, evaluation, and effects on learning. I applied my evaluation framework and the HARUS questionnaire to evaluate two features of HAR in learning support. The first feature of HAR is presenting information related to the real environment. I developed the application called FlipPin, a HAR system for situated vocabulary learning. Aside from conducting a summative usability evaluation of FlipPin in Chapter 4, I also explored how presenting virtual information on a real environment affects the subjects' memorization and motivation. The second feature of HAR is AR X-ray, which provides a useful visualization for learning the interior of a target object. AR X-ray is a developing technology and requires formative evaluations to adapt it to classroom use. Based on my evaluation framework, I explored the possible usability issues of depth perception and legibility in Chapter 5. I then narrowed my evaluations to near-field legibility, a factor affecting both human cognition and usability.

1.1. Evaluations in AR Prototype Development

Several types of evaluations are necessary for developing novel user interfaces with limited guidelines or standards. Insights generated from evaluations are used as inputs to iteratively improve a novel interface. Figure 1.1 shows the three types of evaluations in the user-centered design and evaluation methodology for virtual environments proposed by Gabbard, Hix, and Swan [60]. The evaluations are expert guidelines-based evaluation, formative user-centered evaluation and summative comparative evaluation.

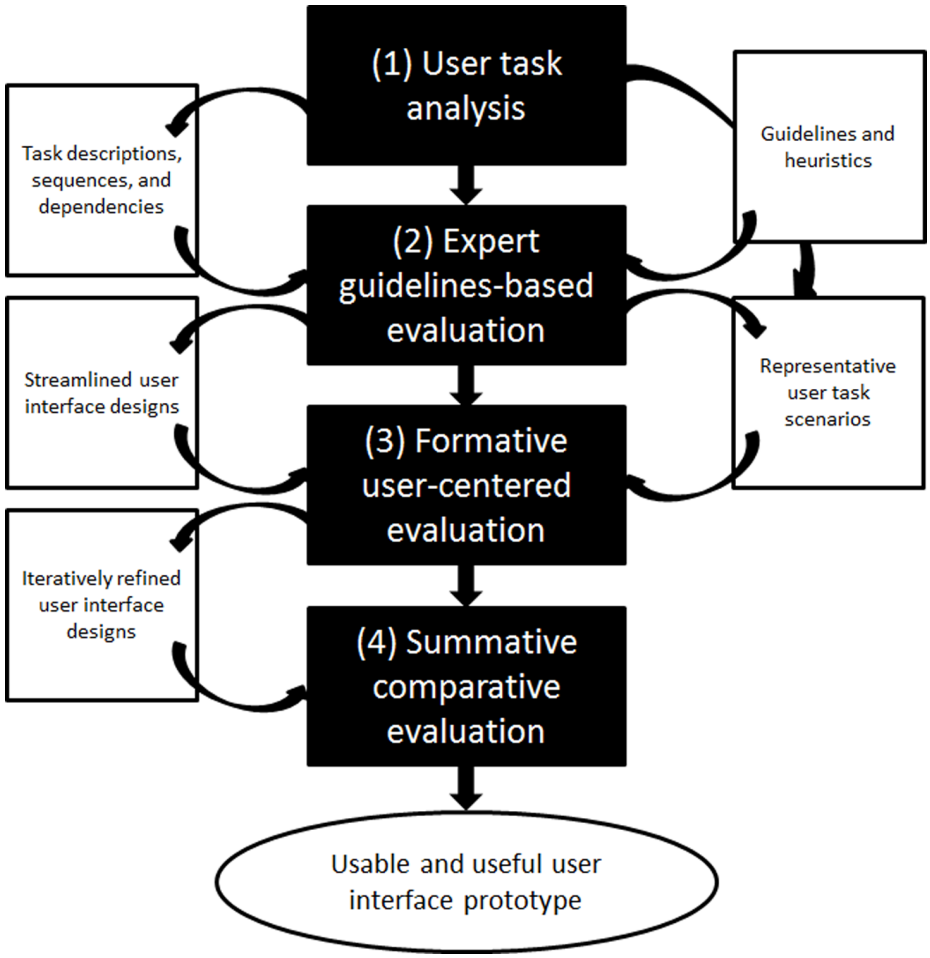

Figure 1.1. User-centered design and evaluation methodology for virtual environment user interaction by Gabbard, Hix, and Swan [60].

Gabbard and Swan [61] note that although this methodology was developed for virtual environments, it still applies for development of AR systems. In the case of HAR systems, conducting the guidelines-based evaluation may be difficult, if not impossible, because there are limited design guidelines for HAR applications. As for conducting the formative and summative evaluations, there are limited studies demonstrating how such evaluations are performed for HAR. More specifically, it would be helpful to summarize known issues of HAR and best practices in conducting evaluations.

1.1.1 Problem and Contribution

The main problem that I am addressing with my dissertation is the lack of evaluation framework and evaluation tools for HAR. In my dissertation, I present a usability evaluation framework for analyzing HAR applications and designing user studies. Moreover, I developed the HARUS, an evaluation tool for measuring the usability, manipulability, and comprehensibility of a HAR application.

In Chapter 3, I discuss the development of my evaluation framework. To accomplish this, I studied the background of HAR applications and I provided my synthesis of previous studies. I then developed the evaluation tool called HARUS based on my evaluation framework. As part of the questionnaire development process, I gathered evidence of validity and reliability of the HARUS in the experiments, as illustrated in Figure 1.2. I also discuss the lessons learned from using the HARUS in these experiments.

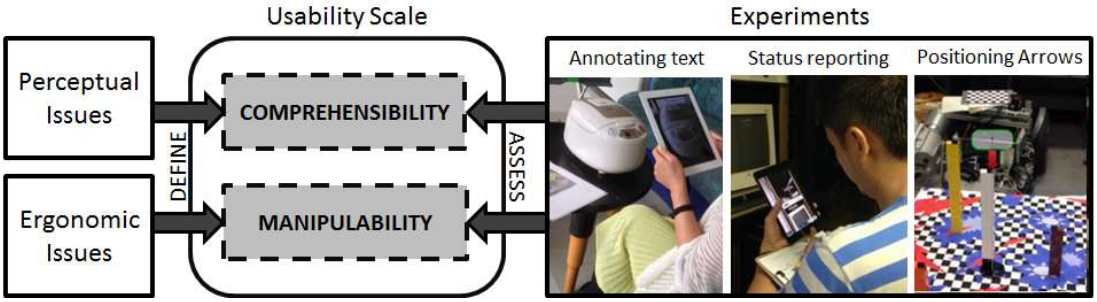

Figure 1.2. The evaluation framework and HARUS are based on common issues of HAR. The validity and reliability of HARUS were assessed in three experiments.

1.2. HAR in Learning Support

Researchers are becoming more interested in using AR for learning support [11]. In AR, computer-generated information is placed in the world as if it co-exist with real objects. It is an emerging technology that is finding applications in education because of its possible benefits to teaching and learning [190]. However, AR's practical uses are relatively not well-understood compared to other technologies [81]. Currently, there are few studies substantiating AR's benefits to learning [72].

For several years now, AR has been considered to be a technology that will play an important role in educational settings. In 2005 [78] and 2006 [79], Johnson et al. forecasted AR as one of the six emerging technologies that will enter mainstream use in educational settings by 2009 to 2011. However, based on the number of publications from 2004 to 2010, AR did not develop as much as other emerging technologies [123].

In 2010, Johnson et al. [84] predicted the adoption of AR around 2012 to 2013 because of the advances in handheld devices, such as smartphones and tablet computers. These handheld devices are already equipped with powerful processors, cameras, and large screens for displaying some virtual data onto the real world scene. Other sensors, such as the GPS sensor working with the gyroscope or compass can identify the phone's location and orientation, thereby displaying relevant content to the user's view.

To summarize research related to AR in education, I conducted my own systematic literature review in Chapter 2. Based on this review, there are limited usability evaluations conducted for AR and HAR. Moreover, the effects of AR on cognition and on learning need to be verified. In my thesis, I focus on the use of HAR for learning support because it is a flexible platform that could be adopted in the near future given the current education practice.

1.2.1 AR as Situated Multimedia

Studies note that AR's strengths and therefore its applicability to education are embodied cognition [193], [90], [92] and interactivity [72], [47] because AR affords new ways of intuitively interacting with information [171]. Aside from embodied

cognition and interactivity, a more fundamental advantage of AR that is not explored as much is the manner of displaying visual information; i.e., AR is a technology for annotating virtual information onto real environments.

Dede [41] explains that AR is useful for supporting ubiquitous learning in authentic environments. Based on the location or other contexts of the user, a system can provide some relevant learning content. In practice, computer-supported ubiquitous learning systems involve the use of handheld devices, such as smartphones [81]. The role of AR in ubiquitous learning is to present the information onto the real environment, thereby creating a stronger connection between the virtual content and the real environment. However, as of the time of this writing, there has been little empirical evidence collected to substantiate or refute AR's potential as a usable carrier of educational content.

In Chapter 4, I developed a HAR system called FlipPin shown in Figure 1.3 for studying vocabularies in a real environment. I then conducted evaluations using the HARUS. Results suggest that using the HARUS may allow users to report more of their difficulties with FlipPin. I also evaluated the learning outcomes and student motivation when using FlipPin. In particular, I tested FlipPin's effectiveness for a word memorization task. Results suggest that using FlipPin may possibly improve retention, attention, and satisfaction.

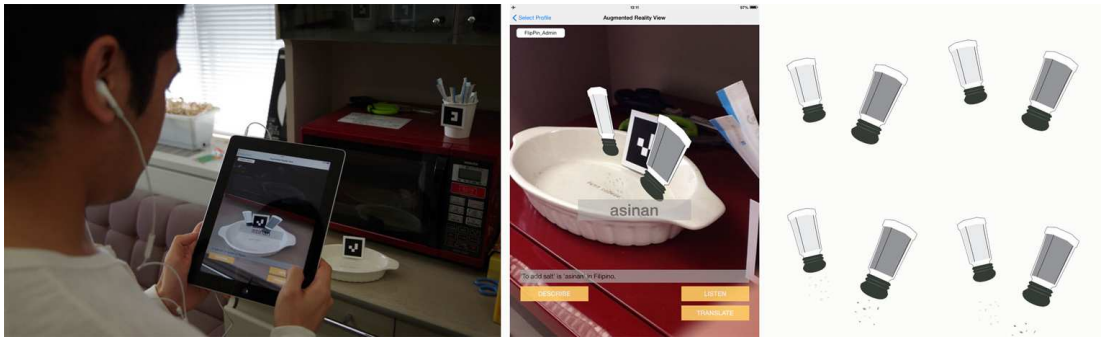

Figure 1.3. A participant studies vocabulary in a refreshment area. FlipPin illustrates an action word using a three-dimensionally registered sprite sheet animation.

1.2.2 AR X-Ray for Educational Settings

Another important affordance of AR is to "visualize the invisible," such as unobservable scientific concepts [190]. An object may be practically invisible due to occlusion, such as internal organs and engines of machines. To view these hidden objects, AR X-ray can be used to provide the illusion of looking inside a target object.

AR X-ray is a novel visualization for education. It is necessary to investigate how it affects the students' perception and the suitability of the state-of-the-art AR x-ray to the teacher's practice. Current implementations of AR X-ray have not yet been tested extensively, particularly in short distances. In response, I conducted evaluations of AR X-ray, as illustrated in Figure 1.4. In Chapter 5, I discuss my user studies on edge-based AR X-ray with students and teachers. I learned that the most important quality of AR X-ray for near-field distances (within arm's reach) is legibility. Legibility is a factor affecting comprehensibility, which I define in my usability evaluation framework and measure with the HARUS. Legibility or being able to distinguish symbols or figures is a prerequisite to understanding a visualization. As such, I conducted user studies on comparing the legibility of two AR X-ray methods, namely, edge-based and saliency-based.

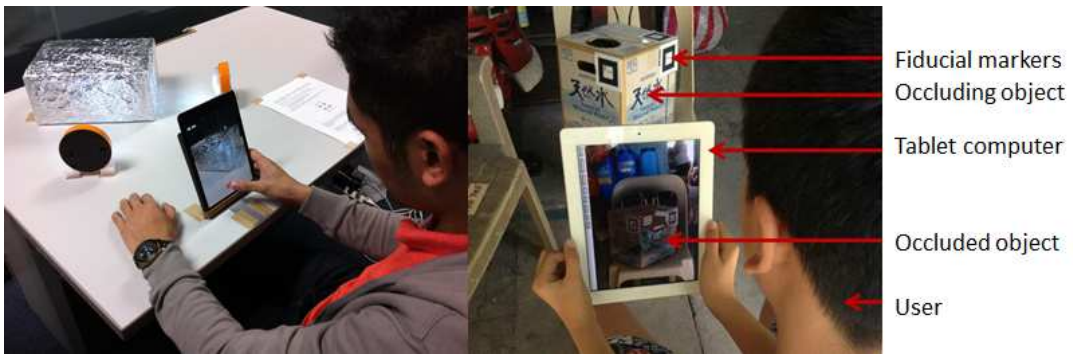

Figure 1.4. Participants were asked to give feedback on several AR X-ray methods.

1.3. Graphical Outline

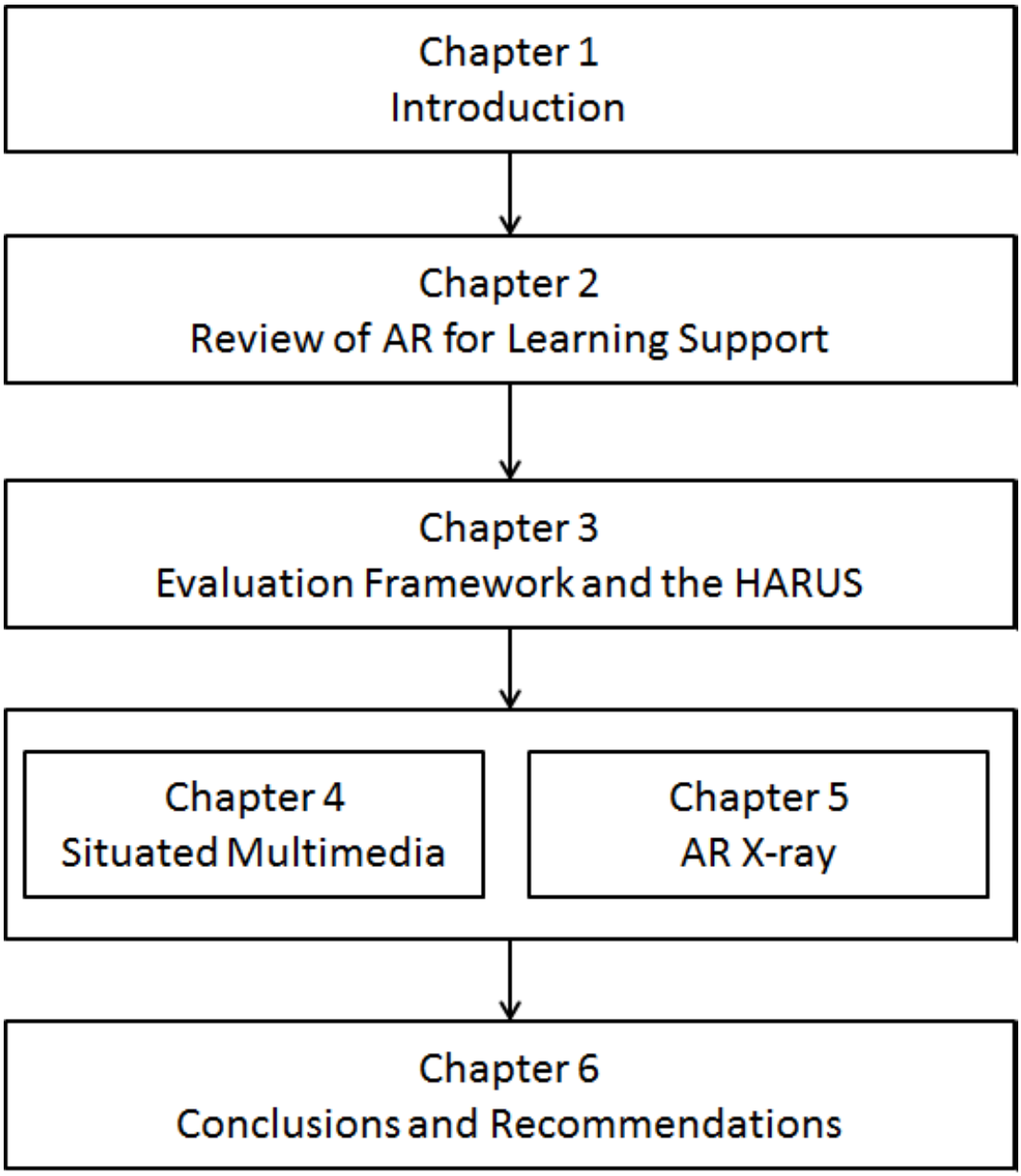

Figure 1.5. Graphical Outline of this Dissertation

CHAPTER 2

Review of AR in Learning Support

There are several research works applying augmented reality (AR) to primary and secondary education, referred to as K-12 education in countries, such as the United States and the Philippines. In this chapter, I reviewed such AR systems that intended to complement existing curriculum materials for K-12. I found 87 research articles on AR for supporting learning in reputable publications. Forty-three of these articles conducted user studies, seven of which compared the effects of AR with more traditional ways of learning. In my meta-analysis of AR learning support, results show that AR systems have achieved a widely variable effect on student performance from a small negative effect to a large effect, averaging to a moderate effect.

I also performed a qualitative analysis on the design aspects for AR systems including display hardware, software libraries, content authoring solutions and evaluation techniques. Based on the existing literature, I conclude that AR incurs three inherent advantages, namely, real world annotation, contextual visualization, and vision-haptic visualization. I explain these advantages through exemplifying prototypes and ground these advantages to multimedia learning theory, experiential learning theory, and animate vision theory.

2.1. Background of the Review

Technology affordances affect instructional design and the manner of teaching. Aside from the content, Dede [40] argues that the technological media (such as computers) have affordances which change the learning experience. Thus, it is important to study the effects of integrating technology in educational settings and how such technologies can be maximized to improve learning. In an attempt to show whether or not people learn better with technology, Tamim et al. [176] conducted a second-order meta-analysis of various technological approaches (using computers for word processing, computer-assisted instruction, distance education, simulation, and hypermedia) against computer-free approaches to learning. Based on 25 meta-analyses representing 1055 primary studies for the past 40 years, Tamim et al. have shown that technology slightly to moderately improves student performance (Cohen's d = 0.35).

The development of AR and its related technologies allows us to create novel AR educational content [11]. Advances in hardware computing power, real-time tracking, graphics rendering, and AR authoring tools contributed to AR applications for educational settings. As a primary goal of the review, this chapter aims to gauge the effect of AR educational content on learning.

Although there are many educational AR prototypes in the current literature, only a few are developed by interdisciplinary groups and base their work on learning theory. Even if the current state-of-the-art execution of AR educational content is effective, it can only be replicated to other contexts if a guideline exists for applying AR for education. As a second goal, I summarize the properties of AR that led to improved learning outcomes. I enumerate the affordances of AR for learning and discuss learning theories relevant to future AR educational content. In addition, I discuss the state-of-the-art implementation and evaluation of AR prototype systems.

2.2. Related Work

AR offers a different set of affordances from more traditional interfaces. Therefore, it will be used differently from other technologies when it is applied for learning

support. To maximize the use of AR, we need to leverage its natural capabilities.

Before reviewing AR applications in learning support, its important to clarify the definition of AR. Azuma [10] defines AR to be when "3-D virtual objects are integrated into a 3-D real environment in real time." First, the definition requires the combination of virtual elements and real environments. It is helpful to think of AR as part of a virtuality continuum introduced by Milgram and Kishino [130]. On one side of the virtuality continuum is the purely real environment, whereas on the other side is the purely virtual environment. AR sits between these two extremes. The second requirement is three-dimensional registration such that the virtual elements are aligned to the real environment. The last requirement is real-time interactivity with the virtual elements. The virtual elements must behave like a real object in the real environment. This may mean, but is not limited to, the AR system responding to changes in the perspective of the user, changes in lighting conditions, occlusion and other physical laws.

In current practice, many applications do not use three-dimensional virtual objects. Instead, they add two-dimensional images on flat surfaces like table-tops and books. Furthermore, many applications do not have perfect integration or three-dimensional registration. The quality of implementation of integration in the current literature varies from imitating the effect of AR [167] on a full integration in an outdoor environment [177]. In the former, the effect of AR is simulated only by flashing relevant information on a screen. It does not employ any kind of tracking. On the other hand, some systems integrate more sophisticated computer graphics on a complex real environment.

2.2.1 AR for Learning Support

AR affords different ways of interaction with information which can be used to support learning. For this review, I relaxed definition of AR to accommodate more prototypes that could help us understand how AR can be used for education. Chang et al. [24] enumerate contents that AR systems can facilitate in various subjects like physics, chemistry, geography and mathematics. They also suggested the use of AR in educational games for primary education. Aside from these contents, Lee [106] mentioned the use of AR systems for astronomy, biology, geometry and cultural heritage. Billinghurst and Dünser [16] explain that these

kinds of content depend on the abilities of AR to illustrate spatial and temporal concepts and emphasize contextual relationships between real and virtual objects. Moreover, AR provides intuitive interaction, enables visualization and interaction in 3D, and facilitate collaboration.

For example, the work of Matsutomo et al. [125] implemented an AR system that demonstrates the advantages mentioned by Billinghurst and Dünser. In their AR system, virtual magnetic fields were integrated with painted blocks acting as magnets (Figure 2.1). Students can move the magnets around to see how various orientations would affect the shape of the magnetic fields. In this example, the space covered by the magnetic field and its variation in time are illustrated via AR. The magnetic field moves with its magnet and changes its shape as it approaches another magnet. The magnets can be moved by hand, thereby providing tangible interaction. Lastly, this kind of system allows face-to-face collaboration wherein students can discuss the learning material in front of them.Participants were asked to give feedback on several AR X-ray methods.

2.2.2 Hardware, Software, and Content

Designing AR systems involves hardware, software, and content. The hardware dictates the computing power and the physical interfaces for input and output. Current AR systems use desktop computers and handheld devices, such as smartphones as the AR platform. Researchers using desktop computers have three options for the display, namely, a computer monitor, an overhead projector, or a head-mounted display (HMD). The choice of device alone affects which software and content would be appropriate. On one hand, desktop systems have bigger screens and higher computing power. On the other hand, handheld devices are more personal and more mobile.

The software should maximize the computing power of the hardware, manage the content display, and handle user inputs. The unique aspects of real-time tracking and three-dimensional rendering are mostly achieved using either open source or commercial AR libraries. AR libraries are good enough for specific applications. Currently, there are many open source and commercial AR development kits suitable for many types of platforms. Among those mentioned that were reported in the literature I reviewed are: ARToolkit (FLARToolkit,

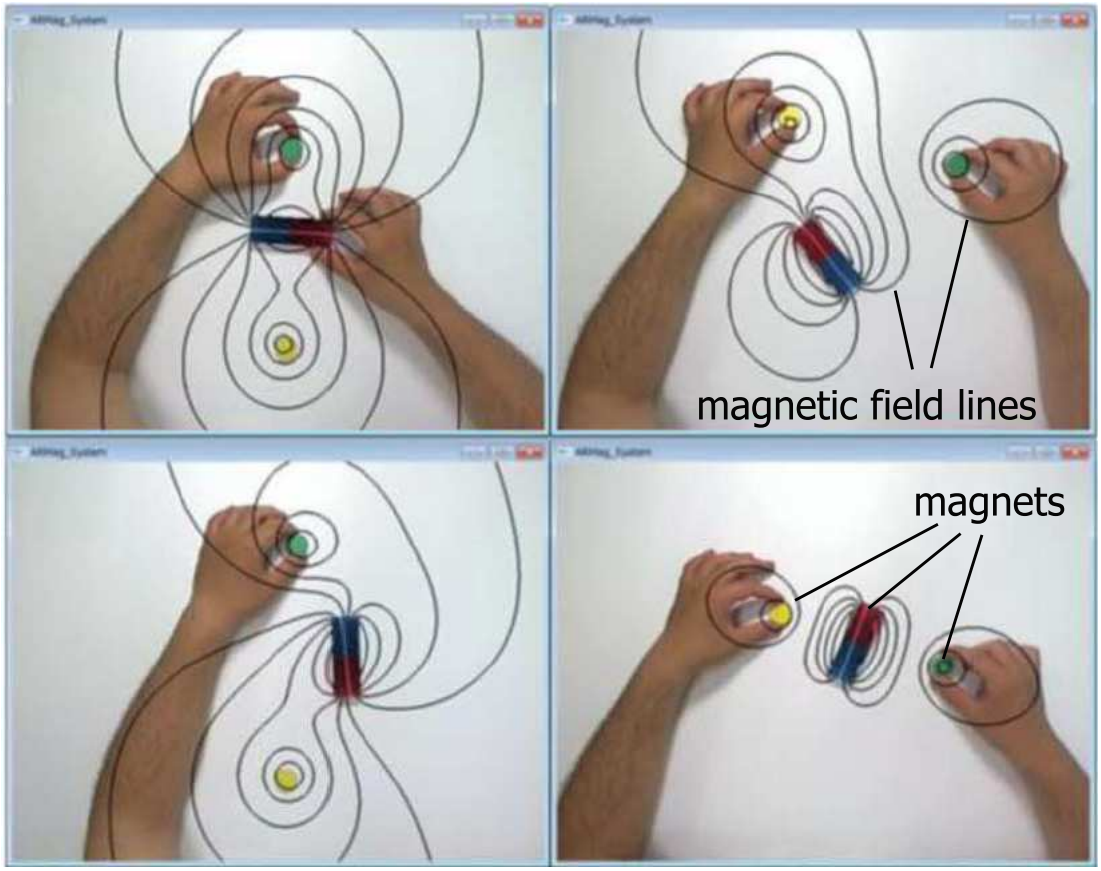

Figure 2.1. The AR system by Matsumoto et al. [125] demonstrates the five abilities of AR that designers can leverage. Virtual lines representing magnetic field lines are added onto magnets.

NyARToolkit), Eyesweb, HUMANAR, Junaio, Opira Registration Library, Popcode, Wikitude, and Zooburst. More recently, AR applications are implemented using the Augmented Reality Vuforia SDK together with the Unity Game Engine.

Content-related aspects include instructional design, authoring tools, and content management tools. This chapter discusses learning theories as basis for effective learning experiences with AR. It discusses the design practices in the current literature to identify what has worked for other researchers.

Content is largely affected by the authoring tools available. Authoring tools are interfaces that allow users (e.g. teacher) to create AR content. In cases

wherein the user is not familiar with programming, which is the more common case, simple authoring tools are necessary to allow the user to create content.

Content management tools are tools for handling content from data storage to device display. AR content can be stored in the desktop PC itself. In cases wherein the desktop PC is in a network, some prototypes have used a server internal to the school. For some commercial AR development kits, a service for hosting the virtual data is available. Delivering location-aware content to HAR is a big technical challenge. However, there are existing solutions that are already explored under the fields of mobile learning, computer-supported ubiquitous learning, and adaptive hypermedia. Existing technologies such as the Hewlett-Packard Mediascape Toolkit (mscape) can be used to deliver location-aware content [172]. For example, Li et al. [110] implemented an interactive mobile learning system that can fetch multimedia content by sending an image of a real object or by entering a predefined geographical area. Chang et al. [26] have explored a network architecture for fetching relevant data about a target learning object found in the real environment. Their goal is to provide a location-adaptive handheld AR (HAR) system. For remote education, Cai and Song [22] discussed a one-to-many remote video learning infrastructure. In this platform, the students can receive lectures from the teacher, and they can also receive AR content for home-viewing using a simple set up involving a handheld device and printed markers.

2.3. Approach for Systematic Review

I conducted a systematic literature review based on the work of Schmid et al. [162]. For my review, I conducted both a quantitative analysis and a qualitative analysis.

2.3.1 Quantitative Analysis

I conducted a literature search on May 30, 2012, in the IEEE Xplore Digital Library. The search string used was: ("augmented reality") AND (educat* OR instruct* OR learn* OR teach* OR train*). To complement the candidate articles found in the IEEE Xplore Digital Library, the search string "augmented reality"

was used to search the publications listed by the Centre of Learning Sciences and Technologies [1]. Most of the 74 journal titles were accessible through the following online bibliographic databases: EdITLib Digital Library, IATED Digital Library, Inderscience, Sage Journals, ScienceDirect, Springer, Taylor & Francis Online, and Wiley Online Library.

The search is limited to journal articles and conference proceedings that are written in English and accessible before June 2012. A total of 503 articles (458 conference proceedings, 42 journal and magazines, 3 early access articles) were gathered from this initial search in the IEEE Xplore Digital Library. Another 150 articles were retrieved from other online bibliographic databases.

The focus of this survey is AR systems for primary and secondary education. Thus, for the research paper to be included, the following criteria must be met:

1. The research paper must have at least a preliminary AR system.
2. The AR system should be applied to learning information or skills.
3. The content should be relevant to grade school or high school.
4. The full research paper is publicly accessible.
5. The paper reports an effect size or provided a means to calculate the effect size; i. e., both mean and standard deviation values are available.

Applying these criteria resulted in 7 articles. From these 7 articles, I computed effect sizes using the formula:

$$d = \frac{\bar{x}_e - \bar{x}_c}{s} \tag{2.1}$$

where $\bar{x}_e$ is the mean of the experimental treatment, $\bar{x}_c$ is the mean of the control, and s is the pooled standard deviation:

$$s = \frac{\bar{s}_e + \bar{s}_c}{2} \tag{2.2}$$

where $\bar{s}_e$ is the standard deviation of the experimental treatment, $\bar{s}_c$ is the standard deviation of the control. I interpret the calculated effect size based on Cohen's recommendation; i. e., an effect size of 0.8 or higher is considered large, around 0.5 is moderate, and around 0.2 is small.

2.3.2 Qualitative Analysis

I conducted the same search as Section 2.3.1. I applied the same inclusion criteria except the fifth criterion requiring an effect size. This search resulted in 87 articles, with 62 articles indexed by IEEE. These 87 articles do not represent 87 unique prototypes because some papers discuss the same prototype.

Moreover, not all these prototypes strictly adhere to the definition of AR of integrating three-dimensional virtual objects onto real environments in real-time. For the purposes of gathering insights in implementing and evaluating AR prototypes, I include the prototypes that use images instead of three-dimensional virtual objects. I also included prototypes that simulate the effect of AR but did not implement tracking of the target object.

I prepared a questionnaire to facilitate the data gathering. The questionnaire has four main parts, namely, publication details, prototype description, use of AR, and design and results of the user study. The publication details refer to the title of paper, name of authors, name of publication venues, etc. The prototype description covers hardware, software, and content descriptions. The use of AR refers to the possible functions of technology and the natural affordances of AR. Schmid et al. [162] listed some of the primary functions of technology in the education setting. For example, technology is commonly used to enrich and/or increase the efficiency of content presentation. The works of Brill and Park [19], and Blalock and Carringer [17] have identified some natural affordances of AR as exemplified in the previous literature. For example, many AR applications use the annotation capability of AR to rapidly and accurately identify objects in the real world. The design and results of the user study refer to the description of the control and experimental groups, the construct being measured, the effect of AR on that construct, etc. For example, aside from student performance in pre-tests and post-tests, other aspects of the learning experience, such as motivation and satisfaction were observed.

The clarity of the questionnaire was evaluated by having two researchers use it separately on 20 papers out of the 87 that pass the inclusion criteria. There were only minor misunderstandings of the questionnaire and these were clarified before proceeding to read the remaining 67 papers. Each of the 67 papers was read only by one researcher.

Table 2.1. Studies Evaluating Student Performance with Effect Sizes

Ref.	Description	Content	Participant	N	Effect
[116]	AR situated learning around the campus	English	Grade school students	67	1.00
[77]	Physics props are annotated with measurements and graphs using AR.	Kinematics graphs	High school students	80	0.86
[124]	With spatial ability training using AR	Spatial ability	University students	49	0.71
[47]	AR annotated print out replicas of art pieces	Renaissance art	High school students	69	0.67
[108]	Collaborative AR learning wherein students simulate collision.	Elastic collision	University students	36	0.58
[76]	AR learning using magic book	English	Grade school students	>30	0.37
[28]	AR situated learning in the library	Library skills	Grade school students	116	-0.28

2.4. Results of the Quantitative Analysis

Eleven articles evaluated AR prototype systems by conducting experiments to compare the performance of students who use their system versus a non-AR approach. Seven of these articles allow the computation of an effect size. The seven AR systems and their corresponding effect sizes are summarized in Table 2.1. AR systems achieved a widely variable effect on student performance ranging from a small negative effect to a large effect. The mean effect size is 0.56. The four additional articles that conducted other student performance evaluations are listed in Table 2.2.

Table 2.2. Other Studies Evaluating Student Performance

Ref.	Description	Content	Participant	N	Result
[169]	AR magic book	Solar system	High school students	40	29% increase
[143]	In situ AR game	Math game	Grade school students	123	No Sig. Diff.
[131]	Projection AR for note-taking	Eulerian graphs	University students	20	No Sig. Diff.
[50]	Collaborative AR with 3D shapes	Spatial ability	High school students	215	No Sig. Diff.

2.4.1 Affordances of AR

Researchers designed their AR systems to take advantage of the following affordances of AR technology:

1. Real world annotation - displaying text and other symbols on real world objects. For example, [77] annotates a real moving ball with values of velocity and the corresponding graph.

2. Contextual visualization - displaying virtual content in a specific context. For example, [28] uses AR to teach library skills by adding virtual information to a library.

3. Vision-haptic visualization - enabling embodied interactions with virtual content. For example, [124] allows the user to view a three-dimensional model on a marker which can be manipulated with bare hands.

2.4.2 Strategies for AR Use

Aside from the inherent affordances of AR, strategies have been applied to create more effective AR systems. Researchers have used the following strategies:

1. Enable exploration - designing AR content that is non-linear and encourages further study. For example, [108] allows students to try out different kinds of scenarios of collision of two balls and see if the collision will happen in the way they hypothesize it to be.

2. Promote collaboration - designing AR content that requires students to exchange ideas. For example, in [143], students were given different roles and asked to negotiate with each other to arrive at a solution.

3. Ensure immersion - designing AR content that allows students to concentrate more and be engaged at a constant level. For example, in [47], students were able to concentrate more using AR as opposed to a standard slide presentation.

2.4.3 Recommendations for AR in Learning Support

The mean effect size of 0.56 of AR systems to student performance should be interpreted carefully. On one hand, it is a good snapshot of the effect of AR technology when used in educational scenarios. However, we must not think of AR as a homogeneous intervention in the learning process because it has a wide design space. The seven articles presented in Table 2.1 include different display devices, content, and experimental design. Moreover, reading these papers individually also reveals that factors, such as instructional design, may have played a crucial role in the success of the AR system.

Learning objectives, pedagogy, teaching expertise, subject matter, grade level, consistency of technology use, and other factors may have a greater influence compared to the unique capabilities of AR [176]. However, as Dede [40] argued, technology affordances affect how the content is designed. The changes in instructional design may either be because of imperfect control of the variables, or because of the technology used. These findings should be interpreted carefully and should only be used as a guide, specifically because the effect sizes vary widely. For future AR learning support systems, I recommend the following:

1. Measure learning that can be attributed to AR systems. To do this, the AR system must be used by an experiment group who will receive the

AR treatment and compared to a proper control group. Instructional design, pedagogical approaches, teaching practices, and other factors should be carefully controlled so that only the AR intervention is made variable. Imperfections in controlling these aspects should be taken into account in the interpretation of the results. A heuristic for this is to ensure that both the AR approach and the AR-free approach are both the best possible design for the particular content. I adhere to this recommendation in my evaluation of situated vocabulary learning in Chapter 4.

2. Report the effect size, or report both the mean and standard deviation of the performance of students. The effect size is a good measure to compare across the design space for AR systems. However, not all research articles that we reviewed report an effect size.

3. Apply the inherent advantages and suggested strategies for AR as needed by the educational scenario. These insights guide the appropriate use of AR.

2.5. Results of the Qualitative Analysis

Eighty-seven papers were found in the current literature when I applied the inclusion criteria to the initial search result. The graph in Figure 2.2 shows the distribution of the publication year of these papers. Starting 2007, there is an increasing number of papers discussing AR prototypes. This review included papers published until June 2012. It does not include papers from July to December 2012. Of these 87, 72 are conference papers, whereas 15 are journal articles. Sixty-one papers are indexed by IEEE Xplore, whereas the other 26 are found in other digital libraries. Based on these 87 research articles, I summarize insights on display devices, content creation, and evaluation techniques. I then discuss the related theories supporting the benefits of AR for supporting learning.

2.5.1 Display Devices

Choosing the appropriate display is an important design decision. In the current literature, there are four types of AR systems based on the device used for

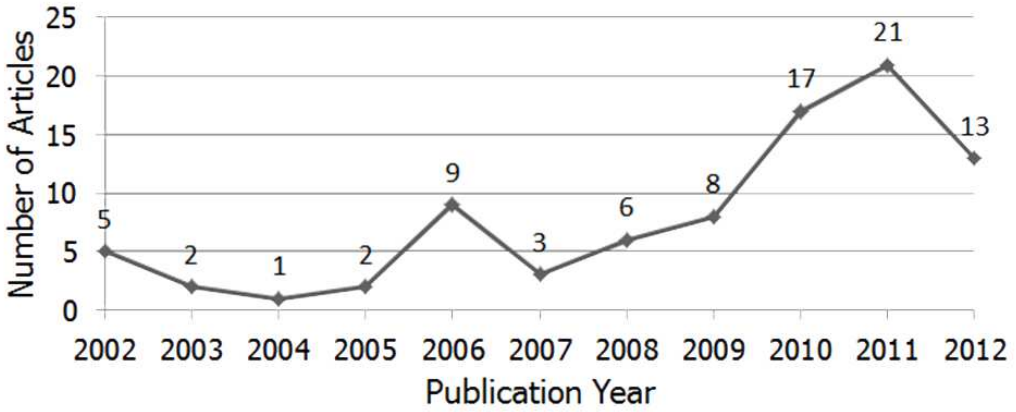

Figure 2.2. Publication of AR Systems for Learning Support per Year until June 2012.

display, namely, computer monitor, handheld devices (smartphone, tablet, etc.), overhead projector, and head-mounted display. Table 2.3 lists display devices with exemplifying AR systems and their corresponding contents.

Researchers have also distinguished the displays as either using a *mirror metaphor* or *glasses metaphor*. In [166] and [157], researchers have made this distinction as perspectives. The glasses metaphor is a first person perspective or AR that is based on what the user can see in front of him. On the other hand, the mirror metaphor is a third person perspective wherein the user becomes an observer of himself.

The mirror metaphor is when a screen appears to be a reflection facing the user, except the user can also see the virtual images integrated to his or her reflection. Figure 2.3 (left) shows an example of the mirror metaphor. We see the reflection of the person as well as the virtual information (vertebral column).

Desktop computers with large monitors are usually used for the mirror metaphor. The mirror metaphor has been applied in AR systems to provide students with some compelling learning experiences. For example, Blum et al. 2012 [18] used the mirror metaphor in presenting an X-ray-like application wherein the user is given an illusion of being able to see inside his body. This kind of system would

Table 2.3. Sample AR Systems with Corresponding Display Devices

Display	Ref.	Content
Desktop monitor	[47]	Visual art pieces
	[18]	Human anatomy
	[54]	Chemistry concepts
Handheld devices	[177]	Butterfly life cycle
	[65]	Electrical circuit
	[93]	Architectural history
	[67]	Physical education
Overhead Projector	[154]	Spelling
	[191]	Playing the drums
Head-mounted display	[109]	Chinese characters
	[169]	Solar system
	[84]	Endangered animals

be useful for students studying human anatomy and sports science to help them connect their understanding of human movements and muscles. In this type of application, the mirror metaphor becomes advantageous because the content is about studying the human body itself.

The glasses metaphor refers to displays wherein a user appears to be looking into the world with a pair of special glasses. In this case, virtual information is integrated to what the user sees in front of him. Figure 2.3 (right) shows an example of the glasses metaphor. Three devices have been applied for AR systems under the glasses metaphor:

1. Head-mounted Display – In [169], Sin and Zaman used the glasses metaphor to present virtual heavenly bodies on AR markers which they can manipulate. Students wore a head-mounted display so that both hands would

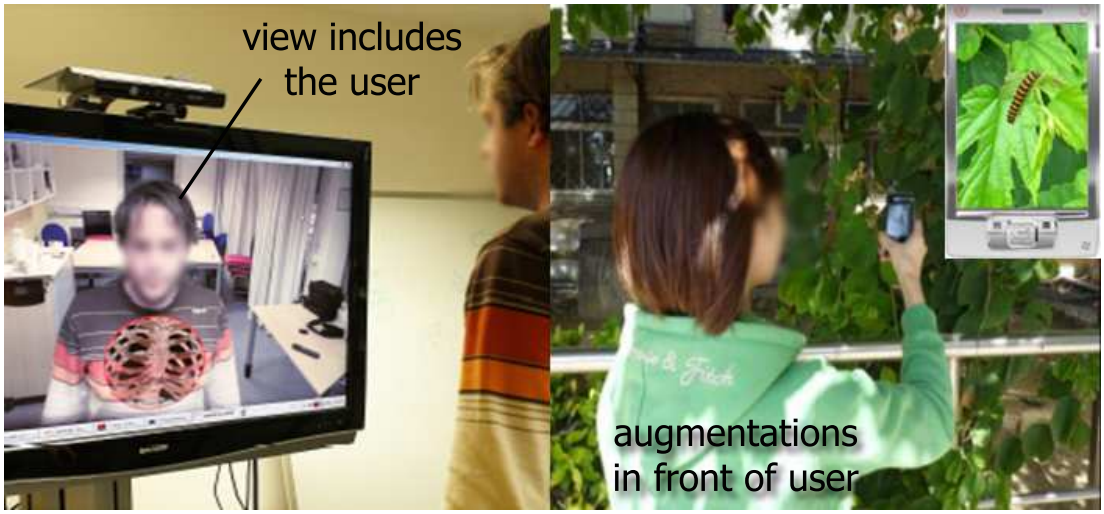

Figure 2.3. Mirror Msetaphor [18] and Glasses Metaphor [177].

be free to handle the markers containing the virtual solar system. Another study by Shelton and Hedgely [166] had argued that this visualization is advantageous because students can easily understand concepts, such as day and night, when they can test for themselves what happens when one side of the earth is occluded.

2. Handheld Devices – In the work of Tarng and Ou [177], students can view virtual butterflies in a real school garden to understand the butterfly life cycle. The use of handheld devices is advantageous because students need to move around an area. In this case, all the processing, data connectivity, and display are found in a single device. Some researchers point out that nowadays, many people own smartphones and tablets which are ideal platforms for AR. These handheld devices are equipped with fast processors, graphics hardware, large touchscreens, and various sensors like camera, GPS, compass, and accelerometer [16].

3. Projector – Projector-based AR affords more user movement in a confined space, say a room, than using desktop systems with monitors. However, it does not afford as much movement as when using handheld devices. Furthermore, the display of projector-based AR is bigger than computer

monitors and smartphone screens. The projector-based system have been successfully used to create a training system for playing a drum set [191], wherein, the drums are annotated with signals for when to hit them. Researchers have pointed out that desktop computers and overhead projectors are already available in most schools making them convenient hardware for AR systems.

2.5.2 Content Creation

The main concern in creating AR content for AR systems are authoring tools and content design. Developers of AR systems usually use AR libraries such as the ARToolkit to create the prototype. However, teachers need authoring tools that would allow them to create content without having to be proficient in programming.

According to Wang et al. [184], authoring tools for non-programmers can be low-level or high-level. Low-level tools require some coding or scripting skills, whereas, high-level tools use visual authoring techniques. Both types usually would make use of drag and drop interfaces and menus. Currently, there are several authoring tools for any type of AR application targeting non-programmers such as DART, ComposAR, AMIRE, and MARS which are discussed briefly in [184]. A basic authoring tool would be BuildAR [184] which allows the teacher to scale, translate, and position virtual objects with respect to a fiducial marker. For example, virtual art pieces can be displayed in a real room to create an AR museum experience [175].

In the current literature, some researchers have developed ways to author AR content for learning. Researchers are exploring mainly three kinds of educational material, namely, magic books, learning artefacts, and location-based content. Researchers consider the book metaphor as one of the main modes of how AR will be used in education [16]. Using the book as the real element, additional virtual content can be added onto it using AR making it a magic book. Students are already familiar with books which afford many natural ways of interaction like flipping pages and viewing the book at different perspectives. For example, Jee et al. [76], [75] talks about an authoring tool for markerless books. The work of Vate-U-Lan [179] used a commercial authoring tool, Zooburst [3], for making

an AR system based on the magic book metaphor.

Another trend in the current literature is the use of learning artifacts wherein students learn about a particular real object. Rahman et al. [151] developed an authoring tool that would allow a teacher to present a small object on a camera and annotate that object with virtual information using a live video visualization. Aside from a desktop computer, they used a depth camera to capture the color image and corresponding depth information. They used the "polygonal annotation scheme" wherein the teacher could draw a polygon on the screen where he or she could see the object he or she wants to annotate with information.

In [168], Simeone and Iaconesi have pre-trained a system to recognize parts of a sculpture called the Minkisi. Users can then annotate virtual information on the real parts of the Minkisi via a desktop system. They describe their use case to involve multiple types of users, such as curators, teachers, and students. Each of them can annotate information on specific parts of the artifact based on various online sources. Users can then evaluate each other's information based on accuracy and usefulness. Therefore, content is developed through communication among the different author-consumers with the physical object as the central point of conversation. Figure 2.4 shows the Minkisi and the two parts where students can annotate information.

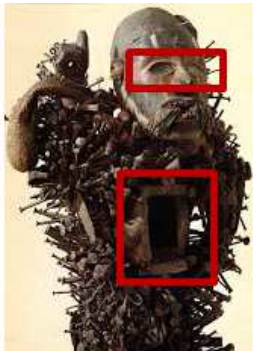

Figure 2.4. The Minkisi artifact and the Parts of Interest [168]

Location-based games take advantage of providing relevant information to a place such as buildings and artifacts. Researchers have shown that novel e-

learning experiences can be provided by delivering content relevant to a place (e.g. history and language). Chang et al. [25] have demonstrated this using handheld devices equipped with RFID tag readers and GPS positioning to deliver information regarding objects found in the real world. Although the game itself represents most of the pedagogical design, AR is an effective display for learners because it integrates the virtual information directly onto the target real world object. Moreover, AR has the advantage of using the camera for detecting and tracking an object in the real world, as opposed to using RFID readers which are not readily available in handheld devices. Furthermore, putting RFID tags on the real world may not be feasible in cases wherein the real objects should not be touched, such as some historical artifacts.

AR can offer location-based content usually used for outdoor AR experiences using handheld devices. Klopfer and Sheldon [96] have developed the AR GameBuilder which allows teachers to create educational AR mobile games in a specific locality. Their system offers predefined maps and GPS coordinates to place virtual objects such as characters in a specific place in the world. According to Klopfer and Sheldon, their platform also allows students to transfer an existing game to another place, such as their own city.

AR authoring itself is seen as an educational experience. For example, Billinghurst and Dünser [16] have used BuildAR to let students create their own AR scene. In this process, students develop skills in mathematics, drawing, and story-telling. In this case, it is important that the authoring tool is usable for primary school students. In the case of the prototype of Simeone and Iaconesi [168], the content is developed in a conversational style wherein the cultural artifact becomes the center of the discussion. The students could verify each other's inputs and the teachers can direct the discussion based on the learning objectives. The prototype by Odeh et al. 2012 [137] allows students to add virtual wires onto a real circuit and conduct simulated experiments. Such systems make the student feel the reality of the components and instruments used in an engineering experiment without physically accessing the laboratory. Moreover, the remote laboratory is available to them for much longer time than a regular laboratory. Lastly, Klopfer and Sheldon have tested the AR GameBuilder and the students have successfully used it for creating linear games. Table 2.4 summarizes AR

authoring activities found in the literature.

Table 2.4. Authoring Activities in AR Learning Support

Ref.	Authoring Activity
[137]	Students can add virtual wires on a real circuit in a remote laboratory set up.
[76], [75]	Teachers can author e-learning applications using a markerless magic book metaphor.
[168]	Teacher, students, and the curator can all access and add information related to a cultural artifact.
[32]	Real props for physics classes, such as balls, carts, and rods are augmented with virtual information such as text and arrows.
[96]	Teachers and students can create and access educational games using predefined maps and GPS coordinates, virtual characters, and file management support.
[175]	Teachers can make a web-based virtual museum by selecting 3D models, associating them to markers, and arranging them in a room.
[151]	Allows a video annotation approach in order to catalogue and add virtual information on physical learning artifacts in a scene.
[170]	The CONNECT Visual Designer allows educators to specify the interactions the learner can have within the AR environment by creating rule-based scenarios.

2.5.3 Evaluation Techniques

Of the 87, 43 papers have performed user studies on the system to observe ease of use, satisfaction, immersion, student motivation and performance, among others. The number of students involved in the study varied from 4 [31] up to 419 [68]

with a median sample size of 36 students [108]. The proper choice of evaluation method for an AR system depends on the purpose of the evaluation. In my review, I observed two primary purposes, namely, to show whether or not an AR system is beneficial to learning, and to quantify some user experience or usability issue and discover possible improvements.

Researchers need to demonstrate the benefits of using their AR system. Thus, they compare either the performance or the motivation of students when using an AR system (the experimental treatment) and when using a more traditional medium of instruction (the control). To measure student performance, the students take a test to measure their mastery of the content. Scores of students belonging to the experiment group and control group are then compared to see any differences in learning. Such comparison between AR system users and non-users are summarized in Table 2.1. Aside from possibly improving student performance, AR systems can be used to increase the motivation of students in educational settings. Abstract constructs such as motivation can be measured by valid and reliable questionnaires, such as the Instructional Materials Motivation Survey (IMMS) [94] and Intrinsic Motivation Inventory (IMI) [2]. In Chapter 4, I compared both measures of performance and motivation to demonstrate the benefits of using AR system in a memorization task.

Researchers also evaluate their AR systems to measure some aspect of user experience and discover possible improvements to the current prototype. In this evaluation, user study participants are observed while they use the AR system, and asked questions in the form of an interview or questionnaires after. Survey questionnaires are the most commonly used evaluation tool in the current literature. Questionnaires are designed to measure a construct such as the user's feelings of satisfaction, enjoyment, or immersion while using the system. After researchers decide on a construct to observe, they either use existing questionnaires, or create their own questionnaire.

Some questionnaires have been tested for validity and reliability; i. e., these questionnaires have been previously shown to accurately measure the the construct of interest. Table 2.5 and 2.6 lists several questionnaires that have been used for AR systems. Currently, there is a need for valid and reliable questionnaires to accurately measure relevant constructs to AR systems for learning

support (e.g. immersiveness). Moreover, evaluation frameworks and usability questionnaires are needed to iteratively improve AR systems. In Chapter 3, I discuss my evaluation framework for HAR and the Handheld AR Usability Scale (HARUS).

Based on my review, some researchers have used the ISONORM which is a general software usability questionnaire [173], [91]. Using this questionnaire, they were able to observe aspects of interface design, such as conformity with user expectations, controllability, error tolerance, self-descriptiveness, suitability for learning, and suitability for the task.

Among the most observed constructs are ease of use, usefulness, and intention to use. In the current literature, researchers usually ask directly if a system is easy to use, if the user thinks it is useful, and if they would use the same system for other subject matters. Therefore, most of the available literature measure perceived ease of use and perceived usefulness. However, it is possible to check for objective measures of ease of use such as counting errors when using the interface and time on a certain task.

Aside from using questionnaires, other evaluation methods also have their own advantages depending on the context of evaluation. Such methods are as follows:

1. *Interviews* are useful for learning about qualitative data that cannot be captured by written responses to questionnaires. For example, interviews were useful in learning about technology acceptance [173], [37]; possible benefits of AR to the current practice of teachers [95], [33]; and learners' opinion about technology including perceived ease of use, perceived usefulness, intention to use, etc. There are also cases in evaluating AR systems wherein interviews would be preferred compared to questionnaires. In cases wherein the respondents are young children or persons with disabilities [194], it is better to conduct interviews to communicate more effectively with the participant.

2. *Observing and coding overt behaviors* have been adopted by several researchers to see how their target user would interact with an AR system. Observation is done to reveal possible improvements for better performance and satisfaction of the user. Behaviors can be divided into two: verbal and

nonverbal. Verbal behaviors can be specific keywords, expressions, questions, or statements a learner says while using the AR system. Nonverbal behaviors include facial expressions (frowning, smiling, surprise, etc.) or body language (fidgeting, leaning close to the AR interface, scratching the head, etc.) [178].

3. *Expert review* was used by Margetis et al. [122] to evaluate touch-based interactions with AR system based on the book metaphor. They employed 4 usability and interaction design experts to perform heuristic evaluation with questionnaires based on the work of Nielsen and Mack [136]. The main goal of the expert review is to identify potential usability problems, and check conformity against the five dimensions of usability: effective, efficient, engaging, error tolerant, and ease of learning [149]. Currently, it is difficult to conduct expert reviews for AR and HAR because there are limited design guidelines to use as references. In Chapter 7, I discuss my efforts and future work toward establishing design guidelines for HAR applications intended for learning support.

2.6. Discussion of AR Annotations

The most basic uses of AR is the annotation of real world objects and environments. Merriam-Webster defines annotation as "a note added by way of comment or explanation." The data type being annotated is usually text that explains a concept. Many AR applications use text as the virtual information being overlaid to the real environment. However, annotation with AR is not limited to text. It could also involve other symbols and icons. This includes, but is not limited to, arrows and basic shapes, such as circles and lines, used to highlight or direct a user's attention. In this thesis, AR annotation is the juxtaposition of real world objects or environments with virtual text or virtual symbols that help explain a concept to a user. In Chapter 3, I used text labels and arrows. In Chapter

Table 2.5. List of Studies Using Previously Validated Questionnaires

Ref.	Metrics or Constructs	Tools
[47]	attention, confidence, relevance, satisfaction	Instructional Materials Motivation Survey [94]
[70]	enjoyment, competence, usefulness, tension	Intrinsic Motivation Inventory [2]
[152]	challenge, collaborativeness, competition, ease of use, movement, rewards, situated learning	Constructivist Multimedia Learning Environment Survey [121], Preferences for Internet Learning
[108]	collaborativeness, interest, perceived skill development	Learning Effectiveness [5]
[21]	use of computer, use of video game, science process	Self-Efficacy in Technology and Science Short Form [6]
[27]	attitude to e-learning, e-learning experience	Technology Acceptance Model [38]
[173]	controllability, ease of use, learnability, self-decriptiveness	ISONORM Usability Questionnaire [148]
[91]	conformity with user expectations, controllability, error tolerance, self-descriptiveness, suitability for learning, suitability for the task	ISONORM Usability Questionnaire [148]

Table 2.6. List of Studies Using Original Questionnaires

Ref.	Metrics or Constructs
[169]	ease of use, effectiveness, learnability
[124]	attractiveness, ease of use, usefulness
[83]	ease of use, enjoyment, perceived skill development
[84]	ease of use, enjoyment, usefulness
[175]	enjoyment, perceived presence
[7]	wearability
[82]	ease of use, engagement, perceived presence, usefulness
[118]	ease of use, intention to use, usefulness
[117]	ease of use, usefulness
[170]	attitude, ease of use, interest
[8]	ease of use, intention to use, learnability, perceived correctness and responsiveness of system
[163]	ease of use, intention to use, perceived correctness and responsiveness of system
[103]	ease of use
[37]	ease of use, expectations of AR, perceived affordances, usefulness
[52]	ease of use, perceived efficiency, usefulness, preferred subjects to use AR
[191]	comfort, enjoyment, intention to use, interest, perceived skill development, usefulness

4, I used text labels and simple sprite sheet animations. Lastly, in Chapter 5, I studied how annotations inside target objects can be made more legible for the users.

In AR systems that use AR annotation, a set of real objects become part of the learning content. The set of objects is augmented with text information or other symbols with the purpose of providing a learning experience. The virtual information is the text or symbol, whereas the real environment is the set of objects. To be consistent to the definition of AR, AR annotation requires tracking the physical objects such that the text information appears as labels that follow the real object. I apply this in all the systems discussed in Chapters 3, 4, and 5.

The benefits of AR annotation can be explained using multimedia learning theory [127]. In this theory, multimedia refers to words (written or spoken) and pictures. Multimedia learning theory has three assumptions, namely, dual channels, limited capacity, and active processing. The first assumption is that there are two separate channels for visual information and auditory information. The second assumption is that both these channels can only accommodate a limited amount of information at a given time. Lastly, the third assumption is that humans are active learners. Incoming information from the channels are processed by organizing them into coherent mental representations and integrated to previously acquired knowledge. Based on these three assumptions, Mayer [127] has derived and empirically proven design principles in authoring multimedia learning materials. Of these principles, the following are directly related to AR annotation applications, namely, *Multimedia Principle, Spatial Contiguity Principle, and Temporal Contiguity Principle*.

Multimedia learning theory can be extended to AR annotation by doing two substitutions:

1. The set of real objects replaces the picture.

2. The virtual texts and symbols replaces the words.

From this theory, it can be argued that learning with AR annotated objects is better than learning about the same object with other reference material such as a manual or a separate online source. For example, in learning how to play a guitar, it will be better to learn about the finger positions highlighted onto

Table 2.7. Multimedia Learning Principles Supporting the Effectiveness of AR Annotated Objects [127]

Principle	Extension to AR Annotation
Multimedia and Time Contiguity	People learn better from annotated virtual words onto physical objects than from separate multimedia (e.g. illustrated manual) and physical objects.
Spatial Contiguity	People learn better when corresponding virtual words and physical objects are presented near rather than far from each other on the screen.

the guitar, than referring to a sheet summarizing the finger positions for each chord. By the definition of AR annotation, three empirically-proven principles, namely, Multimedia Principle, Temporal Contiguity Principle, and Spatial Contiguity Principle guarantee that learning with AR annotated physical objects will lead to better learning performance compared to more traditional ways of learning. The extensions of these principles to AR annotation are shown in Table 2.7.

The principles of multimedia learning theory were tested both on printed materials and computer-assisted instructions. It has not yet been tested for AR annotation applications for learning. However, Fujimoto, et al. [56] has demonstrated how the annotative abilities of AR can ease memorization tasks. In their study, the memorization abilities of users were tested when they memorized symbols by annotating information near the target object (Display 1) against displaying the information on a random place while connected by a line (Display 2) and on the same place, say at the top left of the display (Display 3), as shown in Figure 2.5.

Fujimoto, et al. conducted two types of memory tests: identification and association. In these tests, each of the participants are shown 10 symbols one at a time. The identification test asks the participants to identify the 10 symbols they just saw from 25 symbols. Whereas, the association test asks the participants

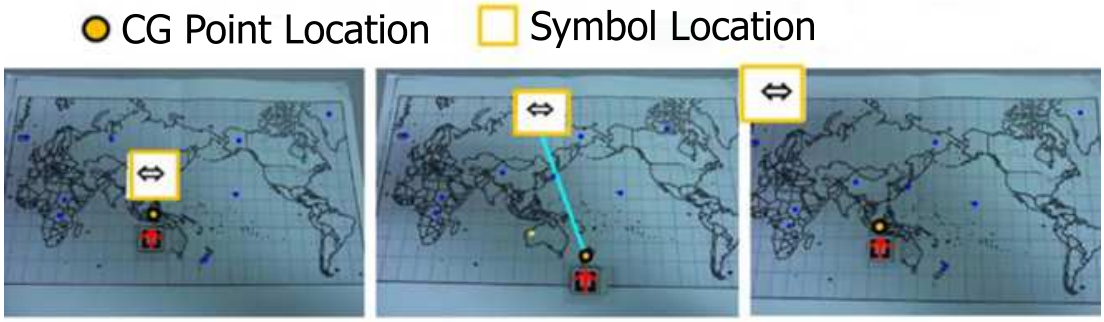

Figure 2.5. Three display methods were tested by Fujimoto et al. [56]. On the left is AR annotation where a label is placed near the relevant part of the map. The middle and right are control scenarios wherein a label is displayed randomly on the map with a connecting line and a label is displayed consistently at the top left of the HMD screen without a connecting line.

to identify where in the map they saw the image. In both tests, Fujimoto et al. measured the accuracy of answers, as well as the time it took for the participant to answer.

Results show that annotating the information on top of an object (Display 1) allowed the users to memorize the labels significantly better than when displaying the information on random places (Display 2), or on the same place on a display (Display 3). In the identification tests, participants were able to achieve an accuracy of 99% with Display 1, 95% with Display 2 and 96% with Display 3. The bigger difference is in the answering time wherein users of Display 1 answered within a shorter time of 45 seconds, compared to 53 seconds and 52 seconds for Displays 2 and 3, respectively.

In the association tests, participants were 87% accurate with Display 1. Whereas, they are only 70% and 75% accurate with Displays 2 and 3, respectively. Furthermore, participants who used Display 1 finished the test in 96 seconds, compared to 112 seconds and 99 seconds for Displays 2 and 3, respectively. All of these tests were proved to be statistically significant in the work of Fujimoto et al. Annotating virtual information near an object makes perception easier, thus it may be possible to utilize AR for better information presentation in educational

settings.

One example of AR annotation is the work of Simeone and Iaconesi [167]. In their work, they trained their system to recognize 3D parts of a model airplane (Figure 2.6.a) so that they can display relevant information for each 3D part. Their system makes use of a personal computer equipped with a webcam. The virtual information can be viewed on the computer monitor together with the real environment including the airplane model and the user. The authors mentioned two use cases. First, instructions on how to assemble the several pieces into an airplane can be annotated onto the parts of the plane. When a user picks up a piece and puts it near the webcam, an instruction relevant to that part is displayed at the bottom of the screen. Instead of the student going back and forth from a manual to the actual objects, the airplane model pieces can be augmented with the manual instructions.

For the second use case, the airplane model can be layered with several kinds of information that the students can access by focusing on specific parts of the plane. The information was taken from various online sources. This prototype is limited in its annotating capabilities because the system does not have a tracking feature. With a tracking feature, the annotated information can follow the real object wherever it is on the screen. However, for the purposes of a prototype, this work is a good approximation of how AR toys can be used in the near future.

Instead of text information, other symbols and shapes can be used to annotate objects. In physics education, magnetic field lines (Figure 2.1) [125] and directions of forces acting on an object (Figure 6.c) [170] have been annotated to real objects like magnets and carts, respectively. With this feature, students can visualize abstract phenomena like magnetic field and force.

Another set of compelling examples can be found in AR systems with the goal of teaching how to play musical instruments. AR applications have been developed to teach people how to play the guitar [132], drums [191], and piano [70]. In [132], a desktop system was used to render a virtual hand on top of a real guitar. The student can then position his hands correctly on a guitar. Instead of the student translating a chord sheet indicating which strings to press at which fret, this information is already annotated on the guitar itself. In [191], a projector-based AR annotation was used to indicate on a drumset which drum to

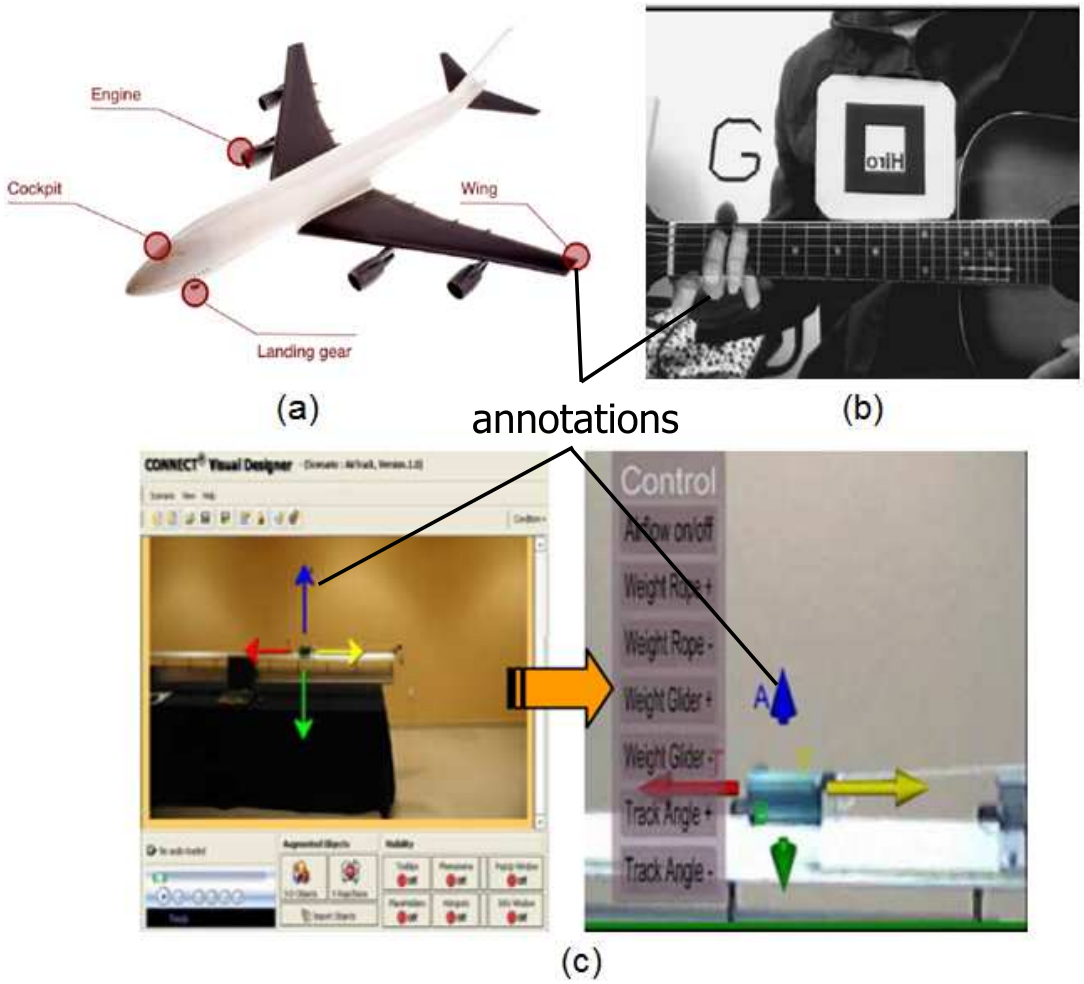

Figure 2.6. AR systems demonstrate real world annotation. (a) shows the parts of the airplane that can be recognized and annotated with words in the work of Simeone and Iaconesi [167]. (b) shows the virtual hands and letter annotated on a real guitar [132]. (c) shows a cart augmented with arrows representing forces acting on it [170].

hit by projecting circles on the top of the appropriate drum. Instead of a teacher demonstrating to the student and pointing which drum to hit, this information can be augmented directly onto the drum set. Lastly, in [70], a desktop AR annotation system is use to demonstrate finger positions on a piano. Instead of

a piano teacher demonstrating the proper finger positions, this information can be augmented on a keyboard. These systems do not intend to replace formal training with music teachers. However, these systems are very useful for self-studying outside of the formal class and for music teachers to create learning materials for their students. The other examples of AR systems using annotation are listed in Table 2.8.

Table 2.8. Examples of AR Annotation

Ref.	Real Objects	Virtual Annotations
[47]	Printed replicas of Renaissance art	Text details relevant to art
[28]	Library	Verbal hints from a virtual character about the structure of the library
[125]	Magnets	Magnetic field lines
[14]	University campus	Signages
[70]	Piano	Highlights correct finger positions
[74]	Go Boardgame	Boardgame pieces added in strategic patterns
[168]	Cultural Artifact	Text details relevant to anthropology
[191]	Drumset	Highlights of sequence for hitting drums
[167]	Airplane model	Description and instruction relevant to the airplane part
[132]	Guitar	Highlights the finger position of chord
[170]	Cart and other objects	Forces acting on the object as it moves
[32]	Ball	Arrows representing instantaneous velocity, acceleration, and centripetal force

2.7. Chapter Summary

AR has unique affordances that can affect the learning experience. Developments in AR technology have enabled researchers to develop and to evaluate AR systems for learning support. These developments encompass hardware, software, and the authoring of content. Currently, AR systems have a mean effect size of 0.56 to student performance with wide variability due to the many possible ways to use AR, as well as, differences in experimental design.

In the course of the development of AR systems, researchers must test their prototypes for benefits in the learning process, and for usability. Such tests must use appropriate control groups, report an effect size, and use standard evaluation frameworks and tools such as usability questionnaires. AR systems have been evaluated through student performance tests, survey questionnaires, interviews, observations of user behavior, and expert reviews. Currently, there is a need for evaluation frameworks and evaluation tools specifically for AR systems. Such evaluations are necessary to iteratively improve the AR systems. In Chapter 3, I discuss my evaluation framework for HAR and provide my own valid and reliable questionnaire.

Based on my review, there are three inherent affordances of AR to educational settings, namely, real world annotation, contextual visualization, and vision-haptic visualization. Furthermore, researchers have used design strategies, such as enabling exploration, promoting collaboration, and ensuring immersion, to create compelling learning experiences. Depending on the learning objective, these affordances and strategies can be employed to create successful AR sytems. In particular, real world annotation may reduce cognitive load based on multimedia learning theory.

CHAPTER 3

Evaluation Framework and Usability Scale

Development of novel interfaces such as handheld augmented reality (HAR) systems requires three types of evaluations, namely, guidelines-based evaluation, formative evaluation, and summative evaluation [60], as discussed in Chapter 1 and illustrated in Figure 1.1. To support these evaluations of HAR, I propose an evaluation framework that defines two new usability constructs, namely, manipulability and comprehensibility. Based on my evaluation framework, I developed the HAR Usability Scale (HARUS), a questionnaire that measures manipulability, comprehensibility, and usability of a HAR system for a specific user group performing a specific task. In this chapter, I focus on exploring the validity and reliability of HARUS. I then highlight some benefits of using this questionnaire compared to other standard questionnaires that were not specifically designed for HAR systems.

3.1. Background

Handheld devices, such as smartphones and tablet computers, now have powerful processors, large screens, and built-in location sensors and cameras. These features of handheld devices make them convenient platforms for augmented reality

(AR) the seamless integration of virtual objects to real environments. Handheld augmented reality (HAR) affords many new ways of interacting with digital content. It is finding applications in various industries such as entertainment, marketing and sales, education and training, navigation, tourism, and social networking. Although several applications have been adopted by general consumers, HAR remains limited and researchers are continuously developing more intuitive interactions using handheld devices.

Usability refers to how well target users can use a functionality of a system [135] to accomplish a specific task. Usability studies are important for iteratively improving AR systems [61]. Among the widely used evaluation technique in user studies are subjective measurements using questionnaires, user ratings, or judgments. For AR systems, researchers have used the System Usability Scale and the NASA Task Load Index for quantifying general system usability and workload, respectively. For handheld devices, the Mobile Phone Usability Questionnaire (MPUQ) enumerates the various questionnaires for common uses of mobile phones [155]. These questionnaires have been previously evaluated and studies support their validity and reliability. However, these standard questionnaires do not consider specific perceptual and ergonomic issues common to HAR systems. As such, researchers complement these evaluation tools with their own questionnaires. These questionnaires are not always tested for validity and reliability. Moreover, the questions tend to be specific to the features of researchers' HAR system.

In response to the lack of valid and reliable evaluation tools for HAR, I developed the HAR Usability Scale (HARUS) which is composed of two subquestionnaires, namely, the manipulability scale and the comprehensibility scale. I designed the questionnaires based on my analysis of HAR systems. I then evaluated the validity and reliability of HARUS in three experiments. I discuss in this chapter some insights gathered from using HARUS in user studies. Researchers and professionals involved in developing HAR applications can directly use my questionnaire to evaluate their own HAR applications. They can also modify the questionnaire with considerations of my evaluation framework and the insights presented in this chapter.

3.2. Related Work

The main points of evaluation in mobile AR are in its unique perceptual and ergonomic issues [174]. Previous mobile AR systems involved carrying computers with backpacks while wearing head-mounted displays and other peripherals. Currently, mobile AR can also be implemented in single handheld device. This new platform lead to new perceptual and ergonomic issues of AR that need to be evaluated in user studies [102].

Several researchers have studied perceptual issues in AR with respect to enabling devices, such as head-mounted display (HMD), handheld device, and projector-camera system. Kruijff et al. considered the human visual processing and interpretation pipeline in summarizing these perceptual issues [101]. They associated these perceptual issues to implementation issues found in the choice of real environment, capturing, augmentation, display, and individual user differences. Moreover, they described several issues and disadvantages arising from the form of handheld devices.

Handheld devices may refer to cellular phones, smartphones, tablet computers, ultra-mobile computers, etc. Currently, many handheld devices have powerful processors, large LCD screens, and built-in cameras. These features allow researchers to implement AR in one compact device. Although handheld devices are useful for many applications, Kruijff et al. listed the following disadvantages: less visibility of the LCD screen, lower fidelity, difference in disparity planes, higher latency, and smaller screen sizes compared to HMD and projector-camera systems [101]. Through my systematic literature review, I support these insights with perceptual issues reported by actual users when testing HAR applications.

Aside from perceptual issues unique to HAR, I also consider ergonomic issues specific to the behavior of people using HAR. Among the several interactions afforded by HAR, the most common use is the magnifying glass metaphor [153]. In this metaphor, the users hold the handheld device in front of them, the screen faces the user and the camera points to a scene. This kind of interaction is very different from conventional uses of handheld devices. As such, previous tools used for mobile devices such as the Mobile Phone Usability Questionnaire (MPUQ) [155] do not consider such interaction. Although there are some overlaps between MPUQ and HARUS, particularly in questions related to cognitive load

and control, standard questionnaires for mobile phones do not give emphasis on fatigue associated with the unique visualizations and gestures when using HAR.

Veas and Kruijff evaluated several handheld platforms to understand and address the ergonomic issues of HAR [182]. They describe HAR ergonomic issues to be an interplay of issues in pose, grip, controller allocation, weight, and size. The goal of their design is for the users to hold a particular pose while gripping the handheld device. Aside from viewing interactions, they also considered input interactions, such as having additional controllers. As expected, the size and weight of the device and the whole system are important considerations in HAR.

Given these two types of issues, it follows that the goal of design for HAR is to have no perceptual and ergonomics issues. In my evaluation framework, I refer to these qualities as *comprehensible* and *manipulable*, respectively. In other words, a perfect HAR application would score 100% on measures of comprehensibility and manipulability. I approximate HAR usability to be equivalent to a linear combination of comprehensibility and manipulability; i. e., I assume usability to be the average of these factors. I then provide evidence that these are sound estimations.

3.3. Approach

To create the HAR Usability Scale (HARUS), I first developed the evaluation framework which involves analyzing the background and conceptualizing the usability constructs. I then chose the format and data analysis and assessed the validity and reliability of HARUS, as recommended by Radhakrishna [150].

1. Analyzing the background – I conducted a systematic literature review to explore the common problems experienced by users when using HAR applications.

2. Conceptualizing the usability constructs – I defined two factors that I want to observe, namely, *comprehensibility* and *manipulability*. Comprehensibility is the ease of understanding the information presented by the HAR system. Whereas, manipulability is the ease of handling the HAR system

as the user performs the task. Comprehensibility and manipulability correspond to the perceptual and ergonomic issues in HAR, respectively. Thus, my framework assumes that the usability of a HAR system is approximated by comprehensibility and manipulability factors.

3. Choosing the format and data analysis – My questionnaire is patterned from the System Usability Scale (SUS) [107]. Moreover, it follows the questionnaire design rules prescribed by Fowler and Cosenza [55]. I designed the questionnaires to be answerable using Likert scales, similar to the SUS. In other words, I asked users to indicate how much they agree or disagree to the statement presented to them by rating a scale from 1 to 7. Only 1 and 7 are labeled, with 1 labeled as "Strongly Disagree" and 7 labeled as "Strongly Agree". I used a 7-point Likert scale because the audience of my experiments are sophisticated enough to distinguish subtle differences in these scales, as recommended by Krosnick and Presser [100]. I ordered the statements such that the positively-stated and negatively-stated statements are alternating.

4. Establishing validity - I assessed the validity of HARUS by looking at its concurrent validity, a kind of criterion-oriented validation procedure [34]. Concurrent validity is demonstrated when a questionnaire correlates well with objective measurements (time on task, etc.) or subjective measurements (SUS, etc.) that have been previously validated. As such, validity is a matter of degree, not all or nothing [129].

5. Establishing reliability - I measured the reliability or the precision of HARUS by computing the Cronbach's alpha – a measure of internal consistency of a questionnaire [100].

3.4. The Evaluation Framework

To create HARUS, I first conducted a systematic literature review of HAR systems to identify the common usability issues reported by AR researchers in their work. Based on this review, I define two usability constructs that affect HAR systems, namely, manipulability and comprehensibility.

3.4.1 Systematic Review of HAR Systems

I used the search string *handheld AND "augmented reality" AND evaluation* to search the ACM Digital Library for relevant research papers. This search resulted in 959 papers which I narrowed to 10 articles (column 2 of Table 3.1) by applying the following inclusion criteria:

1. Must discuss a HAR application
2. Must conduct a user study
3. Must be the latest article on that HAR application

I read the papers with focus on listing issues raised by users and issues observed by experimenters or expert reviewers. I listed these issues encountered by users in Table 3.1.

3.4.2 Definition of Manipulability and Comprehensibility

In my evaluation framework, I classify usability issues as either a manipulability issue or a comprehensibility issue. Manipulability is the ease of handling the HAR device. Manipulability issues refer to difficulties related to gripping, posing or operating the HAR device while the user performs a specific task. In the current literature, examples include:

1. Difficulty in holding the device in positions that allow the best tracking of the real environment and correct rendering of virtual objects.
2. Physical fatigue and stress to body parts often subject to the duration of use, size and weight of the handheld device, and required gestures.
3. Difficulty in entering information on the device.

Whereas, comprehensibility is the ease of understanding the information presented by the HAR system. Comprehensibility issues refer to the difficulties in understanding the novel visualization of HAR in a small screen space. In the current literature, examples include:

Table 3.1. User Issues in HAR Systems

References	Usability Issue
[105], [49], [187], [134], [142]	The tracking is unstable due to the ambient light, bad sensor fusion, or mishandling of the user.
[105], [49], [187], [134], [133]	The virtual objects are not well-registered.
[105]	The application is lagging or has intolerable latency.
[187], [133], [142]	The content was excessive and/or has poor quality.
[49], [142]	The display induces too much cognitive load.
[142]	The download time of the content is too slow.
[49], [161], [180], [46]	The screen is not legible due to outdoor ambient light.
[105], [180]	The screen is not legible due to reflection or glare.
[161], [46]	Depth is overestimated or underestimated.
[181] [161]	The application causes fatigue after extended use.
[181], [180]	The device is too bulky or too heavy.
[181], [180]	Hand interactions are difficult to perform.
[105], [142]	The application is not responsive and/or provides no feedback.
[180]	The keypad is too small.

1. Difficulties in understanding the information arising from imperfections in tracking and rendering, such as imprecise or unstable three-dimensional registration and underestimation or overestimation of depth.

2. Cognitive difficulties arising from excessive content, lagging display, and unresponsiveness of the application.

3. Difficulties related to doing multiple tasks (etc. walking while navigating) and environmental factors, such as ambient light and glare on the screen.

3.5. The Usability Scale

To evaluate the manipulability and comprehensibility of a HAR system, I developed the manipulability scale and the comprehensibility scale which can be used as separate questionnaires. The statements in the HARUS are derived from the issues listed in Table 3.1. When combined, these questionnaires form the HARUS, as shown in Table 3.2.

I score the HARUS similar to the SUS [13]. I designed it to have a two-factor structure representing comprehensibility and manipulability. For each factor, multiple questions are asked to help the users evaluate various aspects contributing to their experience with the HAR system. The HARUS is intended to measure the usability of a HAR system given a target user group and a confined task.

The HARUS is composed of 16 statements (Table 3.2) that roughly correspond to commonly encountered problems in HAR applications. Users were then asked to rate their agreement by using a 7-point Likert scale. To compute the HARUS score, I apply a similar method for computing the SUS score [13]:

1. For the positively-stated items, subtract one from the user response. For the negatively-stated items, subtract the user response from seven.

2. Add all these converted responses.

3. Divide the sum by 0.96 to have a score ranging from 0 to 100.

The HARUS has a two-factor structure. Statements 1 to 8 are measures of manipulability, whereas statements 9 to 16 are measures of comprehensibility.

Table 3.2. The HAR Usability Scale

	Manipulability Scale:
1	I think that interacting with this application requires a lot of body muscle effort.
2	I felt that using the application was comfortable for my arms and hands.
3	I found the device difficult to hold while operating the application.
4	I found it easy to input information through the application.
5	I felt that my arm or hand became tired after using the application.
6	I think the application is easy to control.
7	I felt that I was losing grip and dropping the device at some point.
8	I think the operation of this application is simple and uncomplicated.
	Comprehensibility Scale:
9	I think that interacting with this application requires a lot of mental effort.
10	I thought the amount of information displayed on screen was appropriate.
11	I thought that the information displayed on screen was difficult to read.
12	I felt that the information display was responding fast enough.
13	I thought that the information displayed on screen was confusing.
14	I thought the words and symbols on screen were easy to read.
15	I felt that the display was flickering too much.
16	I thought that the information displayed on screen was consistent.

It is important to note that these statements are not exhaustive [128]. Rather, they are measures belonging to an extensible set of indicators for manipulability and comprehensibility. Similarly, I do not claim that these 16 questions and two constructs are the strict operationism of usability in HAR. They are measures belonging to an extensible set of indicators for usability. However, I showed evidence in the succeeding three experiments that this set of measures is a good approximation of usability in HAR.

The SUS [107] is composed of 10 statements that breaks the question "Is this system easy to use?" into several aspects of the system. Similarly, the concept of HARUS is to break down the questions "Is this system easy to handle?" and "Is the information presented easy to understand?" so that the users find it easier to give their feedback. When users are asked general questions like "Is this system easy to use?", they would not know how to weigh various aspects of the system to come up with a single rating. They can give better feedback if they can rate smaller, more specific aspects. These ratings can then be accumulated to gauge their answer to the bigger, general questions.

There are many areas of HAR applications, including advertising, navigation, work support, scientific visualization, etc. Some may argue that the main factors affecting usability will vary according to the application area. Some may say that the purpose of the HAR application is different, thus the requirements are different. I offer two arguments why HARUS can be used to all application areas.

First, HARUS is not intended to give an overall evaluation of a HAR application. Rather, it evaluates the suitability of HAR application to target users and tasks. Usability evaluations are always with respect to the user and the task [135]. For example, a sophisticated HAR application for work support might have a lower HARUS score than a crude HAR application for advertising because the tasks in work support are more difficult. This is fair because it is possible to create a crude application that addresses the needs of a user for a specific task and it is also possible to create a sophisticated application that does not address the needs of a user for some other task. The evaluation is always relative to the user groups and the tasks.

Second, I applied the best effort because I considered issues in as much application areas as possible. I based HARUS only on known issues because I cannot

predict future issues that will arise in new application areas. These known issues will still be the problem in HAR applications in the next coming years. Furthermore, I applied multiple experiments with multiple benchmarks including both objective and subjective measures of usability. The experiments are both practical and general.

3.6. Validity and Reliability of HARUS

I evaluated the validity and reliability of HARUS in three experiments. The tasks are annotating text, status reporting and positioning arrows.

3.6.1 Experiment 1: Annotating Text

In this experiment, users evaluated an application for annotating text on real objects found in the environment.

Some HAR applications aim to create AR content in situ. In the work of Langlotz et al. [104], the users create virtual content directly onto the environment by using only a smartphone. The usability of such an authoring system can be evaluated using the SUS [107], although some information considering perceptual and ergonomics issues are lost. Time on task can also be used to evaluate this system because people who will find the application difficult to use would tend to spend more time.

This experiment tests the ability of HARUS in evaluating a simple HAR content authoring tool. I evaluated the concurrent validity of HARUS by checking its correlation with SUS, a previously validated subjective measure. Furthermore, I benchmarked against time on task, an objective measure. My hypotheses are as follows:

H1. HARUS and SUS have a positive relationship.

H2. HARUS and time on task have a negative relationship.

3.6.1.1 Experimental Platform

I implemented a simple HAR authoring tool for annotating text on real objects (Figure 3.1, right). I used the PointCloud SDK[1] to detect some natural feature points in the environment. To register feature points, the user must move the handheld device from side-to-side (Figure 3.1, left). Once the system detects enough feature points, the user can add a text label onto the scene.

The application runs on iPad 2 tablets (A5 processor, 512MB DDR2 RAM, 32GB, 601 grams). I used the back camera (640x480 pixels, 30 fps) for sensing, and 9.7 inch LED-display (1024x768 at 132 ppi) for display. The interface was built using standard interface elements of iOS 6 such as labels, textfields, keyboard, etc., as shown in Figure 3.1 (middle).

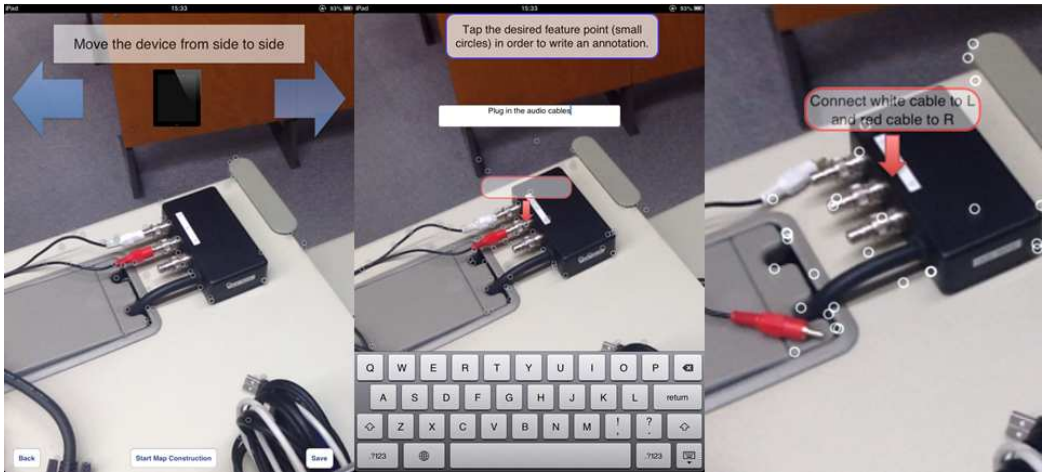

Figure 3.1. Simple HAR Authoring Tool for Annotating Text

3.6.1.2 Design and Procedure

Eighteen voluntary participants with ages ranging from 22 to 41 years old (M=27, SD=4.0) participated in this experiment. First, the experimenters demonstrated how to use the authoring tool. The participants were then asked to annotate English translations on a rice cooker and annotate trivia on a paper bill (Figure 3.2). No time limit was given to do the tasks and the participants were free to give

[1]http://developer.pointcloud.io/sdk

up. I offered this option because I found out in a pilot study that some people fail to do the registration procedure. After finishing the task or giving up, the participants answered the SUS and the HARUS questionnaires. Nine answered the SUS first, whereas nine answered the HARUS first. I took note of the time on task and I calculated the HARUS and SUS scores, as described in Section 3.5.

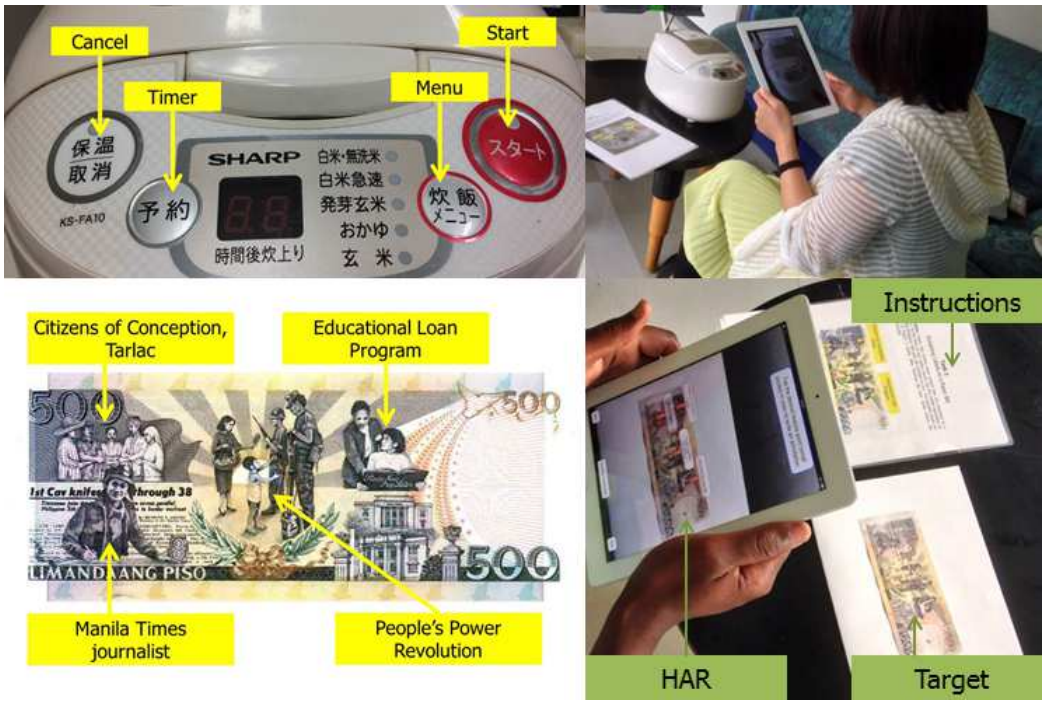

Figure 3.2. Authoring Tasks

3.6.1.3 Results of Validity

Fifteen participants finished the task with an average time of 8.1 minutes (SD=2.5). The 18 participants gave the HAR authoring tool an average SUS score of 62 (SD=22) and an average HARUS score of 65 (SD=16). These scores are below the acceptable SUS score of 70 and above [13].

The HARUS score has a very strong positive relationship with the SUS score and a strong negative relationship with time on task (Table 3.3). Both findings are significant, thereby supporting hypotheses 1 and 2. These findings are indicative of the concurrent validity of HARUS.

Table 3.3. Correlations (r) of HARUS, SUS and Time on Task

	1	2	3
1. HARUS	1.00		
2. SUS	0.87***	1.00	
3. Time on Task	-0.51*	-0.59**	1.00

* significant at 0.05 level
** significant at 0.01 level
*** significant at 0.001 level

3.6.2 Experiment 2: Status Reporting

In this experiment, users evaluated an application for viewing virtual notes on real devices for writing a report.

HAR applications commonly require users to read virtual information associated with real environments. This information could be an advertisement, scientific data, historical information, etc. Aside from the SUS, this kind of commercial application can be evaluated based on the "Affective Aspects and Media Properties" (AAMP) construct of the MPUQ [155], such that, an easy to use product would elicit positive emotional responses. Of the 14 statements measuring AAMP, I chose 8 that were relevant to the task.

Furthermore, in work-related tasks, a useful HAR should lead to better work output. This experiment also checked the relationship of the HARUS with the verbosity of the status report. I assume that writing more words on the report means that the report is more comprehensive and is thus of better quality. My hypotheses are:

H3. HARUS and SUS have a positive relationship.

H4. HARUS and AAMP have a positive relationship.

H5. HARUS and report verbosity have a positive relationship.

3.6.2.1 Experimental Platform

The HAR application enables users to view text annotations on real objects (Figure 3.3, left and middle). The application runs on iPad mini tablets (A7 processor, 512MB DDR2 RAM, 16GB, 308 grams). The back camera (640x480 pixels, 30 fps) is used for sensing, and a 7.9 inch LED-display (1024-by-768 at 163 ppi). The application uses the PointCloud SDK for tracking and the standard user interface elements of iOS 7 for the display.

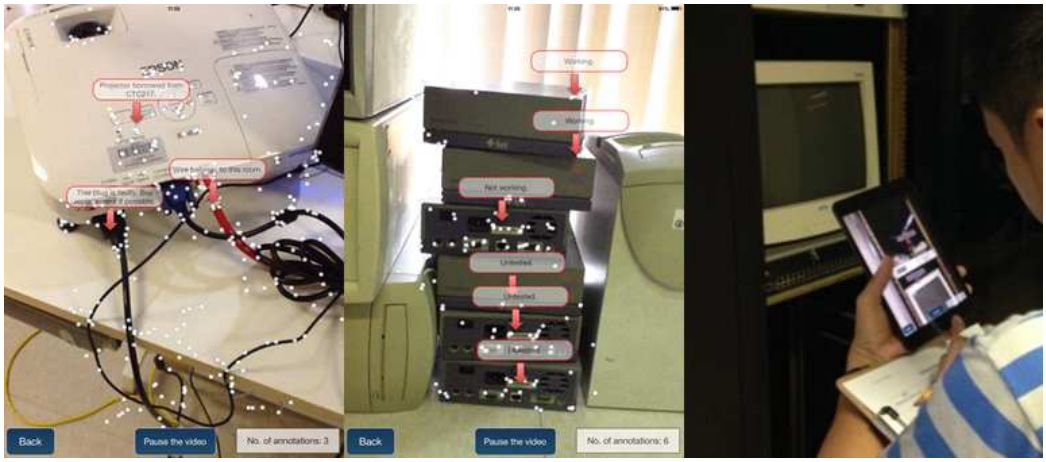

Figure 3.3. HAR for Viewing Annotations on Equipment

3.6.2.2 Design and Procedure

Twenty voluntary participants with ages ranging from 19 to 46 years (M=28, SD=8.1) participated in this experiment. Before performing the task, I explained AR and its enabling technologies to the participants by using videos and slides. I then demonstrated how to use the HAR application for viewing text annotations. The participants assumed the role of a newly-hired maintenance staff. Their first job is to report on the status of equipment by viewing the annotations made by the previous maintenance staff. They then filled the report form that has three columns – device, description of issue, and recommended action. To make the report, the participants need to gather information from the HAR and the devices, such as model, serial numbers, brand, etc. I gave them a time limit of 15 minutes to finish the task. I finally asked them to answer three questionnaires:

HARUS, SUS, and AAMP. I computed the AAMP score similar to the method for SUS.

3.6.2.3 Results of Validity

This kind of work-support task is not limited to those that use head-mounted displays. Several researchers apply HAR because it is less intimidating and easier to share, thereby facilitating collaboration with co-workers [161]. This task is suitable for HAR because writing the report requires both information displayed by the HAR system and information gathered from the real environment such as the description of the device. The natural interaction pattern I observed is as follows: First, the participants find a suitable angle that would reveal the virtual information. They then freeze the screen and settle to a more relaxed pose. Lastly, they switch between reading the screen and inspecting the device when writing the report.

Only one participant was not able to finish the report under 15 minutes. The rest were able to finish the report with an average time of 9.9 minutes (SD=1.9). The participants made reports consisting of an average of 73.5 words (SD=19.5) about 13 individual devices. They gave the HAR an average SUS score of 80 (SD=11) which is an acceptable SUS score. The average HARUS and AMMP scores were 74 (SD=13) and 80 (SD=13), respectively.

Table 3.4. Correlations (r) of HARUS, SUS, AAMP and Verbosity

	1	2	3	4
1. HARUS	1.00			
2. SUS	0.79^{***}	1.00		
3. AAMP	0.75^{***}	0.82^{***}	1.00	
4. Verbosity	0.12	0.23	0.43^{*}	1.00

* significant at 0.05 level

*** significant at 0.001 level

The HARUS scores have a very strong positive relationship with the SUS and AAMP scores (Table 3.4). Both results are significant, thereby supporting hypotheses 3 and 4. These results are indicative of the concurrent validity of HARUS. I did not find any significant relationship between HARUS and verbosity probably because having low word count could mean both lacking in information (bad quality) or simply concise (good quality).

3.6.3 Experiment 3: Positioning Arrows

Positioning virtual objects is one of the most important tasks in authoring AR contents. Currently, HAR has no established interaction metaphors thus various methods of doing specific tasks need to be evaluated. One such task is adjusting the 3D position of a virtual object in the real environment. For this experiment, I implemented a device-centric method similar to the work of Anders Henrysson and colleagues [66]. When the user selects the virtual object on the screen, the position of the virtual object becomes fixed relative to the movement of the device. As such, the user can drag the virtual object by moving the handheld device in any direction.

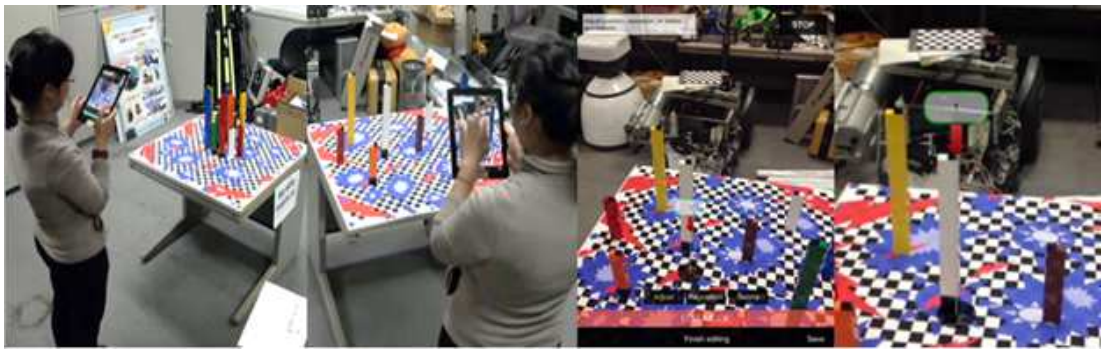

Figure 3.4. Adjusting Arrows to Target Pillars with Different Heights

3.6.3.1 Design and Procedure

I asked the participants to position arrows on top of each of the eight pillars in Figure 3.4. They did this twice, once with the pillars concentrated in the center (Figure 3.4, left most) and once with the pillars farther apart (Figure 3.4, second

from left). After the task, I asked all 23 of them to answer HARUS. Of the 23, I only asked 7 to answer the SUS to save time. I took note of the time on task and I measured the distance of each arrow to the target pillar as the positioning error. In this experiment, I focused on comparing HARUS with time on task and total positioning error with the following hypotheses:

H6. HARUS and SUS have a positive relationship.

H7. HARUS and time on task have a negative relationship.

H8. HARUS and positioning error have a negative relationship.

3.6.3.2 Results of Validity

The participants spent an average time of 18.2 minutes (SD=7.8) on the task with an average total positioning error of 22.6 mm (SD=9.3). On the average, the participants gave the application an SUS score of 57 (n=7, SD=19) and they rated the application 58 (SD=15) on the HARUS. Based on seven participants, HARUS has a very strong positive relationships with the SUS (Table 3.5). Furthermore, based on 23 participants in Table 3.5, HARUS has a strong negative relationship with the total amount of error but not with time on task. In other words, participants who were less accurate with positioning the arrows tend to give lower usability scores to the application. For this experiment, I found evidence supporting hypotheses 6 and 8, but not 7.

3.6.4 Reliability of HARUS

After showing evidence of the validity of HARUS, the last step in developing an evaluation tool is to measure its reliability. I computed the Cronbach's alpha to assess the internal consistency of my questionnaire based on the responses in the experiments. All of the alphas are between 0.7 to 0.9. Thus, the HARUS has good internal consistency, as shown in Table 3.6.

Table 3.5. Correlations (r) of HARUS, SUS, Time on Task and Positioning Error

	1	2	3	4
1. HARUS	1.00			
2. SUS	0.90**	1.00		
3. Time on Task	0.87***	-0.10	1.00	
4. Positioning Error	-0.51*	-0.35	-0.59**	1.00

* significant at 0.05 level
** significant at 0.01 level
*** significant at 0.001 level

Table 3.6. Cronbach's Alpha (α) in Three Experiments

Experiment	α	Interpretation
Annotating Text	0.83	Good
Status Reporting	0.83	Good
Positioning Arrows	0.88	Good

3.6.5 Evaluation of Manipulability and Comprehensibility Scales

Through the three experiments, I showed the validity and reliability of HARUS. I demonstrated concurrent validity by providing evidence supporting six out of the eight hypotheses. In all three experiments, the participants were able to answer the questions consistently as measured by the Cronbach's alpha. Aside from these main findings, I explored some interesting relationships between the factors of HARUS and other variables in my experiment.

The HARUS can be decomposed into two scores: the manipulability score based on questions 1 to 8, and the comprehensibility score based on questions 9 to 16 (Table 3.2), which was computed similarly as the HARUS score (Section 5).

In previous user studies of HAR applications, perceptual and ergonomic issues are described to be interrelated in some user issues mainly because the manner of handling the device affects the quality of visualization. Moreover, instability in tracking makes the tasks longer. As such, participants report fatigue specifically in the arms and hands.

In my experiments, I observed manipulability and comprehensibility to be interrelated but moderate enough to be used as different scales. Table 3.7 summarizes the correlations of these two factors of HAR usability. I only observed a strong positive relationship in the positioning arrows task. For both annotating text and status-reporting tasks, the correlations were moderate and not significant. As such, my guess is that manipulability and comprehensibility in HAR depend on the target users and on the tasks. Therefore, these two factors must be observed independently from each other. The HARUS is suitable for this observation.

Table 3.7. Correlations (r) of Manipulability and Comprehensibility in Three Experiments

Experiment	Pearson's r	Interpretation
Annotating Text	0.40	moderate
Status Reporting	0.34	moderate
Positioning Arrows	0.61**	strong

** significant at 0.01 level

Scoring the manipulability and comprehensibility factors of the HARUS reveals additional insights from the three experiments. In all three experiments, comprehensibility has a stronger positive relationship with SUS compared to manipulability. The difference was small for the authoring scenario (Table 3.8). However, the difference was pronounced for viewing text (Table 3.9) most probably because of the nature of the task. In the tasks for authoring text and positioning arrows, manipulability was very important to the whole usability of the interface because it required the users to do difficult hand movements such as

moving the device, positioning information in 3D, and typing some text. On the other hand, these input interactions are not considered in the status reporting tasks. The focus of these tasks was to understand the information presented to the user.

Table 3.8. Correlations (r) of HARUS Factors, SUS, and Time on Task in Annotating Text Scenario

	1	2	3	4
1. Manipulability	1.00			
2. Comprehensibility	0.40	1.00		
3. SUS	0.72^{***}	0.75^{***}	1.00	
4. Time on Task	-0.41^{*}	-0.45^{*}	-0.59^{**}	1.00

* significant at 0.05 level
** significant at 0.01 level
*** significant at 0.001 level

In the status-reporting task, I did not find strong positive relationship between the HARUS score and the verbosity of the reports. However, a strong positive relationship exists between the manipulability score and verbosity. In other words, people who found the HAR easy to handle tend to write more words on their report. I find this to be logical and I think that there are trade-offs in user performance for activities that use the hands (e.g. handling the HAR and hand-writing a report). I plan to investigate this more in my next experiments.

Although consistent with the results of HARUS, I observed differences in the strength of correlations in the two factors for positioning arrows. Manipulability has a strong negative relationship with positioning error compared to comprehensibility. In other words, those who found the HAR device easy to handle tend to make less error. Intuitively, one would guess that the usability of a device-centric positioning interface would be affected more by the manipulability of the device rather than comprehensibility.

Given that the manipulability and comprehensibility statements can be used

Table 3.9. Correlations (r) of HARUS Factors, SUS, AAMP, and Verbosity in Status Reporting Scenario

	1	2	3	4	5
1. Manipulability	1.00				
2. Comprehensibility	0.34	1.00			
3. SUS	0.58**	0.70***	1.00		
4. AAMP	0.54*	0.68***	0.82***	1.00	
5. Verbosity	0.41*	-0.19	0.23	0.43*	1.00

* significant at 0.05 level
** significant at 0.01 level
*** significant at 0.001 level

Table 3.10. Correlations (r) of Comprehensibility, Manipulability, SUS, Time on Task, and Total Positioning Error in Positioning Arrows Scenario

	1	2	3	4	5
1. Manipulability	1.00				
2. Comprehensibility	0.61**	1.00			
3. SUS	0.83*	0.84*	1.00		
4. Time on Task	-0.07	-0.03	-0.10	1.00	
5. Positioning Error	-0.62**	-0.49*	-0.35	-0.21	1.00

* significant at 0.05 level
** significant at 0.01 level

as separate scales, I evaluated the internal consistency of the responses in three experiments. Although the Cronbach's alphas are slightly lower than those of HARUS, both manipulability and comprehensibility have good internal consis-

tency, as shown in Table 3.11. Thus, these two HARUS factors can be used as separate scales in cases wherein researchers are only interested to measure these factors.

Table 3.11. Cronbach's Alpha of Manipulability and Comprehensibility in Three Experiments

Experiment		α	Interpretation
Annotating Text	Manipulability	0.71	Good
	Comprehensibility	0.74	Good
Status Reporting	Manipulability	0.81	Good
	Comprehensibility	0.80	Good
Positioning Arrows	Manipulability	0.81	Good
	Comprehensibility	0.82	Good

3.7. Chapter Summary

HAR is novel interface that has a high potential for becoming a mainstream technology. It is useful for delivering various content in many fields of application, such as education. The development of new interaction metaphors and HAR systems must also be accompanied with the development of new evaluation tools and frameworks. Valid and reliable questionnaires are important for conducting user studies to iteratively improve HAR systems.

I designed HARUS with sub-questionnaires, comprehensibility scale and manipulability scale, based on ergonomic and perceptual issues of HAR. This approach is advantageous because there are cases wherein standard questionnaires like the SUS do not capture the unique issues in HAR. Moreover, distinguishing between perceptual and ergonomic issues reveals that comprehensibility and manipulability are separate constructs. The usability of a system can suffer more from one of these two separate constructs, therefore, efforts in improving on one could significantly improve the whole system.

I showed the validity and reliability of HARUS in three experiments. My experimental scenarios arise from my own interest in using HAR for real world annotation in the near-field. As such, my experiments are not exhaustive of various HAR scenarios. In particular, I do not have experiments of HAR systems for the far-field. As such, it would be interesting to see if HARUS is also valid and reliable for other specific cases wherein HAR is commonly applied.

HARUS is a tool for evaluating HAR applications with users as they perform specific tasks. HARUS aggregates usability, comprehensibility and manipulability into a single score. This score can be used by researchers and professionals to compare between iterations of the same system, to prioritize among several features of the same system, and to benchmark against previously evaluated implementations of a HAR system. In Chapter 4, I used HARUS to evaluate my own HAR system.

CHAPTER 4

Evaluations of Situated Multimedia

Augmented reality (AR) is useful for presenting information situated in real environments. However, there are few research works exploring the design and evaluation of AR for learning support. In this chapter, I treat AR as a type of multimedia that is situated in real environments; i. e., I use multimedia learning theory as a framework. I discuss my experiences in developing a HAR learning support system for one specific use case – situated vocabulary learning. As part of my summative evaluation, I use the HARUS discussed in Chapter 3 as well as other measures of usability, memorization, and motivation. Results suggest that AR may possibly be better for the retention of words and it may possibly improve student attention and satisfaction.

4.1. Background

AR is the seamless integration of virtual objects and real environments [10]. In AR, computer-generated information is placed in the world as if they co-exist with real objects. It is an emerging technology that is finding applications in education because of its possible benefits to teaching and learning [190]. However, AR's practical uses are relatively not well-undertood compared to those of virtual

reality and other technologies [81] because few research work has been conducted to substantiate its benefits to learning [72].

AR is useful for presenting the explicit relationship of virtual contents to objects found in the real world. Researchers have shown some evidence that presenting digital information together with the context of a real environment helps memorization [57], [56]. They argue that AR has the potential to ease cognitive load and that using AR allows users to form memory retrieval cues based on the real environment. Although, these findings have not been tested on learning scenarios, it points to the possible advantage of using AR as the presentation method for situated learning. Currently, handheld devices like smartphones are already equipped with cameras and other sensors, enough processing power, and large screens for delivering AR learning experiences [16]. For example, Kamarainen, et al. [88] used a HAR system to support a fieldtrip in a local pond.

As of the time of this writing, though, there has been little empirical evidence collected to substantiate or refute AR's potential as a usable carrier of educational content. In my review in Chapter 2, I found only seven research articles reporting evidence of AR's effectiveness in improving learning outcomes. I observed that AR's impact on learning outcomes vary from a small negative effect to a large postive effect. There are many factors attributed to this variation, such as the comparison being made and the appropriate matching of the technology to pedagogical needs. However, even with the current state of AR, researchers already report that AR has positive effects on motivational factors of attention and confidence [165].

AR is a good match for teaching culture and languages because it can be used for presenting information relevant to places [116], [115]. In this research, I limit language learning to vocabulary learning as the target of my HAR system. I based the requirements of my system on multimedia learning theory, previous vocabulary learning systems, and teacher's feedback on AR. Because AR is a kind of multimedia that is situated in an authentic environment, multimedia learning theory [127], [126] can be applied for designing and evaluating AR's benefits to learning. After implementing the system, I conducted system usability evaluations using the System Usability Scale (SUS) and my HAR Usability Scale (HARUS). In my investigation, I reiterated some guidelines for applying AR to

education, as well as added my own design goals. Finally, I evaluated student learning outcomes and student motivation with my application.

The goal of this chapter is three-fold: I would like to develop an AR application, test its usability, and test its effects on learning. To these ends, I demonstrate my development and evaluation framework for prototyping AR learning experiences. I apply AR to the task of memorizing vocabulary words. I then test ARs effectiveness as a platform for a memorization task and examine its impact on student motivation.

4.2. Related Work

The general public is becoming more familiar with AR mainly because of AR browsers used for conveying a variety of location-based information [64]. Currently, people use AR browsers to see virtual annotations integrated with a live video feed of the real environment. This integration promotes easier understanding of location-related information, such as names of buildings, distances of restaurants, and arrows for navigation [56]. In the case of situated vocabulary learning, instead of displaying names and direction, I designed a system that displays words and animations to teach new vocabulary words that are relevant to the objects inside the environment.

4.2.1 Vocabulary Learning Systems

Mastering a foreign language relies heavily on building vocabulary necessary for listening, reading, speaking, and writing [193]. Several creative approaches have been developed to support such vocabulary learning, including hypertext annotations in e-learning [29], collaborative multimedia [80], word games [114], virtual environments [144], and interactions with robots [189]. The instructional designs for these prototypes leverage three main strategies, namely, repetition, engagement, and context. Acquiring new words requires repeated exposure to those words [185]. This includes both memory rehearsal (e.g. pronouncing the words several times) and spaced exposures (e. g. encountering the words several times in conversations) [42].

Several sophisticated systems have been developed in order to support context-awareness in learning [139], [30], [145]. Context is important to vocabulary learning because students can use it to form stronger associations between the new word and objects in the real world [138]. In vocabulary learning, context can take many forms. Researchers have used personalized learning systems that tailor-fit the vocabulary content to students' internal context, i.e. their current level of competence [192]. Researchers have also built vocabulary applications that leverage external contexts, such as studying in a library or eating in the cafeteria [164].

Situated cognition argues that knowledge cannot be abstracted from the situation from which it was learned. Learning is always embedded in the activity, context, and culture from which the knowledge was developed [20]. Learning vocabulary words from dictionary definitions and a few sample sentences is inferior to conversations and meaningful bodies of text. Words that students find useful and words they actually use have better chances of getting acquired. Systems for situated vocabulary learning take advantage of situated cognition by selecting words that are associated with the environment and teaching only the words that are useful. Researchers are taking advantage of near-transfer or applying the knowledge learned in a specific situation to an almost similar context [48]. In situated vocabulary learning, the words are learned in the context of its use thus facilitating knowledge transfer. Moreover, it encourages the students by illustrating the relevance of the vocabulary words.

Language is always situated in activities that are bound to an environment with its accompanying physical, social, and cultural aspects. In two case studies, Wong and Looi [188] asked students to take pictures that illustrate English prepositions and Chinese idioms. For nine weeks, students used mobile phones to take pictures in school and at home. They then annotated the pictures with sentences. These sentences were shared and revised with classmates making the activity collaborative. In their study with 40 students, they have gathered 481 photo-sentence pairs, 124 revisions, and 134 comments. Although the students enjoyed the activity, they observed that there is a wide variability in student participation. Students contributed an average of 12.0 (SD=25.9) pictures, and each offered the revision of 3.1 (SD=7.3) sentences.

Researchers explain that ubiquitous, context-aware systems are useful for providing the necessary situated cognition [20] to language learning. To provide location-aware systems, researchers have described wireless positioning techniques and content distribution using the WLAN within their campus [69], [4], [53]. Using the campus WLAN and WCDMA, Liu [116] provided the content for HELLO, an English language learning system. The system detects the location of the user using QR codes spread around the school. At each location, students practiced conversations with a virtual learning tutor. In their user testing with 64 students, they report that the students who used the situated language learning approach scored higher (M=89.4, SD=7.5) compared to those that used printed materials and audio recordings (M=81.3, SD=9.6). This large effect size (d=1.0) is attributed to practicing English in real-life situations, as well as, encouraging the creative abilities of the students in handling conversations.

Instead of using WLAN positioning techniques and QR codes, Edge et al. [51] took advantage of the sub-categories of Foursquare as the classification of the establishment the user is currently in. They then generated the vocabulary words that are frequently associated with that establishment. Users study these vocabulary words via a mobile application called MicroMandarin. For four weeks, 23 participants used their system to learn Chinese vocabulary words in establishments in Shanghai and Beijing. Of all the participants, 68% felt that the detection of their location was okay to great, and 91% found that the vocabulary content was okay to great.

Similar to MicroMandarin, Vocabulary Wallpaper [39] is a microlearning mobile application that takes advantage of idle times that people spend waiting in different locations. Dearman and Truong [39] prototyped the Vocabulary Wallpaper for casual learning of Italian in three types of establishments within the vicinity of their university. Using GPS or network positioning, Vocabulary Wallpaper determines which of the predefined establishment the user is in. The researchers tested the application with 16 participants using it in four sessions. The results show that the participants can recall an average of 23.3 (SD=17.1) words, and recognize an average of 39.5 (SD=19.3) words out of all the 75 words. Interestingly, the participants significantly (p¡0.05) gained more situated words (M=9.27, SD=6.44; M=7.33, SD=5.68) than words that were designed to appear

more frequently (M=6.73, SD=6.17).

Aside from presenting information related to the user's current environment, TANGO used RFID to tag the objects in the environment to present vocabulary words relevant some objects found in that environment. They equipped a PDA with an RFID readerto scan the environment. Users are presented with a question through the PDA and they answer by tapping their PDAs to the correct object. They evaluated the usability of TANGO in two user studies. In the first user study with six students [140], TANGO has a perceived ease of use of 3.3/5 (SD=1.0) and a perceived usefulness of 4.2/5 (SD=0.4). In the second user study with 16 students [141], TANGO improved its perceived ease of use at 4.3/5 and perceived usefulness to 4.7/5.

Beaudin et al. [15] took TANGO to the next level by detecting more user interactions with objects inside a house. Aside from tagging objects with RFID, they use three more sensors – switches for opening and closing cabinets, water flow detectors for the plumbing system, and piezo-triggered accelerometers to detect movements of objects. Overall, they tagged over 100 objects inside the house with 400 Spanish phrases. The system identifies the users through their mobile phones. When they use a particular object (e.g. open a door, sit on a sofa), the system plays the relevant English word and its Spanish translation. If they want to browse previously encountered content, they can access the phrases through their mobile phones. They asked a couple to use the system for 10 weeks. On the average, the phrases where presented 57 times per hour. However, even at this intense interaction, the couple found it acceptable even for extended use. The male participant recalled 158 of the 274 phrases he encountered, and he correctly guessed 65 out of 126 phrases that were not presented to him. The female participant recalled 79 of the 178 phrases presented to her, and she guessed correctly 26 of the 92 phrases that were not presented to her.

My idea is to use AR for situated vocabulary learning. The most important feature of situated vocabulary learning is the presentation of useful vocabulary words relevant to the current environment. Based on the ARCS model [94], relevance is one of the four factors to consider in creating motivating instructional materials. ARCS stands for attention, relevance, confidence, and satisfaction which are the factors contributing to motivation in using learning materials.

Among Keller's suggestions is relating new information to something the student is familiar with. In my case, I relate words with a familiar environment.

Existing applications can already deliver the relevant and useful information. However, the visualization of information remains on the mobile phone screen. The users are expected to find the relationship of the vocabulary to their surroundings (e.g. by looking around). This relationship is not always obvious. Using AR, I improved the presentation method by annotating real objects with sound, text, images, and animations that are three-dimensionally registered onto the environment. This kind of visualization is beneficial to situated vocabulary learning because it explicitly illustrates the relationship of the vocabulary with the objects found in the current environment.

4.2.2 Multimedia Learning on AR Annotation

In multimedia learning theory, multimedia refers to pictures and words (both written and spoken). It has three assumptions, namely, dual-channels, limited capacity, and active processing. First, humans have two separate channels for perceiving visual and auditory information. Second, individuals can only attend to a limited amount of information at any given time. Lastly, learning only takes place if the learner actively processes incoming information by connecting it to prior knowledge. Multimedia learning identifies five cognitive processes [127], [126] in learning:

1 Selecting words

2 Selecting images

3 Organizing selected words

4 Organizing selected images

5 Integrating incoming information with prior knowledge

Situated vocabulary learning leverages the prior knowledge of places, thereby promoting better learning experiences. Visualizing the information in context-rich environments using AR annotation can aid students in creating meaningful

associations between the content and the real environment. This promotes having a more elaborated knowledge and having more memory retrieval cues. Situated multimedia aids in the cognitive process of integrating incoming information with prior knowledge. This argument is consistent with the findings of Fujimoto et al. [56], [57]. However, AR annotation is also prone to presenting too much information and too much context from the environment leading to cluttered displays [145], [63]. This problem arises because the environment cannot be controlled by the author of the content. Whereas, all other types of multimedia (books, computer applications, virtual environments, etc.) give authors full control of the content. For example, they can make an illustration as abstract or as contextualized as they like by removing or adding details. In the case of AR, the environment is a given and authors of AR learning contents must make use of the environment creatively.

Cluttered displays hamper the cognitive processes of selecting and organizing. As such, in order to benefit from AR visualization, we need to make sure that we design against visual clutter for the HAR application. I can confirm if I am successful or not with the design by conducting usability evaluations [61]. To conduct usability evaluations, I use a general system usability questionnaire called the System Usability Scale or SUS [107]. Another useful tool is my Handheld Augmented Reality Usability Scale (HARUS) discussed in Chapter 3. HARUS has a comprehensibility component which measures the ease of understanding of an AR visualization.

Given that individuals have a limited capacity of information to which they can attend to, Lin and Yu [113] investigated the cognitive load induced by different types of media presentations on a mobile phone. In their study with 32 eight graders, they investigated the use of four multimedia modes, namely, text, text with audio, text with picture, and text with audio and picture. They discovered that the multimedia mode does not have a significant effect on vocabulary gain and retention. However, the learners rated the combined text-audio-picture as the mode that induced the least cognitive load.

Lin and Wu [112] investigated the use of these four multimedia modes in a succeeding study with 423 junior high school students. They did not find any significant differences in vocabulary recognition nor in any interaction between

multimedia modes and learning style preferences of the student. However, the group who used text with audio and picture performed best in listening tests, followed by the group who used text with sound. This result confirms the intuition that audio annotations contribute to the construction of phonological knowledge of words and to the application of this knowledge in listening to sentences. More importantly, results show that the learning effects of the audio were maintained for two weeks with minimal attrition. Based on these works, I implemented features in my AR system that allow users to access text, audio, and pictures.

In a separate study with 121 senior high school students, Lin and Hsiao [111] studied the effects of the use of still images against simple animations in vocabulary learning. Their results showed that the animation group performed significantly better in learning Chinese and English vocabulary words compared with the image group. Thus, recommend the use of animations to illustrate verbs and processes. To facilitate better understanding of vocabulary in my HAR system, I include a renderer for sprite sheet animations. I found this feature to be a simple solution to illustrate verbs.

4.2.3 Practical Considerations in Applying AR

Aside from providing evidence of benefits in the learning process, AR must also adhere to some practical considerations in order to adopt them in schools. Cuendet et al. [35] shares five design principles for adopting AR for classroom use. The five design principles are integrating AR to other class activities, empowering the teacher, providing the teacher awareness of the state of students, flexibility to adapt the activities to evolving scenarios, and minimizing functionalities to what is required at a given time.

Based on a survey with teachers and student in Malaysia, Sumadio and Rambli [173] observed that although most of them experienced AR for the first time, they perceived that the demonstrations presented to them are useful for educational practice. The prototype they showed was an AR learning experience for physics experimentation on heat absorption. Teachers and students expressed that bringing in AR to educational use would make the learning process more enjoyable. The other perceived benefits are better visualization and being able to simulate an experiment before the actual one. From this example, the partic-

ipants suggest that it is better to improve the realism of the virtual objects and expand the prototype to cover other experiments that are within the Malaysian physics curriculum.

Based on my interviews with teachers in the Philippines, AR is perceived to be useful because it offers learning by experiencing some activity that cannot be done now in the classroom. (I discuss this further in Chapter 5.) Although more conventional mediums of instruction will always remain relevant, the teachers would like to take advantage of various technological interventions to connect with their students. Currently, the teachers are interested in using AR to motivate class participation and to hold the attention of students. This sentiment echos the "empowerment" design principle of Cuendet et al. [35] which states that the teacher should remain the central point of class interaction.

However, the teachers also expressed their concern about the use of AR technology. In order to adopt AR technology for the classroom in the next few years, engineers should consider the cost of the technology, usability, and time constraints, including the time to set up and cover the required materials for class. This feedback is related to the minimalism principle of Cuendet et al. [35]. Engineers must limit the functionalities of the system to what is required. Adding more functionalities than required would make AR more difficult to use.

4.3. System Design and Implementation

I created a HAR system that can display any combination of multimedia including image, animation, sound, and text on a real environment. The AR annotations are either labels or sprite sheet animations. I then created two AR applications for learning Filipino and German words in a real environment. I accomplished this by simply filling the HAR system with content for the situated vocabulary learning of Filipino and German words.

4.3.1 Design Goals

To summarize what I discussed so far, I list the following design goals based on multimedia learning, past works on situated vocabulary learning, and some practical considerations for future adoption to educational settings:

1. Minimize visual clutter
2. Support cognitive processes of selecting, organizing, and integrating information
3. Allow interactions with the environment and objects in the environment
4. Present multimodal information, namely, texts, images, and sounds
5. Use animations when appropriate
6. Apply cheap and accessible technology
7. Make the contents easy to create
8. Limit the interactions

4.3.2 The FlipPin HAR System

Figure 4.1 illustrates the package diagram of my system and Figure 4.2 shows the sample interface enabled by my system. The main part of the system is the Controller, which has access to learning contents, sensor (camera), and user inputs. The Controller receives the marker ID and camera view matrix from the Tracker and uses these information to specify the behavior of the on-screen display. The Tracker was built using ARToolKit, and the Renderer was built on OpenGL ES 2.04.

I use ARToolKit [89] to measure the camera pose with respect to the target object. Fiducial markers in the video feed is located using the ARToolKit, which also outputs the marker's ID and the matrix representing the current view of the camera. The image is transformed to the correct view using the matrix, and then it is rendered accordingly using OpenGL ES 2.04.

My AR system runs entirely on iPad tablets. For my experiments, I used the iPad 2 (dual-core A5, 512MB DDR2 RAM, 32GB, 601 g, 9.7 in display, 1024-by-768 at 132 ppi), and the iPad mini (64-bit A7, 512MB DDR2 RAM, 16GB, 331 g, 7.9 in display, 1024-by-768 at 163 ppi). The system works with fiducial markers (Figure 4.2) to determine the target object and the viewing angle of the tablet's back camera. I used the back camera set to 640x480 pixels at 30 fps to

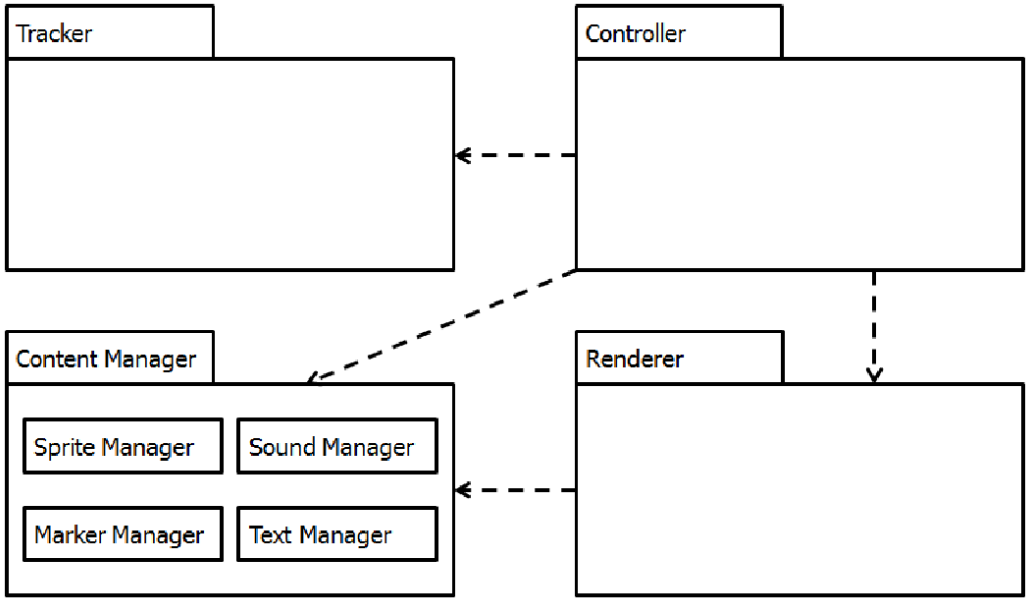

Figure 4.1. Package Diagram of the HAR System

sense the marker and to provide a video feed. After identifying the marker, the system loads the corresponding audio, text, and image. Audio and text can be accessed using buttons (listen, translate, and describe). The images can either be still images or sprite sheet animations (Figure 4.3; Figure 4.1). The images are transformed depending on the camera view and are inserted in the video feed to suggest three-dimensional registration, that is, to give an impression that the graphics co-exist with the real objects.

4.3.3 Situated Vocabulary Content

I used the platform describe in Figure 4.1 to create two situated vocabulary learning systems – one for 30 Filipino words and the other for 10 German words. I based the design of the content from previous works [111], [113], [112]. I used a combination of text, audio, images, and animations as content. The text data are the vocabulary, its translation in English, and the description of the scene (only for the Filipino version). The audio data is the proper pronunciation of the vocabulary as spoken by a native speaker. The image data are text labels,

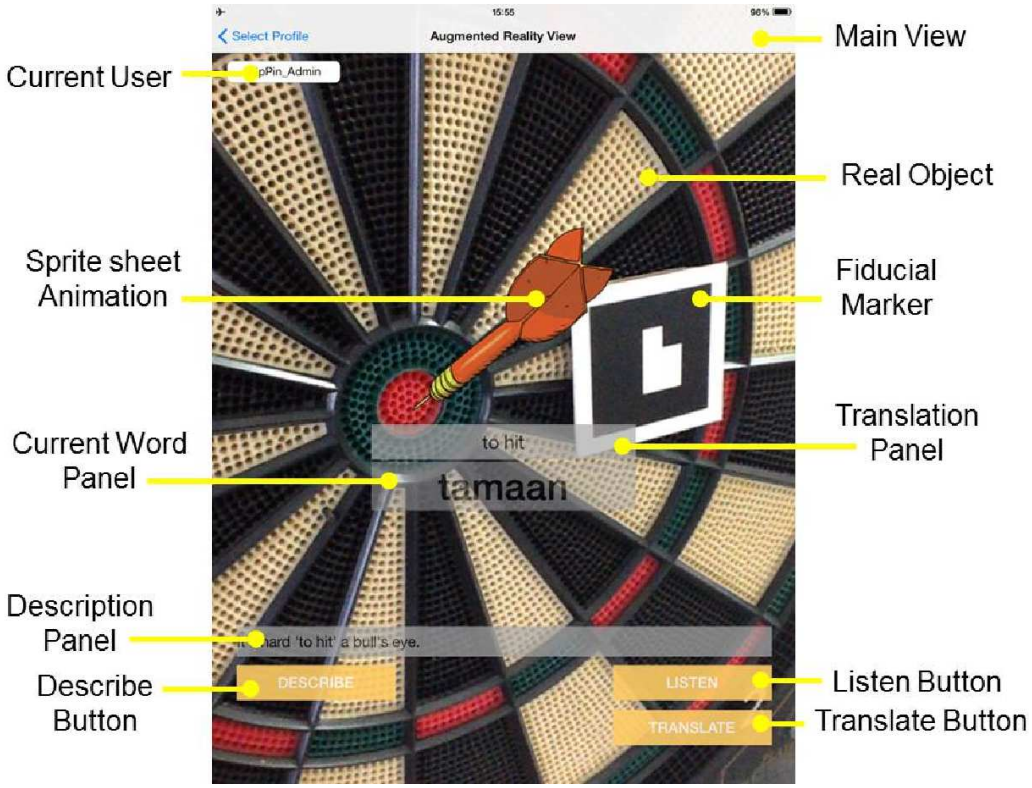

Figure 4.2. Sample Interface for Situated Vocabulary Learning

images, or animations, as shown in Figure 4.3.

4.4. User Studies on FlipPin

I explored the strengths of AR annotations for situated vocabulary learning over a non-AR counterpart (Figure 4.4) in two experiments. In particular, I am interested in the effects of AR on memorization and student motivation. Through these experiments, I evaluate the use of AR annotation for presenting vocabulary content that is situated in the real environment. I compared the AR systems to a non-AR version which is a tablet application that mimics flash card interaction. My comparison does not employ any kind of special instructional design, such as game mechanics and collaboration. As summarized in Table 4.1, users simply point the tablet PC to objects found in their environment when using my AR

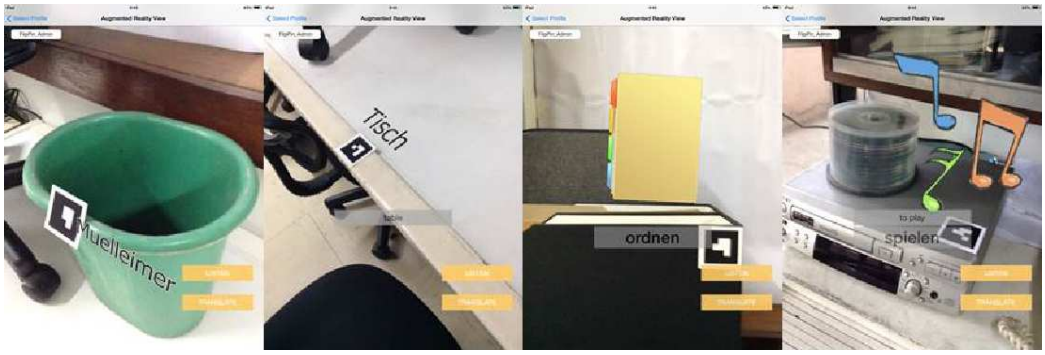

Figure 4.3. Label for Nouns, Sprite Sheet Animation for Verbs

application. On the other hand, the flash cards application allows the user to flip through contents by pressing either next or previous.

I considered inherent features of the interaction as part of the treatment. Thus, I made no attempts to control them. For example, one advantage of an AR learning system is that the students see the real objects in their surroundings even when they are not studying. I imagine this feature to trigger unintended rehearsal of the vocabulary, thereby improving memorization. This unintended rehearsal is part of AR learning. Thus, I did not control this aspect. I did not forbid the students in the AR treatment from visiting the study place when they are not studying.

Another inherent feature is that students tend to cover all the vocabulary words several times in one study session when flash cards are used. The flash cards are sequentially arranged, and students try to go through all the content two to four times in one sitting. Even if this is the case, interventions were not made because it is an inherent feature of the use of flash cards. Moreover, advising the students who use the AR system to view all the content several times will interrupt their natural learning style.

For my experiments, I controlled both location and time constraints. All of my students were only allowed to use the applications inside their respective laboratories. However, the applications are available to them at any time they want to study on that day. Given these features, I had seven hypotheses which I tested for significance in the 0.05 level via student's t-test. The hypotheses are as follows:

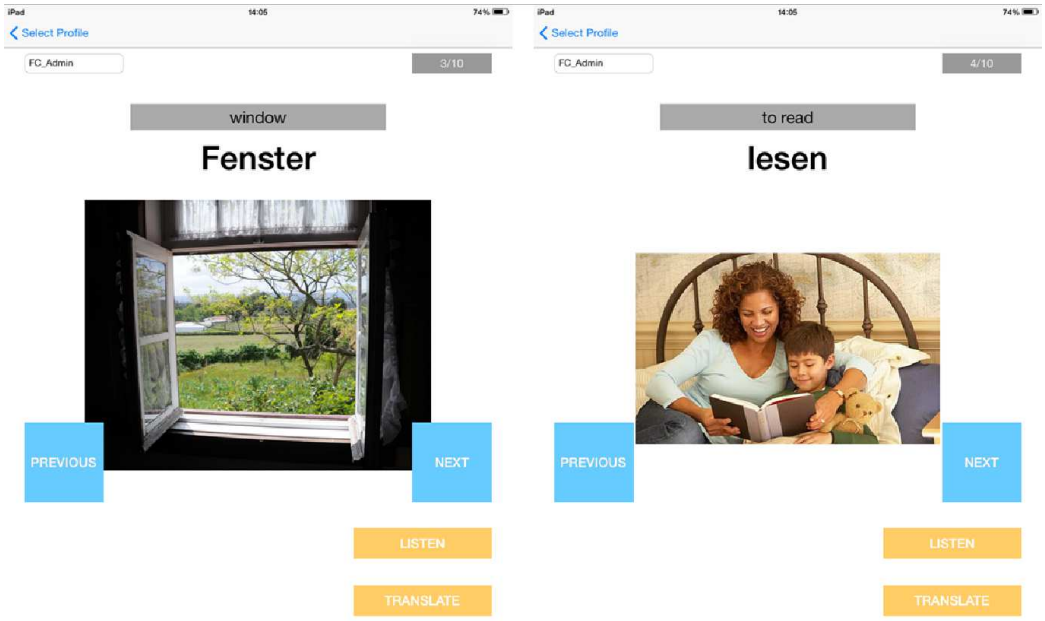

Figure 4.4. Non-AR version of the AR Applications

1. Students will perform worse on a delayed post-test with non-AR compared with the immediate post-test.
2. Students will perform worse on a delayed post-test with AR compared with the immediate post-test.
3. Students will perform better in an immediate post-test with non-AR.
4. Students will perform better in a delayed post-test with AR.
5. Students will rate AR as a more motivating instructional material.
6. Students will maintain their attention better with AR.
7. Students will find the contents presented through AR to be more relevant to them.
8. Students will feel more confident with non-AR.
9. Students will feel more satisfied with AR.

Table 4.1. Summary of Comparison of Two Interfaces for Vocabulary Learning

	AR Application	Non-AR Application
Interaction	Users find an object with a marker. They then point the tablet PC to the marker to reveal the content.	Users press "next" or "previous" to switch between contents.
Inherent Feature	Users can see the markers in their environment even when they are not studying.	Users can quickly go through all the material because they are arranged in a series.
Visual Display	Texts, images, sounds, and animations are displayed in on the real environment.	Illustrations are shown on a white background.
Place and Time	Users can only use it inside their laboratory at any given time.	Users can only use it inside their laboratory at any given time.

4.4.1 User Study 1: Filipino Vocabularies

I adopted a between-groups approach with 31 participants (26 male, 5 female, aged 23 to 42, information science graduate students) to test my application for studying Filipino words. The first languages of the participants are Japanese (13), Chinese (5), Portuguese (3), German, English, Turkish, Bosnian, Indonesian, Finnish, Arabic, Spanish, Nepali, and Wolof. In my experiments, I divided the people into the treatment groups with consideration to the distribution balance of their first languages.

Eighteen participants were recruited from one laboratory. I set up my system inside their laboratory (Figure 4.5) so that they can learn words related to their refreshment area. All of them have experienced using an AR application before, thus AR is not a novel technology for them. Thirteen participants from three laboratories were asked to use the non-AR version. Similar to the AR group, the non-AR group had used AR before and they are familiar with other novel interfaces. I distributed tablet computers to them with the flash cards application installed.

Figure 4.5. Refreshment area with markers (left), Learner using situated vocabulary learning (middle), Learner using non-AR vocabulary learning (right)

The participants used the assigned application for a recommended duration of 10 to 15 minutes per day for five days. The AR version was used inside a refreshment area with a maximum of four people using the application at the same time (Figure 4.5). On the other hand, the learners used the non-AR version wherever they went inside their laboratory office.

For my comparative analysis, I evaluated the usability of the application and the participants' learning outcomes. On the fifth day, the participants answered the questionnaires to measure the perceived usability of the applications. They then immediately took a post-test. After 12 to 14 days, they took a delayed post-test. The immediate post-test (27 items) and delayed post-test (24 items) consists of questions on recognizing the word in a multiple choice question, recalling the translation of the word, and guessing which word fits in different contexts.

Lastly, both AR and non-AR applications logged time-stamped button pushes, words studied, as well as tablet acceleration and orientation based on the built in sensors. I did not notice any burden on the application due to the system logging even after extended use.

4.4.2 User Study 2: German Vocabularies

I adopted a within-subjects design with 14 participants (8 male, 6 female, aged 17 to 20, Filipino undergraduate students) to test the application for learning 20 German words (10 for AR and 10 for non-AR). Each participant used the AR and non-AR versions for a maximum of eight minutes. Seven used the AR

version first, whereas the other seven used the non-AR version first to balance ordering effects. For the AR version, the learners viewed the content on a small area around a laboratory technician's desk. The markers were placed near each other in a small area to minimize the time spent on transferring from one object to another. This was important because I wanted to observe the study time of the students. For the non-AR version, they used the application while sitting inside the same room.

The students are then asked to answer 10 multiple choice questions that test their skill to recognize a word using a recognition game (Figure 4.6). Aside from logging the answer, I also logged the time it took for the learner to answer the question. After taking the quiz, the participants also answered a subset of the Instructional Materials Motivation Questionnaire or IMMS. I picked 30 questions that are applicable to my system out of the 36 questions listed in the work of Huang et al. [71]. IMMS models the extent of motivation one gets from an instructional material by using the ARCS model (Attention, Relevance, Confidence, and Satisfaction). This model had been previously applied to AR instructional materials by Di Serio et al. [47].

4.5. Results and Discussion

My experiments involved a small sample size, thus the results should be interpreted with caution. These experiments should be replicated with a bigger sample size. Nevertheless, these results can guide future design of AR systems and experiments in situated vocabulary learning with AR. In my experiments, I observed significant decrease in immediate to delayed post-test scores with non-AR, but not for AR suggesting that students who learned via AR retained more vocabulary. No significant differences were observed in learning outcomes between using AR and non-AR applications for vocabulary learning. However, students report better attention and satisfaction in using my system. In summary, I found evidence that supports hypotheses 1, 3, 6, and 9 but not 2, 4-5 and 7-8.

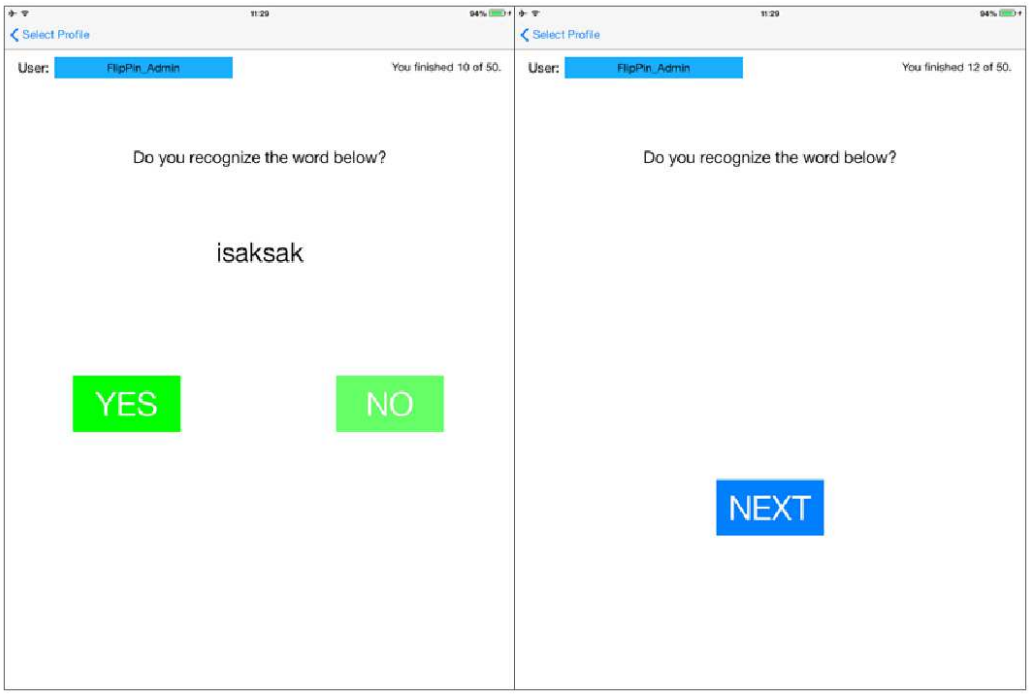

Figure 4.6. Screenshot of the Recognition Game

4.5.1 Comparison of Usability of Applications

I computed the SUS score and its factors from the participant responses in Experiment 1. The results in Table 4.2 show that the AR application has an SUS score of 74, which is close to the score of the non-AR application (80). According to Sauro [160], both interfaces were above average (SUS score>68). Thus, they were both good interfaces. Moreover, the results in Table 4.3 show that my participants did not have difficulty in learning these new interfaces.

I found a marginally significant difference between the two interfaces with a moderate effect size (d=0.63). Despite the differences in usability, using these applications for comparison was reasonable because both represent my best effort and both had above average usability. I achieved a good usability score because I applied previous research in multimedia learning. Furthermore, my current interface features were minimal and the task was simple.

Table 4.2. Summary of SUS Scores

	Application	N	Mean	SD	T value	p value
SUS Score	AR	18	74	12	1.64	0.055
	Non-AR	13	80	6		

Table 4.3. Summary of SUS Factor Scores

	Application	N	Mean	SD	T value	p value
Usability	AR	18	70	14	1.50	0.073
	Non-AR	13	76	7		
Learnability	AR	18	90	13	1.53	0.068
	Non-AR	13	96	5		

4.5.2 Manipulability and Comprehensibility of FlipPin

Aside from the SUS, I used HARUS (Chapter 3) to evaluate the usability of my system. HARUS is specifically design for HAR. It has two factors relevant to HAR, namely, manipulability and comprehensibility. Manipulability corresponds to the ease of handling the device when doing certain tasks. Usability questionnaires for software and mobile phones do not usually cover manipulability because software tends to be stationary and mobile phones tend to be held with a fixed posture. AR, on the other hand, requires the user to move around while pointing their handheld devices at various angles. This can be difficult sometimes due to unstable tracking of the natural environment, among other reasons. The second factor of HARUS is comprehensibility which is the ease of understanding the presented information. Although comprehensibility is common to all types of software, HARUS is designed for users to respond to AR-specific issues, such as the alignment of virtual contents and real environments, visual clutter, depth perception, etc.

Table 4.4 summarizes the HARUS score and its factors. My current prototype scored 61 (out of 100) in terms of overall usability, with a score of 63 on manipulability and 59 on comprehensibility. Compared to the usability score of 74, I think that I got a lower usability score from HARUS because it is more sensitive to AR applications. This current score can be used as a reference for the next

Table 4.4. Summary of HARUS Scores and its Factors

	HARUS	Manipulability	Comprehensibility
AR	61	63	59

iteration of my application. It could also be used as a benchmark for other AR applications for situated vocabulary learning. Through the use of HARUS, I may be able to compare HAR systems more accurately. However, its results should be interpreted with caution because HARUS is a relatively new questionnaire with some evidence of validity and reliability.

One of the straight-forward ways to improve the system is to use lighter devices. Some students reported that the iPad 2 is too heavy and it requires the use of two hands. Another way to improve the manipulability of my system is to use some ergonomically-designed handle for tablets, such as the work of Veas & Kruijff [181].

I think that applying markerless tracking, such as point cloud-based tracking using the PointCloud SDK, would decrease comprehensibility if I cannot detect good enough features to track the environment. Moreover, such feature registration process would be difficult to create if the content authors are teachers. For my current application, simply printing markers and placing them in the environment is an easier and more stable way of tracking the environment. However, I expect both markerless tracking technology and tablet computing power to improve significantly in the next few years. At that time, switching to markerless tracking would be practical.

4.5.3 Comparison of Information Retention

Table 4.5 is a summary of the results comparing the immediate and delayed post-test scores in Experiment 1. For the AR group, six people were not able to take the delayed post-test because they were inaccessible. (They were at their home towns at the time and did not check their emails 12 to 14 days after the study phase.) Both AR and non-AR groups decreased from immediate to delayed post-test scores. The difference for the non-AR group is significant with a large effect (d=0.84). Whereas, the differences for AR is marginally significant, with a small

Table 4.5. Comparing Immediate and Delayed Post-Tests

Application	Post-Test	N	Mean	SD	T value	p value
AR	Immediate	18	71%	20%	1.46	0.058
	Delayed	12	68%	23%		
Non-AR	Immediate	13	86%	20%	3.42	0.001
	Delayed	13	70%	18%		

Table 4.6. Comparing Immediate and Delayed Post-Tests for Nouns

Application	Post-Test	N	Mean	SD	T value	p value
AR	Immediate	18	79%	19%	1.30	0.100
	Delayed	12	69%	25%		
Non-AR	Immediate	13	90%	14%	2.78	0.005
	Delayed	13	71%	21%		

effect size (d=0.14). Thus, I found evidence supporting hypothesis 1 but not hypothesis 2.

These results are consistent with the work of Fujimoto et al. [57], [56] which reports that information associated with a place is better remembered. In my case, vocabulary that's associated with a place is better remembered than those that were abstracted (non-AR). However, I believe that an experiment with high sample sizes is necessary in order to better support this claim, and to better understand how familiar places contribute to the integration process of multimedia learning.

In this experiment, labels were used as annotations for nouns, whereas sprite animations were used as annotations for verbs. I separated the scores related to nouns and verbs. Based on the results shown in Tables 4.6 and 4.7, I observed that AR led to better retention. Using the non-AR application led to a significant difference for nouns with a large effect size (d=1.06). For verbs, I did not observe a significant difference. However, there is a 9% difference between the immediate and delayed post-tests.

Table 4.7. Comparing Immediate and Delayed Post-Tests for Verbs

Application	Post-Test	N	Mean	SD	T value	p value
AR	Immediate	18	64%	23%	0.75	0.230
	Delayed	12	71%	27%		
Non-AR	Immediate	13	82%	24%	1.03	0.157
	Delayed	13	70%	18%		

Table 4.8. Comparing AR and Non-AR

Post-Test	Application	N	Mean	SD	T value	p value
Immediate	AR	18	71%	20%	2.14	0.020
	Non-AR	13	86%	20%		
Delayed	AR	12	68%	23%	0.31	0.380
	Non-AR	13	70%	18%		

4.5.4 Comparison of Post-Tests

Table 4.8 compares the immediate and delayed post-tests in Experiment 1 for AR and non-AR. In the immediate post-test, the non-AR group scored significantly higher with a moderate effect (d = 0.75) compared with the AR group thus supporting hypothesis 3. I observed the same results for both nouns (d=0.66) and verbs (d=0.77), as shown in Tables 4.9 and 4.10, respectively. The breakdown in Table 4.11 shows that the AR group scored lower than the non-AR group in all types of questions. This result is indicative of an overall inferior mastery of content rather than a weakness in a particular question type.

In most practical cases, people do not usually apply their learning immediately

Table 4.9. Comparing AR and Non-AR for Nouns

Post-Test	Application	N	Mean	SD	T value	p value
Immediate	AR	18	79%	19%	1.86	0.037
	Non-AR	13	90%	14%		
Delayed	AR	12	69%	25%	0.27	0.400
	Non-AR	13	71%	21%		

Table 4.10. Comparing AR and Non-AR for Verbs

Post-Test	Application	N	Mean	SD	T value	p value
Immediate	AR	18	64%	23%	2.13	0.020
	Non-AR	13	82%	24%		
Delayed	AR	12	71%	27%	0.24	0.410
	Non-AR	13	70%	18%		

Table 4.11. Immediate Post-Test Scores for Each Question Type

Question Type	Application	N	Mean	SD	T value	p value
With illustrations	AR	18	87%	12%	0.99	0.163
	Non-AR	13	92%	20%		
Recognizing Filipino	AR	18	80%	15%	2.54	0.008
with Choices	Non-AR	13	94%	15%		
Recognizing Filipino	AR	18	64%	30%	1.95	0.031
without Choices	Non-AR	13	83%	24%		
Translating from	AR	18	55%	31%	2.54	0.008
English to Filipino	Non-AR	13	81%	23%		
Transfer word usage	AR	18	75%	19%	2.40	0.012
with choices	Non-AR	13	91%	16%		

Table 4.12. Delayed Post-Test Scores for Each Question Type

Question Type	Application	N	Mean	SD	T value	p value
With illustrations	AR	12	71%	27%	0.26	0.400
	Non-AR	13	73%	16%		
Recognizing Filipino	AR	12	67%	23%	0.70	0.247
with Choices	Non-AR	13	72%	13%		
Recognizing Filipino	AR	12	69%	30%	0.09	0.463
without Choices	Non-AR	13	71%	27%		
Translating from	AR	12	65%	28%	0.10	0.462
English to Filipino	Non-AR	13	64%	33%		
Transfer word usage	AR	12	64%	25%	0.87	0.196
with choices	Non-AR	13	71%	19%		

after studying. Rather, they would use their knowledge after a few days, either for a test or to apply it to a new lesson. As such, the delayed post-test is a more important point of comparison for learning than the immediate post-test. After 12 to 14 days, the significant difference in learning disappeared (Table 4.12). This is consistent with results of Lin and Yu [113] who reported that various multimedia modes did not have significant differences. However, the students did report differences in cognitive load. In experiment 1, the participants are graduate students who may not be sensitive to differences in cognitive load induced by an interface. For experiment 2, I asked a younger group of students to test my interface because they may be more affected by differences in cognitive load induced by interfaces.

4.5.5 Comparison of Post-Tests with SUS as Covariant

Assuming that implementation quality was a factor affecting the learning of the students, I could do fairer comparisons of post-test scores if both AR and non-AR applications have almost the same SUS score. However, I observed a small difference of six SUS points between the AR and non-AR applications. I conducted ANCOVA to take into account this difference in usability.

I can conduct ANCOVA because the difference in SUS score was not sig-

Table 4.13. ANCOVA of Post-Test Scores with SUS Score as Covariant

Post-Test	Application	N	Mean	SD	Adj. Mean	F value	p value
Immediate	AR	18	71%	20%	72%	2.14	0.090
	Non-AR	13	86%	20%	85%		
Delayed	AR	12	68%	23%	69%	0.00	1.000
	Non-AR	13	70%	18%	69%		

Table 4.14. Duration of Application Use (in minutes)

	Application	N	Mean	SD	T value	p value
Usage	AR	18	29.7	10.7	2.88	0.004
	Non-AR	13	55.8	36.5		

nificant. I also checked the homogeneity of variances using the Levene's test. The results of the Levene's test showed that there are no significant differences ($p>0.05$) in variances. The ANCOVA results in Table 4.13 are almost similar to the ANOVA results in Table 4.8. Marginally significant differences were observed in the test scores of AR and non-AR groups for the immediate post-tests. However, there is almost no difference in the delayed post-tests.

4.5.6 Comparison of Usage of Applications

To gain insight to the differences between studying with AR and non-AR applications, I calculated the total amount of time the application is open, and the total number of button pushes for listen, translate, and describe buttons. I found that the non-AR application was used significantly longer compared to the AR application (Table 4.14) – a finding I already expected after observing the participants study on the first day and on the fifth day.

In order to study with the non-AR application, the students had to keep the application open for the entire study period. However, when studying with AR, the students could put the application down and rehearse the words by going through each object in the room and calling out the vocabulary. In this case, using the application becomes unnecessary because the room itself represents the learning material. I think this connection with the digital content and the place

Table 4.15. Frequency of Button Pushing

Button	Application	N	Mean	SD	T value	p value
Listen	AR	18	408	364	1.01	0.160
	Non-AR	13	262	168		
Translate	AR	18	40	23	2.32	0.015
	Non-AR	13	16	23		
Describe	AR	18	69	70	0.35	0.365
	Non-AR	13	58	88		

is one important feature of AR that could be exploited for situated learning.

I also found some differences in the amount of buttons pushed in the AR application compared with the non-AR counterpart. All three buttons (listen, translate, and describe) where used more in general, with the translate button being pushed significantly more. This could mean that AR may be more motivating for students, specifically for maintaining attention as Di Serio et al. [47] reported. In another study, Ibanez et al. [72] reported AR's influence on learners' flow state, specifically on concentration, distorted sense of time, sense of control, clearer direct feedback, and autotelic experience. As such, for experiment 2, I applied the IMMS similar to Di Serio et al. [47] to observe motivation. For Experiment 2, I removed the describe button because students did not use it as much, and I did not see any significant differences in its use.

4.5.7 Comparison of Recognition Test

There was no significant difference between the recognition test between using AR (M=94%, SD=8%) and using non-AR (M=95%, SD=8%) for vocabulary learning. On the average, the non-AR group answered my multiple questions faster (M=2.28 s, SD=0.92 s) than the AR group (M=2.60 s, SD=1.03 s) for each question. However, this difference was not significant.

Table 4.16. Summary of the IMMS Score

	Application	N	Mean	SD	T value	p value
Motivation	AR	14	76	12	1.34	0.096
	Non-AR	14	71	11		

4.5.8 Comparison of Student Motivation Factors

Experiment 2 focuses on evaluating motivation by using the ARCS model. Although two interfaces can arrive at the same learning result, performance in tests should not be the only measure of success in creating interfaces. User experience is another important consideration. As such, I also evaluated the interfaces in terms of its ability to motivate students to learn.

Overall, the difference between the IMMS ratings of AR and non-AR are only marginally significant (Table 4.16). However, looking at the factors of the IMMS (Table 4.17) significant differences were observed in the attention and satisfaction factors. The students report that the AR application catches and holds their attention more than the flash cards application. This is consistent with the observations of Di Serio et al. [47]. Moreover, they report higher satisfaction with their learning experience. The learners were slightly more confident to use flash cards probably because it is a more familiar interface. This finding is opposite of that of Di Serio et al. [47]. The learners rated AR to be higher in relevance by five points, which is attributed to the implicit connection between learning contents and real environment. However, no statistical significance was observed for the relevance and confidence factors.

4.6. Chapter Summary

AR is useful for presenting situated multimedia. In this chapter, I discussed my experience in developing and evaluating an AR system for learning experiences based on real environments. As part of my development process, I drew design goals from multimedia learning theory, past systems for vocabulary learning, and needs of teachers. I then created a HAR system for displaying situated multimedia (text, image, sound, and animation). As a use case of the AR system, I filled the

Table 4.17. Factors of the IMMS Score

Factors	Application	N	Mean	SD	T value	p value
Attention	AR	14	75	14	1.84	0.038
	Non-AR	14	65	14		
Relevance	AR	14	74	14	0.97	0.172
	Non-AR	14	69	13		
Confidence	AR	14	80	12	0.74	0.232
	Non-AR	14	83	8		
Satisfaction	AR	14	77	16	1.71	0.049
	Non-AR	14	66	18		

system with Filipino and German vocabulary contents, thereby creating two AR applications for situated vocabulary learning.

I evaluated the AR applications by combining methods in human-computer interaction, usability engineering, and education technology. I observed differences in immediate post-tests results, with students who used the non-AR application scoring better than those who used AR. This effect is only temporary as both AR and non-AR users have almost equal scores in the delayed post-tests. Moreover, I observed a larger difference between immediate post-test to delayed post-test with the non-AR application. This suggests that using AR resulted to better retention.

Aside from differences in post-tests, the potential of AR lies in the difference in the learning experience, more specifically, reducing cognitive load, improving attention, and increasing satisfaction. My experiments suggest that using AR annotation may lead to better retention, attention, and satisfaction.

CHAPTER 5

Evaluations of AR X-Ray

Augmented reality (AR) annotations usually refer to virtual overlay on top of or beside real objects and scenes. However, advances in AR techniques allows the creation of illusions wherein virtual objects appear to be inside a real object or behind a scene. In this case, we refer to this AR annotation to be internal. Internal annotation is achieved using AR X-ray. Unlike other AR techniques, there are few studies investigating how AR X-ray affects human perception. Aside from human perception, there are few studies exploring the appropriateness of AR X-ray to educational settings.

In Part 1 of this Chapter, I conducted preliminary evaluations of the AR X-ray with teachers and students. I identify legibility as one of the possible usability issues affecting AR X-ray. Legibility contributes to the comprehensibility issues described in Chapter 3; i. e., users cannot understand the information presented to them when they cannot read symbols or distinguish objects presented through the AR X-ray. In standard AR, annotations are overlaid on top of the real world. To position a virtual annotation inside an object, AR X-ray requires partially occluding the virtual annotation with visually important regions of the real object. In effect, the virtual annotation becomes less legible compared to when it is completely unoccluded.

In Part 2 of this chapter, I compare the legibility of two methods for AR X-ray, namely, edge-based and saliency-based. In my first experiment, I explored the tolerable amounts of occlusion to comfortably distinguish small virtual objects. In my second experiment, I compared edge-based and saliency-based AR X-ray methods when visualizing virtual objects inside various real objects. I also benchmarked the legibility of these two methods against alpha blending.

5.1. Part 1: AR X-Ray in Learning Support

Several AR systems depend on "X-ray vision" as the primary use of AR. Sample applications include seeing through a pilot's cockpit floor and walls [58], visualizing ultrasound information within a patient's body [12], looking through buildings in local navigation [158], and studying underground pipes in construction [195]. Despite many advances in prototypes, the realization of AR X-ray remains challenging for general application due to lack of user studies. Currently, researchers are studying various depth cues based on the current understanding of the human visual system, and ways of measuring perceived depth to evaluate usability [120].

Similarly, in educational settings, one important affordance of AR is to "visualize the invisible" such as unobservable scientific concepts [190]. For example, unseen forces acting on an object and magnetism have been illustrated using AR. To teach the concept of force, Sotiriou et al. [170] used AR to integrate virtual arrows onto real carts found in a science museum display. Matsutomo et al. [125] used AR to draw virtual magnetic field lines onto real magnets. In their prototype, the shape of the magnetic field lines are computed in real-time to demonstrate how two magnets affect each other's magnetic field. Aside from naturally invisible concepts, an object may be practically invisible due to occlusion such as internal organs and engines of machines. This can be addressed using AR X-ray or providing the illusion of being able to peer inside a target object.

Recently, researchers [37] have described methods for porting AR X-ray to handheld devices making it practical for mobile AR learning. Applications like Environmental Detectives [97] and EcoMOBILE [88] have recommended a new instructional method that uses mobile AR games to investigate a real environment. According to these researchers, mobile AR supports ubiquitous learning

by facilitating the interaction of students with the real environment. AR X-ray is another interaction with the environment that instructional designers can take advantage of in a location-based game and other educational applications. Using AR, real objects found in an environment becomes a trigger for presenting information.

AR X-ray is a novel interaction for education. Thus, it is necessary to investigate how it affects the students' perception and the suitability of the state-of-the-art to the teacher's practice. State-of-the-art implementations of handheld AR (HAR) X-ray have not yet been tested extensively. Currently, the target is pedestrian applications such as navigation [37], thus the user studies focus on depth perception on the medium-field (beyond arm's reach to 30 meters) and the far-field (beyond 30 meters) distances. Implementations in these distances may not be suitable for the near-field (within arm's reach) distances, which is the case when students approach a real object and try to observe its interior. Moreover, current AR X-ray methods uses occlusion cues which sacrifices the legibility of the virtual object to convey depth. In educational settings, legibility is more important than conveying depth so that students can clearly observe the interior of a real object. For my research, I am interested in studying AR X-ray in the near-field wherein a student approaches a real object and views internal annotations on the object. We want to support this basic interaction which is useful in mobile AR games, and other instructional designs.

In this chapter, I describe the implementation of an edge-based AR X-ray using a tablet computer and evaluate it in the near-field distance with students and teachers for the first time. This AR X-ray interaction can be directly integrated into ubiquitous learning such as location-based games, etc. Furthermore, I describe my participatory approach for using user studies with students and teachers to iteratively refine our application.

5.1.1 Related Work

AR was first implemented with the use of head-mounted displays connected to computers. However, advances in handheld devices and network technologies enabled the use of HAR in the mobile design space. HAR has become a practical technology for ubiquitous learning. Using HAR, students can explore the real

world annotated with virtual information and observe some phenomena which would have not been apparent if not for HAR.

Supporting situated learning is often cited as one of the key advantages of AR to learning [48], [171], and [190]. Situated learning theory explains that learning takes place through the process called legitimate peripheral participation. Legitimate peripheral participation happens when a student increases his participation in a community of experts by interacting with their peers, experts, environments, and artefacts. One example is language acquisition wherein we start as infants uttering our first words, and become sophisticated speakers with thousands of vocabulary and intuitive grasp of grammar. This learning takes place as we interact with our parents, teachers, and friends in various situations and environments.

As a context-aware technology, HAR supports situated learning because it enables students to collaborate with other people and interact with the environment. For example, Klopfer and Squire [97] developed a HAR game called "Environmental Detectives" wherein students pretend to be environmental scientists investigating the spread of a toxin in their campus groundwater. In this research, the game supported collaboration by requiring the students to work in groups in gathering and processing information. Moreover, the game supported the interaction with the environment by guiding the students and virtually drilling wells to get sample groundwater. Another example of supporting situated learning is the "EcoMOBILE" project [88]. EcoMOBILE uses HAR to navigate students around a pond, and annotate relevant virtual information onto the real pond environment. In this study, teachers report that the HAR system promoted student interaction with each other, and with the pond ecosystem. In both Environmental Detectives and EcoMOBILE, students are learning how to solve problems situated in a real enviroment. Compared to traditional classroom instruction, learning with HAR systems enable students to apply their knowledge in real-world contexts more easily [48].

Developing AR X-ray directly contributes to situated learning with HAR. Using AR X-ray, students have a new way of interacting with their environment; i. e., using internal annotations, students can virtually explore the interiors of real artefacts found in the real environment.

5.1.1.1 AR X-Ray

Livingston, et al. [120] defines AR X-ray vision as "the ability to virtually 'see through' one surface to what in reality is hidden from view." AR X-ray is rendering virtual objects onto to the real world such that the virtual object is perceived to be behind or inside the real world object that is occluding it. Several techniques have been offered to provide an illusion of looking through objects with special care into the depth perception of users [120]. Current research studies occlusion cues to suggest depth to users when using AR X-ray in handheld devices.

Sandor, Dey et al. [43] introduced a "melting metaphor" to HAR to reveal occluded points of interest in an outdoor environment. Their application gives the user an illusion of melting buildings to reveal what is hidden behind it. In another study, Sandor et al. [159] compared legibility when using edge-overlay X-ray to when using saliency-based X-ray. Edge-overlay X-ray uses edges found in the real world scene as occlusion cues. In this research, we used edge-overlay AR X-ray and we discussed our implementation in Section 4. Instead of edges, three salient features including hue, luminosity, and motion can be used to create "saliency maps" as described by the visual saliency model of Itti et al. [73]. This saliency map is used to decide which occluding objects in the real world scene should be kept in the scene. Overall, there is no statistically significant difference in legibility between edge-overlay and saliency-based X-rays. Edge-overlay is better for scenes with high brightness and high edge surfaces. Saliency-based is better for scenes with medium to low brightness [159]. Aside from edges and salient features, Zollmann, et al. [195] demonstrated the use of textures found in the environment as occlusion cues. In cases wherein edges, salient features and textures cannot be found in the scene, they recommend the use of synthetic details for compensation. Zollmann, et al. did not compare their X-ray technique with other AR X-ray implementations.

There are very few works on user-based research investigating perceptual issues such as depth perception in HAR. Dey, et al. [37] was first to investigate depth perception in AR X-ray with mobile devices, such as iPhones and iPads. They report that when using the current state-of-the-art techniques for HAR, users underestimate distance for medium-field (beyond arms-reach to 30 meters) and far-field (beyond 30 meters) distances. They did not conduct a study for near-

field (within arm's reach) distances because the target of their work is pedestrian applications such us AR browsers for navigation [45]. Based on empirical research [37], there was no significant difference in depth perception at varying screen resolutions. Users understimated distances of virtual objects from theselves more on the the bigger device (iPad) compared to the smaller device (iPhone). However, using the iPad allowed for better estimation of distances between two virtual objects. One of their most interesting finding is that both the tracking method and edge-based X-ray do not influence depth perception in outdoor locations.

5.1.1.2 Development Models

For developing AR systems, Livingston [119] recommends a two-step approach in solving human factors issues such as depth perception in AR X-ray. First, he recommends to conduct limited perceptual tests that use only the well-designed parts of the user interface. This is to ensure that there are no perceptual issues that will hinder the users to perform higher level tasks with the interface. Second, researchers may proceed with comparing user performance on higher level tasks. This compares the interface against traditional methods for solving particular tasks. If researchers attempt to skip the first step, they risk testing an interface that may have usability issues. Gabbard and Swan [61] supports this by recommending a usability engineering model for AR. They proposed the iterative use of user studies to gain insights for the design. Their model adapts a user-centered design by iteratively refining the interface using feedback from:

1. User task analysis – Requirements of the user tasks are gathered and understood.

2. Expert evaluation – Experts examine paper mock-ups and prototypes based on design guidelines.

3. User study – Users are observed as they perform tasks with the interface.

In the field of educational technology, researchers need to design both the AR interface and the educational experience. To develop a HAR educational game, Klopfer and Squire [97] adapted a development process that integrates

principles from rapid prototyping, learner-centered software, and game design methodologies. Their development had six phases, namely,

1. Brainstorming – They conceptualized a novel educational software platform using HAR.
2. Design of first instantiation – They designed an exemplifying HAR game by envisioning user scenarios.
3. Development of first instantiation – They built a "quick and dirty" rapid prototype of the exemplifying HAR game.
4. Field trials of first instantiation elements – They tested novel elements of the application (e.g. GPS navigation software, concept of AR, basic game functions, etc.) with students and teachers.
5. Classroom implementation of first instantiation – They tested the performance of the whole first instantiation by asking the teachers to use it in class.
6. Platform design for creating next instantiations – Based on the knowledge from 1-5, they created a software toolkit to create similar educational HAR games for different learning scenarios.

Consistent with the work of Klopfer and Squire [97], Dunleavy and Dede [48] recommended the use of design-based research approach to study the feasibility of applying HAR in K-12. Design-based research is a combination of methods that refines educational applications by testing them based on principles from earlier research. Aside from insights obtained from learning theories and video game design principles, design-based research uses results of field testing individual elements and the entirety of the application with the actual target users. This type of formative research iteratively improves a HAR application similar to user studies explained by Gabbard and Swan [61].

5.1.2 Approach

My evaluation involves three main activities, namely, focused group discussions (FGDs) and interviews, first prototype development, and user studies. The FGD

is used to brainstorm about the use of HAR in general, and generate general user requirements for a first prototype. Then, I developed a demonstration of edge-based AR X-ray as the first prototype. Lastly, I used this prototype for studying the student's perception when using AR X-ray.

Figure 5.1 summarizes our activities in a compact model that illustrates the feedback that informs the design of my HAR system. This participatory design involves stakeholders such as teachers and students early in the design to ensure that the application suits their needs. In this model, the AR experts and/or educational technology experts implement a state-of-the-art AR interaction. Teachers and curriculum designers can be the experts that evaluate whether or not it is usable for their context. User studies with students can also be used to gain insights about the current design.

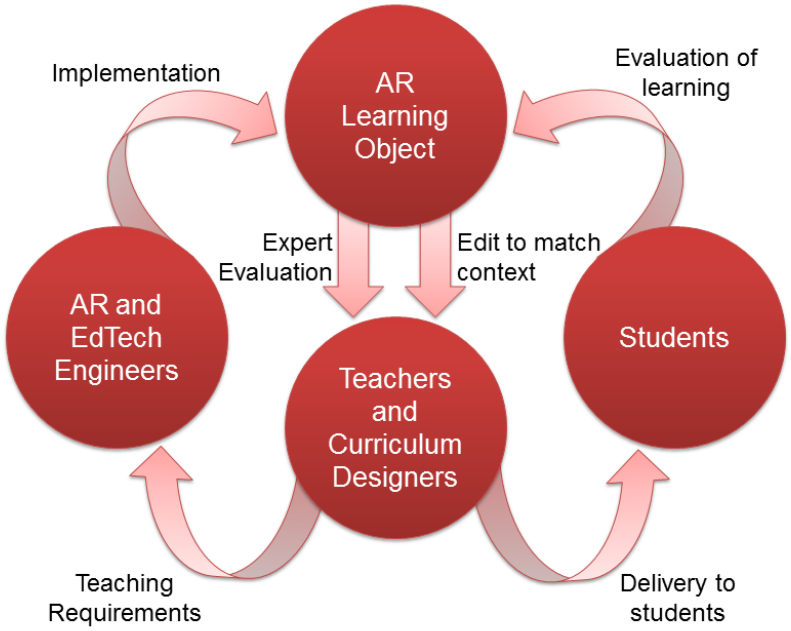

Figure 5.1. Participatory Design of Augmented Reality Learning Object

5.1.2.1 Focused Group Discussions

Two FGDs were conducted in two schools in the Philippines. The first FGD was conducted with the school principal and a school administrator from Spring Christian School in Muntinlupa City, Philippines. The second FGD was conducted with four teachers and two parents from Spurgeon School in Makati City, Philippines. The goal of the FGDs were to identify gaps in learning that could be addressed by AR. Moreover, we discussed foreseen difficulties in classroom implementation, and their willingness to adapt AR in the classroom.

As a starting point of discussion, I selected state-of-the-art prototypes designed for classroom use and sketched the application to the teachers. The applications I chose cover several topics, namely, butterfly life cycle [177], collision in physics citeLC11, human internal anatomy [18], playing the guitar [132] and magnetism [125]. Then, I asked the teachers open-ended questions. (E.g. What concerns come to mind when these applications are presented to you? Why do you think these are useful/not useful? Why is this suitable/not suitable for this particular topic? If you were to use this application, what would you like to remove/add?) The FGDs were conducted using colloquial language, Taglish, which refers to a combination of Filipino and English. This is the more natural conversation style in the Philippines. Each FGD lasted around one hour.

5.1.2.2 Implementation of Prototype

AR X-ray was implemented entirely on iPad 2 tablets with dual core Apple A5, 512MB RAM, and 32GB memory. I used vision-based tracking using the ARToolkit [89] which requires fiducial markers. The target real object is a cube (side = 60 cm) with print on the faces as shown in Figure 5.2 "Target Object." The virtual 3D models displayed in the interior are cultural artifacts as shown in Figure 5.2 "Virtual Object."

I use the edge-based AR X-ray as proposed in [158]. I obtain the image of the target object using the back camera of the iPad with a resolution of 480x640 pixel. Then, I use the Canny edge detector to extract edges from the target object and the scene. This edges are multipled to a radial mask to create an alpha mask. Based on the alpha mask, I blend the target object and the virtual object. The output of this blending is overlaid onto the target object.

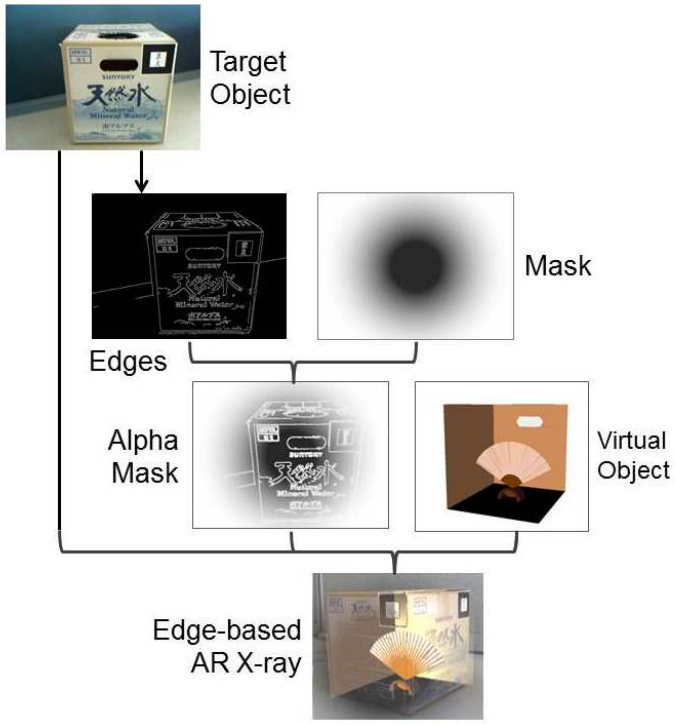

Figure 5.2. Implementation of Edge-based AR X-Ray

5.1.2.3 Constructs in User Studies

I chose three important constructs to AR X-ray, namely, depth perception, legibility, and realism. I then conducted user studies to evaluate my AR X-ray prototype based on these constructs.

Based on the literature [120], one of the important aspects of AR is depth perception. Although Dey, et al. [37] have shown that AR X-ray on an iPad device does not influence depth perception, their work is limited to medium-field and far-field distances. They did not test for the case of near-field or distances that are within arm's reach.

The second construct is legibility of the virtual object. Edge-based AR X-ray overlays edge-like textures on the target object to provide occlusion cues to the user. As such, the technique itself makes the virtual object less legible than in standard AR. Sandor, Cunningham, et al. [158] have confirmed in a test that too

many edges have a negative effect on legibility.

The last construct investigated is the feeling of realism in students. The teachers have noted that aside from the issue of legibility, another aspect that may be confusing for students is the concept of AR itself. They are interested to know how realistic the virtual information is perceived by the students.

5.1.3 User Studies

We employed three simple evaluations of depth perception, legibility, and realism by comparing edge-based X-ray (Figure 5.5.a) and simple virtual overlay (Figure 5.5.b). The goal of the user studies is to compare if the current implementation of edge-based X-ray influences simple virtual overlay which is the standard AR technique. Finally, I interviewed teachers to gather feedback on AR and AR X-ray.

For the three evaluations with students, I test the following hypotheses:

H1. AR X-ray conveys depth more.

H2. AR X-ray is less legible.

H3. AR X-ray is more realistic.

5.1.3.1 Set Up

Figure 5.3 shows a user study participant using the AR X-ray prototype. I implemented the system entirely on the iPad. It uses fiducial markers placed on the target object for tracking. In the user-based evaluations, I asked the users to explore what is inside the box using the application.

The prototype implements edge-based AR X-ray as described in Section 5.1.2.2. Figure 5.4.a shows the real world scene. Figure 5.4.b shows the edges detected from Figure 5.4.a using the canny edge detector. The edge detection is suppressed on the sides as shown in Figure 5.4.c before applying to the final edge-based X-ray in Figure 5.4.d.

Figure 5.5 shows the simple virtual overlay versus the AR X-ray. Figure 5.a shows the simple virtual overlay which does not include occlusion cues to provide an illusion of X-ray vision. This is used as the control scenario for the user studies.

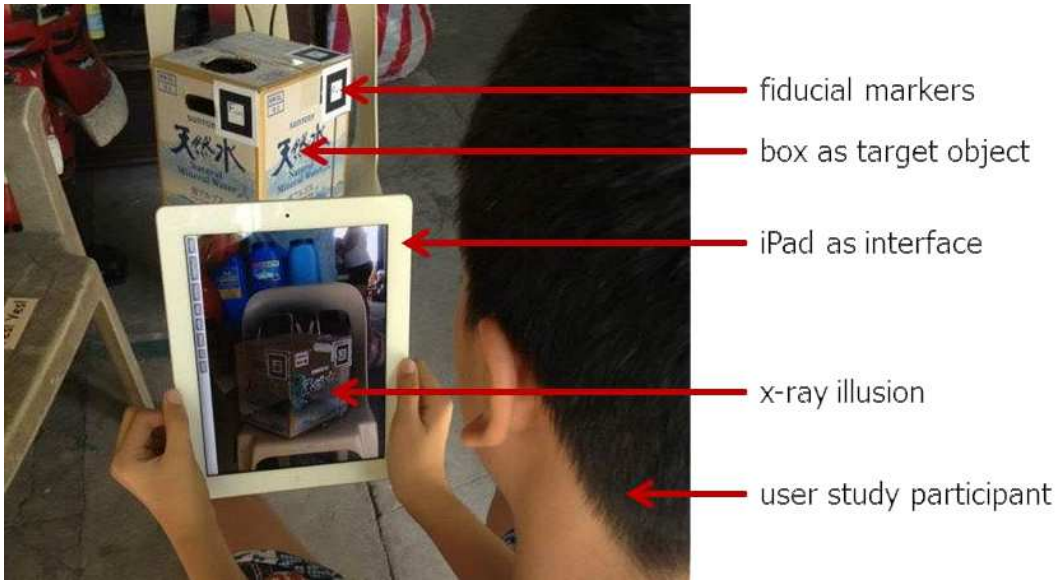

Figure 5.3. Overview of the System

5.1.3.2 Participants, Procedures, Instruments

I conducted empirical user studies with two groups of students and an interview with teachers to gather feedback.

5.1.3.2.1 Study with Students

The first evaluation was conducted with 23 Filipino students (9 male, 14 female, 5-15 years old). The study was conducted either at the participant's home, or the home of a relative or family friend. I obtained permission from the parents to conduct this study, with the parent waiting outside the testing room. Similar to the pilot test, the participants were divided into control-first (8 participants) and experiment-first (15 participants). The students were asked to respond to 6 statements in an interview format:

1. The object is inside the box.
2. The object is easy to see.
3. The object seems real.

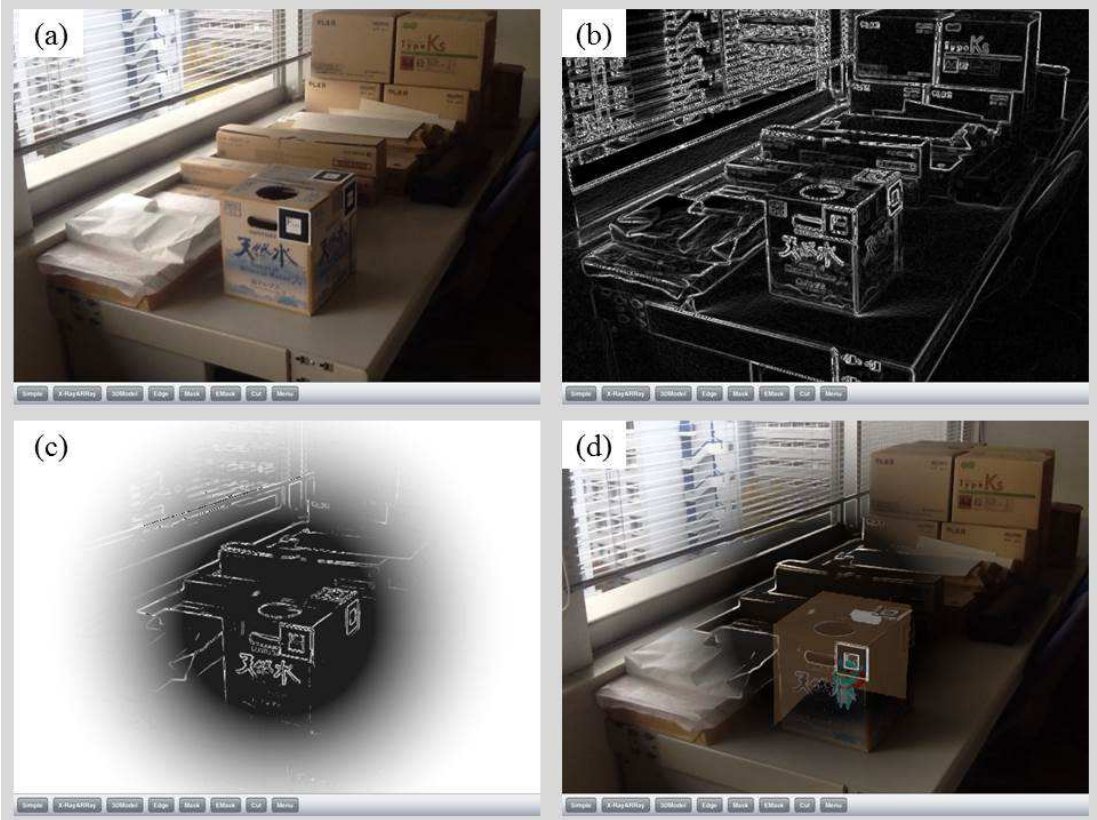

Figure 5.4. Edge-based X-Ray

4. The object seems flat.

5. I can see different parts off the object clearly.

6. My classmates will say the object is real.

The questions were read to the students and then translated in Filipino. Students respond by picking one of five possible answers arranged in a 5-point Likert scale:

1. No! No!

2. No!

3. I don't know.

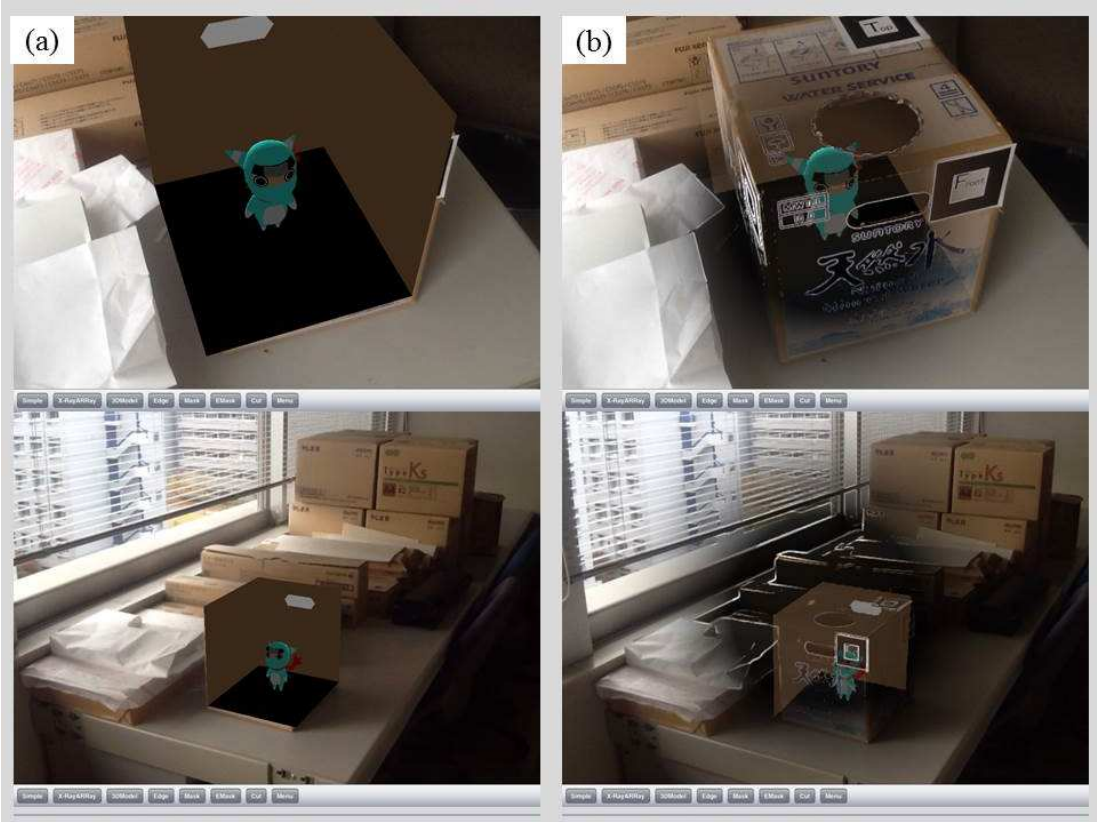

Figure 5.5. (a) Simple Virtual Overlay and (b) AR X-Ray

4. Yes!

5. Yes! Yes!

I included a short 5-10 minute break between evaluating the control scenario and the experiment scenario. Aside from question 4, higher scores would correspond to higher perception of the construct. I inverted the response to question 4 by subtracting it from 6, thereby getting a score that is parallel with the other questions.

Lastly, I conducted a debriefing interview with the participants primarily for them to make sense of the experience. This is particularly necessary because testing AR systems with younger children can be a confusing experience for them. Moreover, I conducted the debriefing interview to gather the students' impression

of AR X-ray.

5.1.3.2.2 Study with High School Students

The second evaluation was conducted with 47 Filipino high school students (21 male, 26 female, 11-16 years old) of Spring Christian School in Muntinlupa City, Philippines. I divided the participants into two groups. The control-only group (n = 21) viewed only the simple virtual overlay. Whereas, the experiment-only group (n = 26) viewed only the edge-based AR X-ray. Both groups were asked to answer the same statements in Section 5.1.3.2.1 translated to Filipino. Participants can respond to these items in a 5-point Likert scale with 1 corresponding to Strongly Disagree and 5 corresponding to Strongly Agree.

5.1.3.2.3 Interview with Teachers

I evaluated the edge-based AR X-ray with teachers in an interview. I demonstrated the edge-based AR X-ray to twelve teachers from Spring Christian School and Spurgeon School in Makati City, Philippines. The interview flow revolved around whether or not the current implementation is appropriate and useful for their practice. I also explained two sketches of possible applications of AR X-ray. The first one is for looking inside the body to see the skeletal system as envisioned by Blum, Kleeberger, et al. [18]. The second one is looking inside plants to see how water is transported. Based on these two examples, teachers were asked to identify advantages, disadvantages and suggestions on these proposed applications.

5.1.4 Results and Discussion

Table 5.1 and Table 5.2 summarizes the responses of the participants in the two empirical studies. All the differences I observed were not significant, except for one legibility question in Table 5.1. This particular finding supports hypothesis 2. AR X-ray is significantly less legible compared to simple virtual overlay with a moderate effect size (d = 0.63). However, I did not observe this result in the other questions corresponding to legibility.

All twelve teachers expressed their interest in learning materials using AR

Table 5.1. Results of Evaluation with Grade School and High School Students

Construct	Treatment	N	Mean	SD	T value	p value
Depth	Experiment	23	3.6	1.3	n. s.	n. s.
(Question 1)	Control	23	3.7	1.2		
Depth	Experiment	23	4.0	1.2	n. s.	n. s.
(Question 4)	Control	23	4.1	0.9		
Legibility	Experiment	23	3.3	1.2	2.1	<0.05
(Question 2)	Control	23	4.0	1.0		
Legibility	Experiment	23	3.6	1.3	n. s.	n. s.
(Question 5)	Control	23	4.0	0.9		
Realism	Experiment	23	2.7	1.3	n. s.	n. s.
(Question 3)	Control	23	3.4	1.4		
Realism	Experiment	23	3.1	1.3	n. s.	n. s.
(Question 6)	Control	23	3.3	1.2		

Table 5.2. Results of Evaluation with High School Students

Construct	Treatment	N	Mean	SD	T value	p value
Depth	Experiment	26	3.4	1.2	n. s.	n. s.
(Question 1)	Control	21	3.4	1.2		
Depth	Experiment	26	3.3	1.0	n. s.	n. s.
(Question 4)	Control	21	3.3	1.0		
Legibility	Experiment	26	4.0	1.1	n. s.	n. s.
(Question 2)	Control	21	3.8	1.1		
Legibility	Experiment	26	3.7	1.0	n. s.	n. s.
(Question 5)	Control	21	4.0	0.6		
Realism	Experiment	26	3.0	1.3	n. s.	n. s.
(Question 3)	Control	21	3.1	1.2		
Realism	Experiment	26	3.7	1.0	n. s.	n. s.
(Question 6)	Control	21	3.4	1.0		

X-ray, and they are willing to undergo some training for using such "high-tech" materials. They believe that some topics can be illustrated more clearly to students when using AR X-ray. According to the teachers, they regularly improve their skills by attending seminars or workshops. This could be a venue for learning about using AR X-ray materials. They identified three key advantages of AR X-ray for teaching:

1. Experiential learning – Eight of the teachers identified "learning by experience" as an advantage of AR X-ray. Aside from learning from illustrations in books, students can be given another kind of experience that catches their attention and motivates them to learn some more.

2. Improved attention – The teachers said that AR X-ray will generate interest in students because many of their students are visual learners. Moreover, AR X-ray runs on handheld devices that their students use for entertainment. As such, they associate handheld devices to enjoyment.

3. Increased motivation – Teachers speculate that AR X-ray will motivate students by first catching their attention, and then generate interest to know more. Using this novel visualization, students may be encouraged to ask more questions, or find answer by themselves in textbooks. In this scenario, AR X-ray complements their practice, and it does not replace the need for teachers or books.

The teachers commented three issues that need to be addressed to successfully integrate AR X-ray to their practice. The first issue is the overhead cost of adapting new technology. New technology entails expenses for the school for the hardware, software, and training requirements. In the case of HAR systems, it would require one device per child. In the Philippines, some private schools have already adapted tablet computers as replacement for some textbooks, and as part of a general computer laboratory class. The second issue is the perception and accuracy of the virtual information. Teachers noted that AR X-ray is highly subject to misinformation. The virtual abstraction provided by AR X-ray may be inaccurate because of poor tracking. For example, in displaying body organs onto a student, the size of the organs and their relative positions will vary depending

on the body type of the students. Inaccurate integration of the virtual objects can misinform students about their bodies. Lastly, AR X-ray can be too interactive for students and thereby consume more time. Learning modules using AR must be carefully planned to work with the time constraint allotted for the lesson.

5.1.5 Summary of Part 1

AR X-ray is a technique for making annotation inside real objects. Internal annotations are useful in HAR educational games and other designs for ubiquitous learning. Furthermore, HAR supports situated learning because it facilitates interaction among students and with the real environment. Using AR X-ray, students can virtually look inside a target real object. Recently, methods for achieving AR X-ray on handheld devices have been proposed making it ideal for ubiquitous learning. However, there is a lack of user studies evaluating the state-of-the-art techniques.

I described an edge-based approach to achieving the effect of "X-ray vision" and evaluated it with students and teachers. Edges from the target object are extracted and used as a mask to select which parts of the target object should be displayed as occlusion cues. I used the ARToolkit for tracking, and I implemented AR X-ray entirely on a tablet computer. As a novel interaction for education, I investigated the student's perception of depth, legibility, and realism in a series of user studies. Results of the evaluation with students show that except for legibility in one study, there are no significant difference between AR X-ray and simple virtual overlay, which is the standard AR technique. I investigate the legibility AR xray more deeply in the next chapter.

The teachers were interested in using AR X-ray because they think that it will give chance for students to learn by experience. They believe that it will arouse interest in students and may lead them to ask more questions in class. However, key issues in expenses, accuracy of virtual information, and teaching time constraints must be addressed to adapt this technology in practice.

5.2. Part 2: AR X-Ray Legibility

AR and its related technologies, systems and applications enable many novel visualizations that can be applied to various fields [23]. Among these novel visualizations, X-ray vision or seeing through an occluding surface [120] leverages the inherent capability of AR to display a combination of real environments and virtual objects simultaneously.

Despite the many advances in prototypes, the realization of AR X-ray remains challenging for consumer applications because of the need for practical hardware for computing and displaying X-ray visualizations properly. Moreover, we need more user studies to explore this rather super human sense that this interaction technique offers. An important point of evaluation for AR X-ray is the legibility of the visualization. Although significant effort has been made to improve the legibility of AR X-ray in [85], [86], and [87], there is little empirical research aimed at exploring legibility in various methods for creating X-ray visualizations. AR X-ray relies on partially occluding virtual objects to convey depth to the user. However, there is a trade-off between depth cues and legibility because occluding the virtual object reduces its legibility. Thus, the task is to find the sweet spot wherein adequate occlusion cues are provided while keeping the virtual object legible.

Both legibility and depth cues are subject to the users' intention. For example, different users may be interested to see some specific parts of a virtual object, but not the rest. Current AR X-ray systems use an image-based approach which performs calculations on images such that the all parts of the images have equal importance. In cases wherein the user is only interested in a specific part, some user input mechanism must be available to favor what the user is interested in. In my thesis, I am interested in using AR X-ray in the near-field (approximately within arm's reach) to support ubiquitous learning. Using AR X-ray, I can make annotations inside real objects. These "internal annotations" are useful for studying human anatomy, situated educational games, etc. as discussed in Chapter 5.

In this chapter, I offer a first exploration of legibility in AR X-ray. In particular, I study legibility in my implementation of edge-based and saliency-based AR X-ray systems. Moreover, I explore using user inputs for improving legibility.

5.2.1 Related Work

AR X-ray employs image-based techniques to preserve parts of an occluding object that are important to understanding it, while removing the rest to reveal occluded objects. Several methods have been applied, among which are edge-based and saliency-based methods. Although some empirical evaluations of perception exists, the focus is on depth perception; whereas the legibility of the visualization remains unexplored.

5.2.1.1 Partial Occlusion

Livingston et al. [120] defines occlusion to be when a closer opaque object (the occluder) prevents light rays bouncing off a farther object (the occluded) from reaching an observer, thereby making the occluded invisible to the observer. Partial occlusion occurs when the occluder blocks only a fraction of the light rays bouncing from the occluded.

Occlusion and partial occlusion are important for perceiving depth in our natural environment. For example, when examining a skyline, people can identify which building is nearer to his position by identifying which building partially occludes another. Similarly, people can understand an object to be inside a translucent container, say a wine bottle, through partial occlusion.

5.2.1.2 Image-based Techniques

Current methods for achieving AR X-ray relies on image-based techniques to determine important regions of an occluding real object. These regions are then preserved by rendering them on top of the virtual object, after the virtual object has been overlaid onto a real environment. AR X-ray requires an importance map which is an image representing the important regions of a real object. Figure 5.6 shows examples of real objects and their corresponding importance maps. For the edge-based AR X-ray, the importance map is based on the edges detected on the object. For the saliency-based AR X-ray, the importance map is based on the visual saliency map [73] of the object.

In [85] and [86], Kalkofen et al. developed a visualization technique that partially occludes virtual objects with edges found on the occluding real object. In

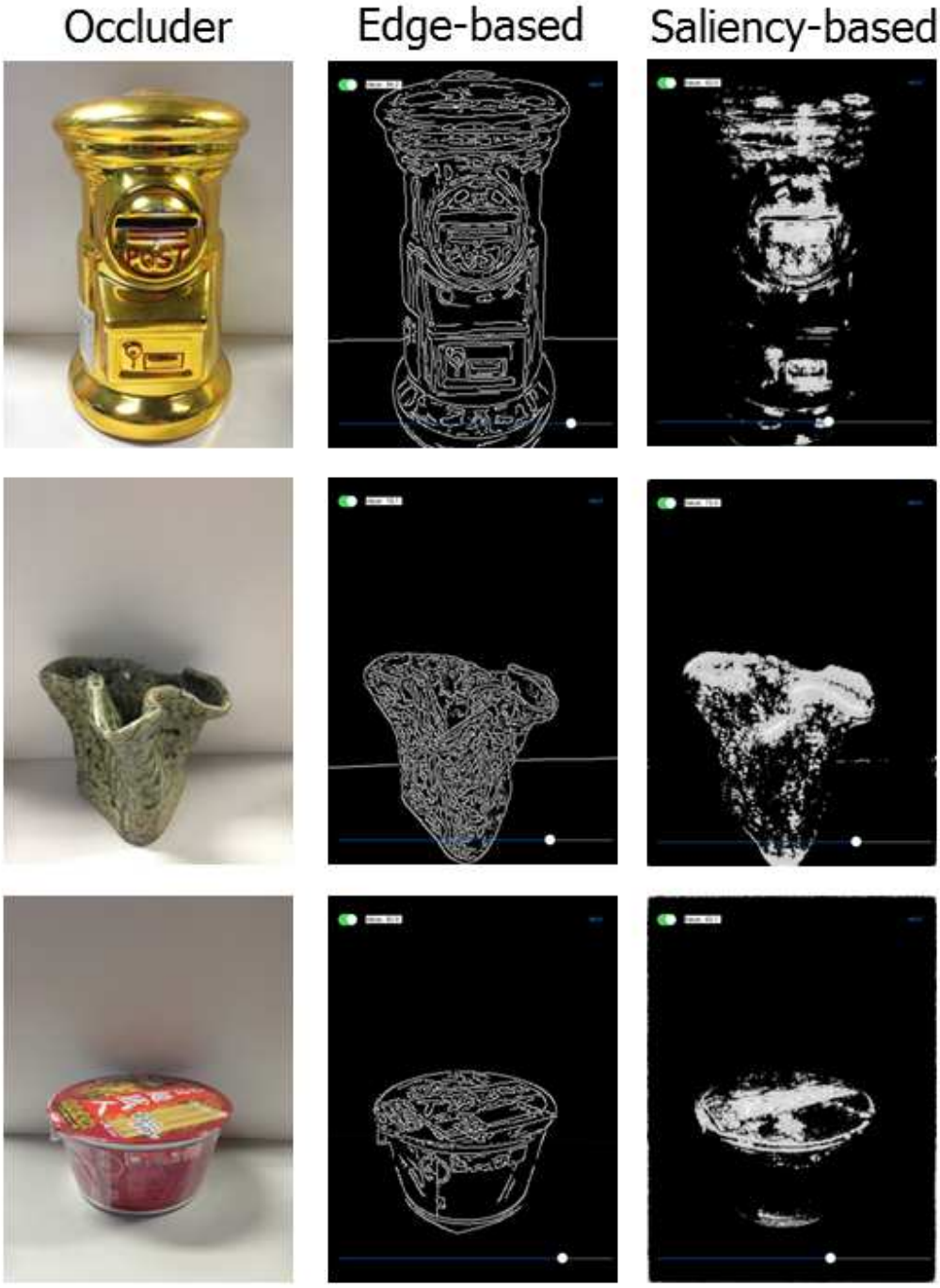

Figure 5.6. Left column shows the possible occluding objects. Middle and right columns are the importance map generated by detecting edges and salient regions, respectively. Darker areas are less important, whereas lighter areas are more important.

their visualization, the virtual objects are being viewed inside real objects which serves as the context of visualization. By maintaining the edges of the occluding object, Kalkofen et al. achieves a context-preserving AR X-ray visualization. They explained that preserving edges solves two important requirements of AR X-ray visualization. First, the edges are able to provide enough depth cues to convey the spatial relationship of the occluder and the occluded. Second, they provide important information about the occluder, such as its shape and prominent features.

Avery et al. [9] applied the edge-based approach of Kalkofen et al. [85] to their AR X-ray system. They used an edge detection filter on live video images to detect sharp changes in the luminance of the occluder. Edges are represented by thin white lines and are then overlaid on live video images of the occluded. They explained that drawing these lines maintains the major shape of the occluder. Moreover, the occluded remains visible because the AR x-ray method uses minimal occlusion cues.

Sandor et al. [158] introduced another method for generating the importance map. Instead of detecting edges, they detect visually important regions based on the saliency computational model of Walther [183]. In their computation of the saliency map, Sandor et al. considered luminosity, color opponency, and motion as observed in the changes in the luminosity channel. This implementation follows the intuition that bright areas, highly-contrasted colors, and moving objects tend to capture people's attention and are thus important to understanding the scene.

Around the same time with [158], Zollman et al. [195] offered a more complex method for creating the importance maps for their AR X-ray. In their work, the importance map is a combination of edges, saliency maps, textures, and synthetic details. Aside from considering edges and salient regions of an image to be important, they also considered highly-textured regions to be important to understanding the scene. In cases wherein few edges, salient regions, and textures are found, their method adds synthetic details to provide occlusion cues.

Kalkofen et al. [87] improved on the work of Zollman et al. [195] by making the method adaptive. After maintaining important parts of the occluder and deleting its unimportant parts, an additional module automatically adjusts the contrast to more clearly separate the occluder from the occluded virtual object.

5.2.1.3 Past Empirical Evaluations

Several studies have conducted empirical evaluations of various methods of achieving AR X-ray. Most of the past research revolves around understanding depth perception in AR X-ray. However, some hints have also been mentioned regarding the legibility issues of AR X-ray as reported by users.

In [9], Avery et al. argues that their edge-based method provides a good sense of depth. However, they note that some visualization design is necessary when displaying the occluded. In some cases, it is possible to confuse the edges found on the occluded to be part of the occluder. This hampers depth perception. They noted that their method is limited because the sensitivity to edges is fixed. As such, if the background is too cluttered, many edges will be drawn, thereby making the occluded difficult to see. However, they did not conduct a formal evaluation of legibility.

In [158], Sandor et al. compared their edge-based and saliency-based AR X-ray systems for the far-field (beyond 30 meters) distance. In their experiment, users are asked to find a target on a 640 x 480 pixel screen. The targets are either big (16 pixels) or small (9 pixels) red circles. Overall, they did not find significant differences on the time taken to find the target. However, participants were significantly faster with the edge-based X-ray than with the saliency-based X-ray when finding the small red circles. This suggests that legibility becomes an issue when the targets become smaller. The participants preferred the edge-based AR X-ray over the saliency-based AR X-ray. However, this difference was not significant. In a follow up experiment, Sandor et al. confirmed that high levels of edges causes problems for edge-based AR X-ray and that high levels of brightness causes problems in saliency-based AR X-ray. In this paper, I explore more on these problems in my experiments.

The work of Kalkofen et al. [85] allows some user inputs to modify the X-ray visualization. This is their pre-emptive solution to address possible problems in depth perception and legibility. Kalkofen et al. allows users to select parts of the real object wherein the X-ray visualization will be applied. This allows the user to specifically input the part of the image wherein they need some help to understand the visualization. Similarly, Zollman et al. [195] foresaw possible problems in depth perception and legibility so they recommended a parameter

that could be modified by the user to adjust the importance map. However, their system did not apply such user inputs.

Instead of user inputs, Kalkofen et al. [87] improved AR X-ray by adding another step that automatically adjusts the contrast of the occluder and the occluded. By adding this step, users performed significantly better in finding targets in their AR X-ray visualization. This approach is straight-forward because it applies the rule that previously important regions of the occluder must remain an important region by adjusting its contrast. Although this automatic method is good for the target acquisition task, it does not consider the intention of the user, such as in the system of Kalkofen et al. [85]. I believe that it is still important to understand the cases wherein AR X-ray visualization methods result in illegible compositions, so that we can provide user inputs to adjust the X-ray visualization.

Dey et al. [44] were first to compare depth perception in AR X-ray when using handheld devices, such as iPhones and iPads. They found that when using the current techniques for handheld AR, people underestimate distances in the medium-field (beyond arms-reach to 30 meters) and in the far-field distances. Overall, they did not observe any effects of AR X-ray on depth perception [46]. Users underestimated distances of virtual objects from themselves more on the iPad compared to the iPhone. However, using the iPad allowed for better estimation of distances between two virtual objects. Participants also expressed their preferences for the iPad over the iPhone for AR X-ray. Lastly, the varying screen resolutions of the iPad and the iPhone did not result in significant differences [44]. I agree with Dey et al. that handheld devices, such as smartphones and tablet computers, are appropriate platforms for AR X-ray. As such, I use tablet computers for my exploration of legibility of X-ray visualization.

5.2.2 Approach

For developing AR systems, user studies are useful for gaining insights to address human factors issues found in novel interaction techniques [61]. In the case of X-ray visualization, comprehensibility or the ease of understanding the presented information will contribute significantly to the overall usability of an AR system, as discussed in Chapter 3. Given the limitation of AR to currently available hardware, Livingston recommends confining human factors experiments to testing

only with the well-designed components of the whole AR system [119].

In exploring legibility, I apply methods that have been used for studying legibility in other AR systems. Not surprisingly, legibility is an issue in annotating real environments with virtual texts. Various methods, such as label separation methods [146] and active text drawing styles [62], have been employed for this purpose. In both these studies, the researchers are proposing an improvement in the legibility of virtual texts drawn in the natural environment. To accomplish this, they followed this pattern:

1. Conduct a literature review to inform the design of the AR system.
2. Create a high-quality prototype that represents the idea. However, implement only the necessary parts of the system for testing [119].
3. Execute an experiment to validate hypotheses or gain insights as to how the idea affects human perception.

I also followed this pattern in this chapter. I conducted a literature review in Section 5.2.1 to discover the proper implementation and find out the current problems reported by users. Then, I implemented two X-ray visualizations based on the edge-based method and the saliency-based method. In my prototype systems, I did not implement the tracking part of AR because I am only interested in how well the composed image of AR X-ray visualization is understood by the users. As such, I assumed that the AR system can be consistently held properly while doing the visual tasks. By doing this, I can separately study legibility issues due to X-ray visualization, and legibility issues due to unstable tracking [119]. In other words, I prevented tracking instability [44] from interfering with performing the visual tasks.

In two experiments, I explored how edge-based and saliency-based AR X-ray techniques affect the performance of users in some visual tasks. Similar to [146] and [62], my data set is composed of multiple perception judgements from the users. I then treat the users as random variables during the analysis. Note that I conducted usability evaluations such that, the object of the experiment is the AR X-ray system and not the human person [135]. The goal of this exploration is not to understand human perception, but to generate ideas on how to improve legibility in AR X-ray systems.

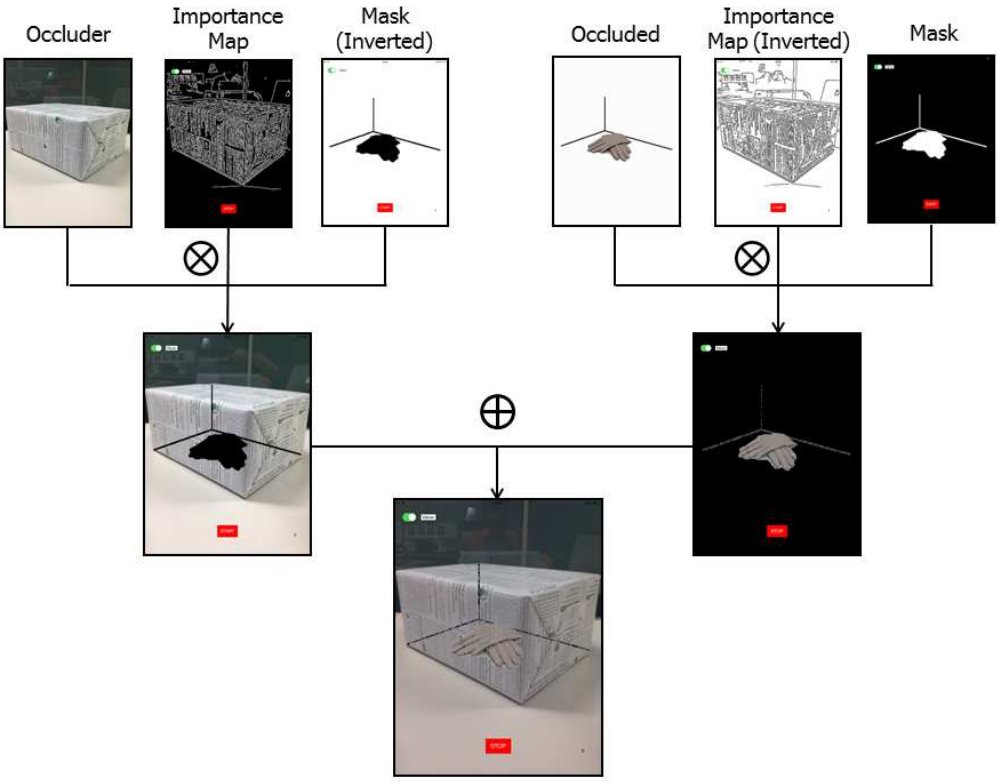

Figure 5.7. AR X-ray performs a series of calculations. Occluder, importance map, and mask (inverted) are multiplied per pixel. Occluded, importance map (inverted), and mask are multiplied per pixel. The resulting images are then added per pixel.

5.2.3 Implementation

I implemented both the edge-based and saliency-based AR X-ray visualizations by following the diagram in Figure 5.7. The calculation requires the images of the occluder and its importance map, and the occluded and its mask. The occluder, importance map, and mask (inverted) are multiplied per pixel. Then, the occluded, importance map (inverted) and mask are multiplied per pixel. The resulting images from these two operations are then added per pixel.

To generate the importance map of the edge-based AR X-ray, I used OpenCV

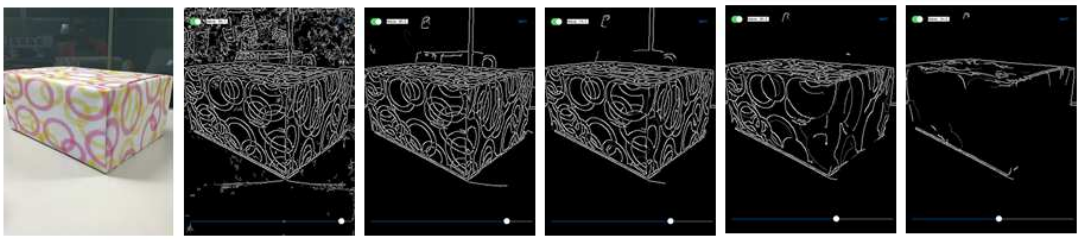

Figure 5.8. Edge-based importance map as threshold is increased using my slider.

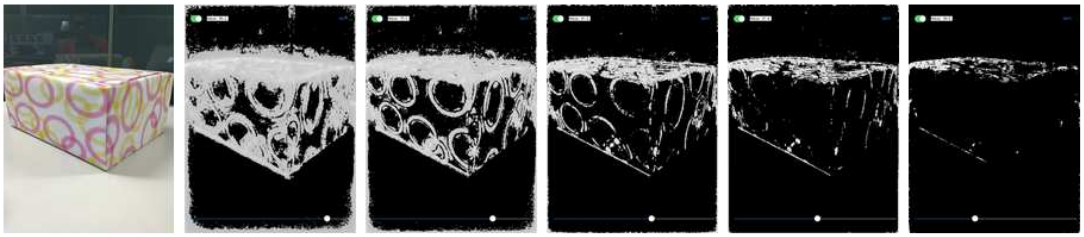

Figure 5.9. Saliency-based importance map as threshold is increased using my slider.

2.4.3[1] to detect Canny edges. To generate that of the saliency-based AR X-ray, I used the visual saliency tracker of Nick's Machine Perception Toolbox.[2]

MY AR X-ray applications run entirely on iPad mini tablets (A7 processor, 512MB DDR2 RAM, 16GB, 308 grams). I used the back camera (357x288 pixels, 10 fps) for sensing, and a 7.9 inch LED-display for the display. I used the standard user interface elements of iOS 7.

5.2.3.1 Thresholding Using Sliders

As a simple mechanism for controlling the amount of partial occlusions, I modify the thresholds for the edge-based and saliency-based AR X-ray techniques using an iOS slider. By increasing or decreasing a threshold, I can decrease or increase the amount of important regions, thereby decreasing or increasing the amount of occlusion. Figures 5.8 and 5.9 show the importance maps generated as the thresholds are increased.

[1] `http://docs.opencv.org/`

[2] `http://mplab.ucsd.edu/~nick/NMPT/`

The Canny edge detector already requires two threshold values for detecting the edges in order to generate the edge-based importance map: a low threshold and a high threshold such that 1) when the pixel gradient is higher than the high threshold, the pixel is treated as an edge; 2) when the pixel gradient is lower than the low threshold, the pixel is not an edge; 3) when the pixel gradient is between the thresholds, then it is treated as an edge only if it is adjacent to a pixel that passed condition 1. In my implementation, I fixed the high threshold to be twice the low threshold as recommended in the OpenCV documentation. Thus, I had only one value to control with the slider.

For thresholding the saliency-based importance map, I first normalized the importance map to have a range of values from 0 to 255. I then suppress to zero the values that are below the threshold.

5.2.4 Experiment

In my experiments, I asked participants to use my system for accomplishing visual tasks that involve a combination of common AR tasks, namely, distinguishing a virtual object and identifying it [119]. I designed two tasks that would generate insights on the appropriate levels of occlusion which does not hamper the understanding of AR X-ray visualization.

5.2.4.1 Participants

I envision that in the near future, AR X-ray will be integrated with handheld devices, such as smartphones and tablets. As such, I chose my participants to be regular smartphone and tablet users. I had a maximum sample size of 14 subjects (10 male, 4 female) with a mean age of 26 (SD = 2). All of them are daily smartphone users. Aside from smartphones, the subjects use tablet computers regularly. Four uses a tablet daily, three uses a tablet a few times a week, and two uses a tablet a few times a month. According to the demographics survey, 12 of the participants have used some form of AR technology before, however only two had previously experienced AR X-ray.

Table 5.3. Summary of Variables

	Independent Variable	Dependent Variable
Task 1	Edge-based AR X-Ray (E) Saliency-based AR X-Ray (S)	amount of partial occlusion (APO)
Task 2	Edge-based AR X-Ray (E) Saliency-based AR X-Ray (S) Alpha Blending (A)	alpha value (α) object identification time (OIT)

5.2.4.2 Variables

The dependent variables in this study were amount of partial occlusion (APO), alpha value (α), and object identification time (OIT). The independent variables were the type of occlusion presented, namely, edge-based AR X-ray (E), saliency-based AR X-ray (S), and alpha blending (A). Alpha blending serves as a benchmark for the two AR X-ray methods. Table 5.3 summarizes the independent and dependent variables for Tasks 1 and 2.

5.2.4.3 Instruments

Each independent variable in Table 5.3 corresponds to an iOS application. In total, five different iOS applications were developed for an iPad mini for this study. Two applications correspond to E and S for measuring the level of tolerance in Task 1 and three applications (corresponding to E, S, and A) for measuring alpha value and identification time in Task 2. Each application consisted of one of the selected AR X-ray visualization methods to see through a real object. Although the real object was captured live using the device's camera, a predefined set of images representing the content of the boxes were used.

For Task 2, I fixed the APO to 60/100 for the two X-ray methods. I chose this threshold by asking two participants (not part of the 14) to perform a task similar to task 1. However, instead of using Landolt C's, I used the objects in Figure 5.13.

I prepared a table with a tablet computer on it. The device was positioned close to the edge with its camera facing the table and its display facing a user

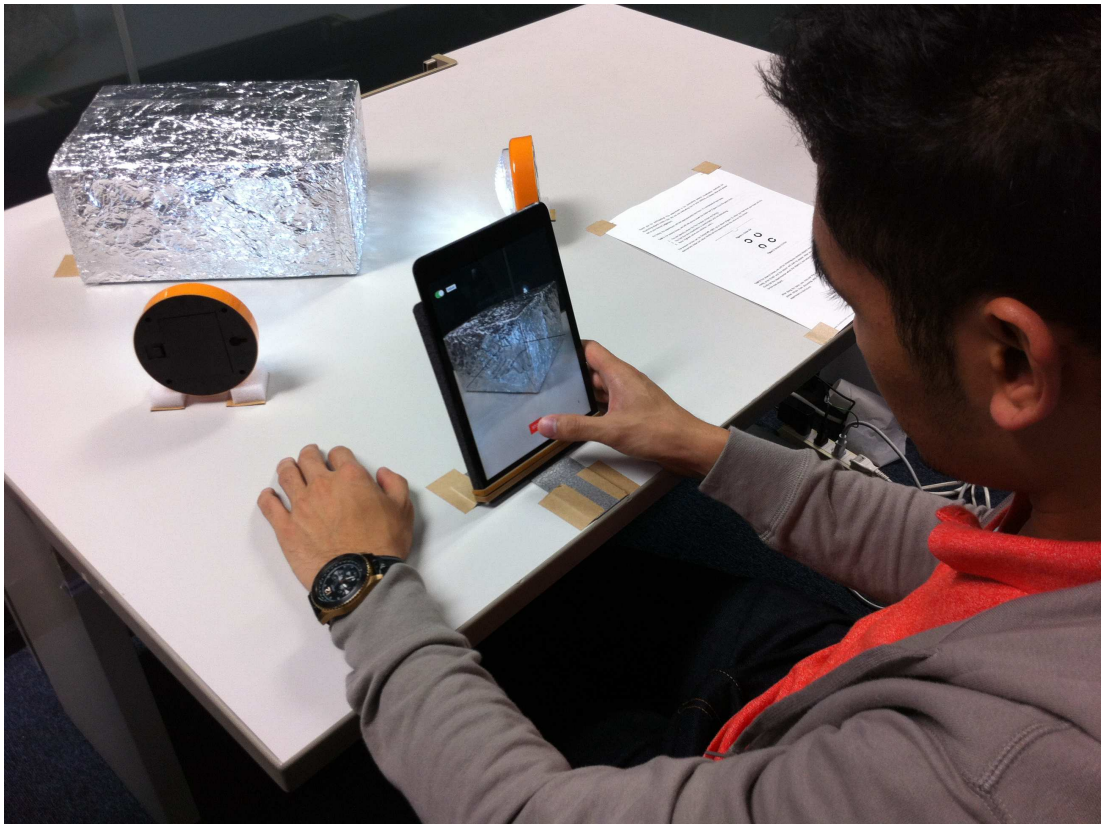

Figure 5.10. User Study Set Up

sitting on a chair (Figure 5.10). I prepared a set of six boxes of the same size (10.6 in x 4.7 in x 7.7 in) with different textures to simulate occlusion conditions that challenge the AR X-ray methods. The boxes were selected to feature different kinds of lighting (Figure 5.11 e-h), edges (Figure 5.11 c-d) and colors (Figure 5.11 a-b). Light sources were also fixed on the table next to the marked position for the boxes. Two of the boxes (sil and crum) were used twice (with and without the lights on), totaling the number of box set ups to eight. Box set ups Red, Brown, Green, and Crum were chosen to challenge edge-based AR X-ray; whereas Pink, Sil, Sil-light, and Crum-light were chosen to challenge saliency-based AR X-ray.

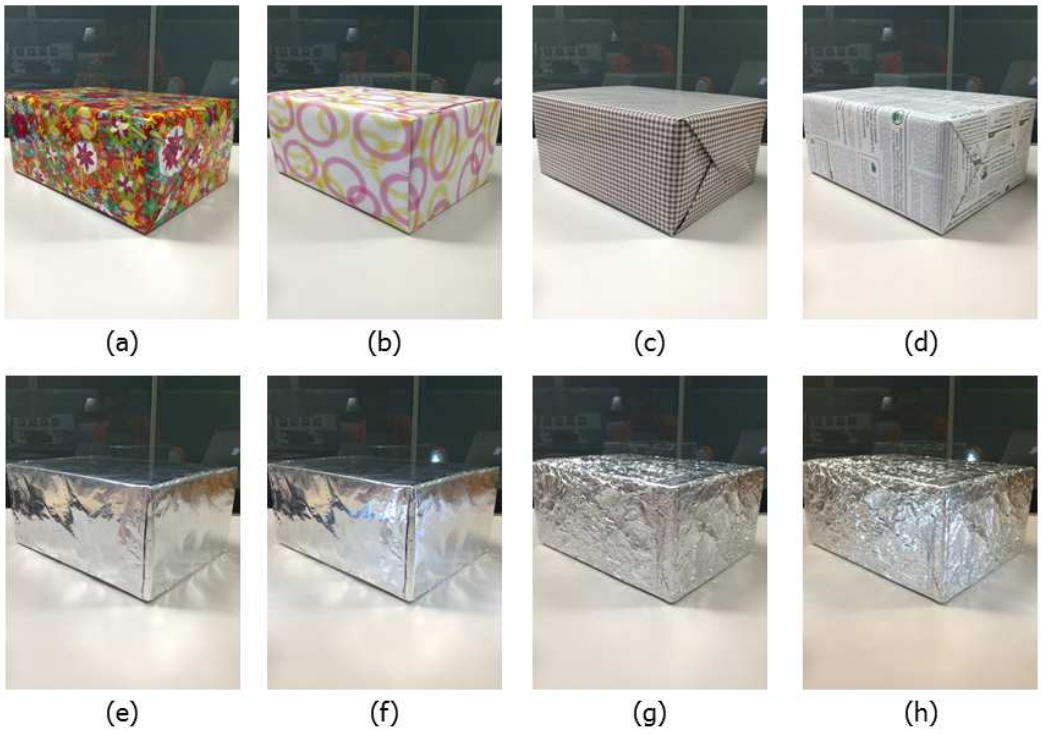

Figure 5.11. Box Set Ups: (a) Red, (b) Pink, (c) Brown, (d) Green, (e) Sil, (f) Sil-Light, (g) Crum, and (h) Crum-Light.

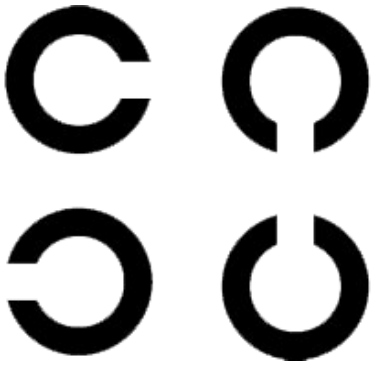

Figure 5.12. Landolt Cs

5.2.4.4 Procedures

Prior to the experiment, I informally screened the subjects based on their vision health. I excluded subjects whose vision was impaired for any reason. I informed the subjects about the content of the research, and I asked them if they are willing to participate. Participants were then presented with two tasks.

Task 1 measured the level of tolerance to occlusions when using the two X-ray methods E and S. I asked the participants to slowly lower the threshold value using a slider. Half of the participants used E first, whereas the other half used S first, with a 5 minute break in between. I instructed them to decrease the threshold until they can finally see clearly all four Landolt C's (Figure 5.12). The participants performed this task twice for each of the eight box set ups. For each time, the Landolt C's were in two sizes, big (21x21 pixels) and small (7x7 pixels). I asked the participants to first adjust the scale, and then say the value corresponding to the threshold level. The values ranged from 0 to 100; 0 has the minimum occlusion whereas 100 has the maximum occlusion.

Task 2 aimed to measure the alpha value and the time taken to identify objects through the two X-ray methods (E and S) and a benchmark (A). For Task 2, subjects had to identify two objects per box set up (one at a time) randomly picked from a predefined set of 23 (Figure 5.13) which were displayed in different positions within the box. I varied the order of A, E, and S for each participant with a 5 minute break in between.

In this task, a start button was presented at the beginning before every object

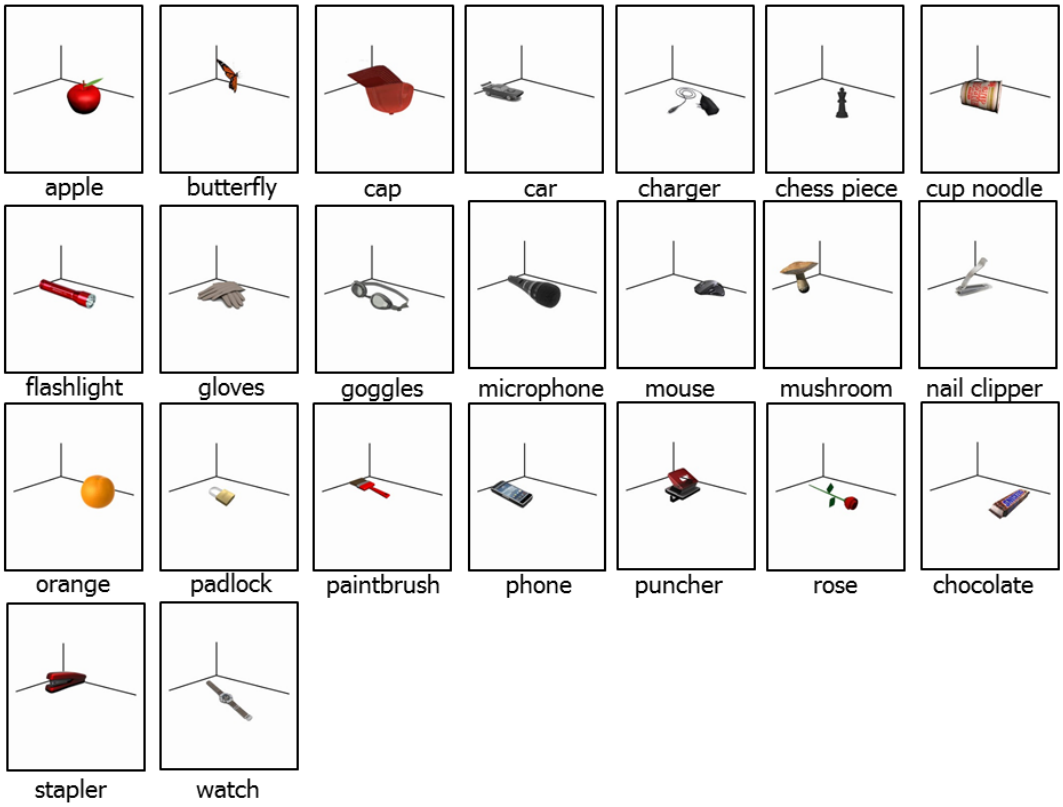

Figure 5.13. Objects for the Object Identification Task

could be identified. Once the start button had been pressed, the parts of the box that did not constitute the occlusion had their alpha decreased over time (decrease of 0.02 per second), thus slowly revealing the object. Participants were asked to press the stop button and say what the object was, once they identified it. I then took note of the alpha value. Note that in this experiment, object identification time (OIT) in seconds, and alpha value (α) are related such that $\alpha = 100 - OIT * 0.02$.

5.2.4.5 Data Analysis

I conduct a within-subjects design for both Task 1 and Task 2. I try to minimize order effects (practice effect, fatigue effect, etc.) by varying the order of E, S, and A, and by having a 5 minute break in between treatments.

I gathered mean scores from 14 subjects in Task 1 and 11 subjects in Task 2. Two means scores were compared for Task 1 using a Paired-Samples t-Test to observe differences between the two X-ray methods E and S. Furthermore, I conducted in depth comparisons between means scores for different box set ups within the same method.

Three mean scores were analyzed for Task 2 using Repeated Measures ANOVA to observe differences among the three methods used (Alpha, Edges, and Saliency). For cases where ANOVA found a significant difference, post hoc tests using the Bonferroni correction were used to discover which specific means differed. All the statistical analyses were run using SPSS^{TM} statistical software at a 0.05 level of significance.

5.2.4.6 Hypotheses

H1. To be legible, smaller objects require less occlusion cues than bigger objects when using both X-ray methods.

H2. The edge-based AR X-ray will be less legible when there are many edges. This corresponds to box set ups Red, Brown, Green, and Crum.

H3. The saliency-based AR X-ray will be less legible for when there are high contrast colors and lighting. This corresponds to box set ups Pink, Sil, Sil-Light, and Crum-light.

H4. Objects will be identified faster when viewing through alpha blending than when viewing with the edge-based AR X-ray.

H5. Objects will be identified faster when viewing through alpha blending than when viewing with the saliency-based AR X-ray.

H6. Objects will be identified faster when viewing through the saliency-based AR X-ray than when viewing with the edge-based AR X-ray.

5.2.5 Results and Discussion

The first task is to recognize small targets. This task is aimed to represent judging small parts or details in the visualization. On the other hand, the second

Table 5.4. Overall APO for Big and Small Landolt Cs

	Size	N	Mean	Std. Dev.	p	Effect Size (d)
Overall - APO	Big	224	51	29	<0.001	0.62
	Small	224	33	27		

Note: APO ranges from 0 to 100. Higher APO means that more occlusion can be tolerated.

task is a higher-level task. From the abstract Landolt Cs, I moved to identifying meaningful virtual objects. The purpose of task 2 is to observe if partial occlusion prevents users from identifying the object inside the box.

In task 1, I have gathered a total of 448 responses from 14 participants. Each participant viewed two sizes of Landolt Cs for each of the two AR X-ray methods on each of the eight box set ups. In this experiment, I found evidence that support my hypotheses 1 and 2, but not 3. In task 2, I have gathered a total of 528 responses from 11 participants. Each participant viewed two objects in each of the eight box set ups using either the AR X-ray methods, or the alpha blending method. Results support my hypotheses 4 and 6, but not 5.

5.2.5.1 Comparison Based on Object Size

I separated the APO specified by the participants according to the size of the Landolt Cs. Overall, they indicated a significantly lower APO for the small Landolt Cs with a moderate effect size, as shown in Table 5.4. This supports my hypothesis 1 that smaller target objects would require less occlusion cues to be legible.

Although this result is not surprising, this result has an important implication for AR X-ray. Current AR X-ray methods extract occlusion cues regardless of the virtual object being viewed and the intention of the user. To improve the legibility of AR X-ray, future methods must consider either or both knowledge of the virtual object or the user's intention. For example, virtual objects can have an accompanying metadata that indicates which areas of the virtual object are small and important. With this information, the AR X-ray method can avoid

Table 5.5. APO for Big and Small Landolt Cs per Box Set Up

	Size	N	Mean	Std. Dev.	p	Effect Size (d)
Red - APO	Big	28	30	22	0.01	0.64
	Small	28	16	22		
Pink - APO	Big	28	70	26	<0.001	1.00
	Small	28	48	17		
Brown - APO	Big	28	39	18	0.03	0.63
	Small	28	27	20		
Green - APO	Big	28	66	22	0.61	0.11
	Small	28	64	14		
Sil - APO	Big	28	63	29	<0.001	1.05
	Small	28	33	28		
Sil-light - APO	Big	28	62	29	<0.001	0.97
	Small	28	33	31		
Crum - APO	Big	28	44	24	0.08	0.50
	Small	28	32	24		
Crum-light - APO	Big	28	36	22	<0.001	0.95
	Small	28	17	18		

Table 5.6. Overall APO for AR X-Ray Methods

	AR X-Ray	N	Mean	Std. Dev.	p	Effect Size (d)
Overall - APO	E	224	41	31	0.55	-0.06
	S	224	43	37		

occluding these parts. Another example would be to detect the area of interest by tracking the user's gaze. Similarly, the AR X-ray method could adapt by prioritizing the legibility of the area of interest.

Aside from incorporating knowledge of the virtual object and the user's intention in controlling partial occlusions, a straight-forward way to improve the legibility of AR X-ray is to control occlusion cues with user input. In my present work, I discussed in Section 5.2.3.1 how I implemented a simple slider for decreasing occlusion cues. Using such user input methods is not only easier to implement, this approach also considers that different users can tolerate different levels of occlusions. In my experiments, I did observe high variability based on the standard deviations listed in Table 5.4.

I can also see the same pattern for each box set up listed in Table 5.5. Except for Green and Crum set ups, users have indicated a significantly lower APO for the small Landolt Cs with a moderate to large effect size. Moreover, users have indicated different APO values for the different box set ups with high standard deviations.

5.2.5.2 Variation in Tolerable Occlusions

In this exploration, I selected various patterns on my box set ups to challenge the AR X-ray method. My results in Table 5.6 shows that participants indicated almost equal APOs for the two methods. In other words, given my thresholding method discussed in Section 5.2.3.1, the participants indicated around the same threshold value for the two AR X-ray methods. However, the users did not indicate the same APO for each box set up listed in Table 5.7.

Based on past empirical evaluations discussed in Section 5.2.1.3, researchers observed that edge-based AR X-ray becomes difficult to understand when there

Table 5.7. APO for AR X-Ray Methods per Box Set Up

	AR X-Ray	N	Mean	Std. Dev.	p	Effect Size (d)
Red - APO	E	28	19	24	0.25	-0.31
	S	28	26	21		
Pink - APO	E	28	61	20	0.43	0.21
	S	28	56	28		
Brown - APO	E	28	28	20	0.04	-0.51
	S	28	38	19		
Green - APO	E	28	63	20	0.52	-0.22
	S	28	67	17		
Sil - APO	E	28	53	35	0.27	0.31
	S	28	43	30		
Sil-light - APO	E	28	48	37	0.55	0.17
	S	28	42	34		
Crum - APO	E	28	33	27	0.15	-0.41
	S	28	43	22		
Crum-light - APO	E	28	25	23	0.57	-0.14
	S	28	28	21		

are many edges detected in the background. In my present work, I also observed that my edge-based AR X-ray was less legible than my saliency-based X-ray for box set ups with many edges. As shown in Table 5.7, participants indicated a lower APO for all boxes with many edges, namely, Red, Brown, Green, and Crum. However, it is only with the Brown box set up wherein the edge-based AR X-ray scored significantly lower with a moderate effect size. As such, only this result supports my hypothesis 2. I observed marginal significance in the Crum box set up. I believe this particular set up is challenging for both AR X-ray methods. Aside from the many edges, the crumpled foil has sufficiently bright regions which hampers the saliency-based AR X-ray.

Past empirical research indicates that saliency-based AR X-ray may suffer when high contrast color or bright lighting are concentrated in one area of the occluding object. For example, the pink box has thick pink and yellow rings on white background. Intuitively, the salient regions would be the entire pink and yellow areas. Another example would be the reflected light in sil-light and crum-light. The LED light caused an entire area to be bright. Intuitively, the salient regions would be these areas with high luminosity.

In my present work, I observed that my saliency-based AR X-ray was less legible than my edge-based X-ray for some of the box set ups with high contrast color or bright lighting. As shown in Table 5.7, participants indicated a lower APO for boxes Pink, Sil, and Sil-light, but not Crum-light. However, all the differences I found are not significant, thus I do not have support for my hypothesis 3. I only observed a small effect size for box set ups Pink, Sil, and Sil-light.

One improvement in the experimental design is to make an even more challenging lighting condition for saliency-based AR X-ray. However, this would also be unnecessary because such intense lighting conditions may not occur frequently in actual settings. In particular, I am interested in indoor applications, such as museums and classrooms where lighting is not as dynamic compared to outdoors.

5.2.5.3 Comparison of AR X-Ray Methods

I computed the mean α for each participant, for each method. As discussed in Section 5.2.4.2, α and OIT have an indirect relationship. Higher α means faster identification time, and vice versa. I summarized the means and standard de-

Table 5.8. Overall Descriptive Statistics for Task 2

Method	N	Mean (α)	Std. Dev.
A	11	74	7
E	11	59	10
S	11	72	4

Note: The α ranges from 0-100.

Table 5.9. Summary of One-way Repeated Measures ANOVA for Alpha Values

	df	df (Error)	F	p	Effect Size (η^2)
Overall	2	20	12.1	<0.001	0.55

Table 5.10. Overall Pairwise Comparisons for Alpha Values

I and J	Mean Diff. (I - J)	Std. Error	p
A and E	14	4	0.009
A and S	2	3	1.000
E and S	-13	3	0.006

viation of α in Table 5.8. Participants were faster in A, followed by S. I then conducted a repeated measures ANOVA to test these means scores. Mauchly's Test of Sphericity indicated that the assumption of sphericity had not been violated, $x^2(2) = 1.073$, p = 0.585.

There was a statistically significant difference in overall α among methods A, E, and S, F(2, 20) = 12.1, p<0.05 as shown in Table 5.9. Post hoc tests using the Bonferroni correction revealed that the mean difference between A and S was not significant. However, there was a significant difference between A and E, and E and S, as shown in Table 5.10. This supports my hypotheses 4 and 6, but not 5.

For the task of identifying an occluded object, participants using the saliency-based AR X-ray performed almost the same as when there are no preserved occlusion cues (alpha blending). However, I observed that edge-based AR X-ray significantly hampers object identification with a large effect size.

5.2.5.4 Object Identification for Different Box Set Ups

I explored deeper into the results of Task 2 by conducting analysis on each box set up. Mauchly's Test of Sphericity indicated that the assumption of sphericity had not been violated, except for the Green box set up $x^2(2) = 8.499$, p = 0.014. As such, I excluded the Green box set up in this discussion.

Table 5.11 shows the descriptive statistics per box set up. The results for each individual box set up is similar to the overall results. It follows the pattern of almost equal α for A and S, with E lower than both. Results for Red, Brown, Crum, and Crum-light are significant with a large effect size ($0.22 \leq \eta^2 \leq 0.46$) as shown in Table 5.12.

Post hoc tests using the Bonferroni correction identified significant differences between A and E, and between E and S. Looking at the pairwise comparisons in Table 5.13, E had a significantly lower α compared to A for Brown, Crum and Crum-light. E had a significantly lower α compared to S for the Brown box set up. Moreover, there is partially significant differences between E and A, and between E and S for the Red box set up.

Red, Brown, and Crum were among the four designs chosen particularly to challenge edge-based AR X-ray. These box set ups have plenty of edges that are kept as occlusion cues. Because of these edges, the occluded object becomes

Table 5.11. Descriptive Statistics for Each Box Set Up

Box Set Up	Method	N	Mean (α)	Std. Dev.
Red	A	11	55	11
	E	11	35	23
	S	11	53	12
Pink	A	11	81	16
	E	11	77	19
	S	11	80	11
Brown	A	11	70	15
	E	11	52	15
	S	11	72	10
Sil	A	11	84	5
	E	11	78	14
	S	11	82	6
Sil-light	A	11	80	9
	E	11	71	14
	S	11	76	10
Crum	A	11	72	8
	E	11	57	18
	S	11	73	12
Crum-light	A	11	67	10
	E	11	40	25
	S	11	57	16

Table 5.12. One-way Repeated Measures ANOVA for Alpha Values (α) for Each Box Set Up

	df	df (Error)	F	p	Effect Size (η^2)
Red	2	20	5.2	0.015	0.34
Pink	2	20	0.2	0.803	0.22
Brown	2	20	8.5	0.002	0.46
Sil	2	20	1.5	0.250	0.13
Sil-light	2	20	1.5	0.241	0.13
Crum	2	20	6.2	0.008	0.38
Crum-light	2	20	5.8	0.010	0.37

less legible. Thus, the participants needed significantly more time to identify the object. On the other hand, the saliency-based AR X-ray did not occlude as much for these four box set-ups.

The box set ups Pink, Sil, Sil-light, and Crum-light were designs chosen to challenge saliency-based AR X-ray. Among these four, the saliency-based AR X-ray had the biggest difference with the Crum-light box set up. However, this difference was not significant.

5.2.6 Summary of Part 2

Currently, many researchers are working on improving AR and its enabling technologies. AR X-ray is a useful visualization technique for many fields of application, including ubiquitous learning. To make successful applications, more user studies must be conducted to further understand how X-ray visualization methods affect depth perception and legibility of the virtual object.

In this chapter, I implemented two AR X-ray methods, namely, edge-based AR X-ray and saliency-based AR X-ray. To create an X-ray illusion, I select important regions of the occluding real object and render it over the occluded virtual object. Logically, the more occlusion cues are placed over the virtual annotation, the more difficult it will be to see the virtual object. As such, it is important to control a parameter to adjust the amount of occlusion cues.

Table 5.13. Summary of Pairwise Comparisons for Alpha Values (α) for Each Box Set Up

Box Set Up	I and J	Mean Diff. (I - J)	Std. Error	p
Red	A and E	20	8	0.113
	A and S	1	4	1.000
	E and S	-18	7	0.090
Pink	A and E	5	8	1.000
	A and S	1	8	1.000
	E and S	-3	7	1.000
Brown	A and E	18	6	0.033
	A and S	-2	5	1.000
	E and S	-20	5	0.011
Sil	A and E	7	4	0.396
	A and S	2	3	1.000
	E and S	-5	5	1.000
Sil-light	A and E	9	5	0.312
	A and S	4	5	1.000
	E and S	-5	6	1.000
Crum	A and E	15	5	0.037
	A and S	-1	5	1.000
	E and S	-16	6	0.056
Crum-light	A and E	27	9	0.042
	A and S	10	5	0.293
	E and S	-17	9	0.249

I implemented these two AR X-ray methods on an iPad mini, using OpenCV and NMPT computer vision libraries. Using these computer libraries, I compute an importance map which represents the parts of the occluder that will be rendered on top of the occluded. For my application, I added a function to adjust the sensitivity for generating the importance map. I allowed the user to control this sensitivity using a slider.

I used the prototypes to explore on the legibility of the two AR X-ray methods. My results confirmed that smaller objects should have less occlusion cues to be legible. I observed that my edge-based AR X-ray was less legible when there are too many edges on the occluding real object. On the other hand, the saliency-based AR X-ray was less legible when there are high contrasts in color or bright lighting. For identifying larger objects, saliency-based AR X-ray allowed the users to perform better than with the edge-based AR X-ray approach. Aside from automated adjustments, I recommend that future AR X-ray systems should have user inputs to adjust the amount of occlusion. This allows the user to fit the visualization according to his intentions and preferences.

CHAPTER 6

Conclusion and Recommendation

Handheld augmented reality (HAR) is an emerging technology in many application areas. To create effective HAR systems, different types of evaluations need to be conducted by researchers, developers, and designers. Currently, there are limited experiences and standards in conducting these evaluations. Conducting evaluations on AR systems is difficult because it relies on the expert knowledge of AR researchers.

In response to this problem, I developed a usability evaluation framework that helps non-AR experts analyze the possible problems in their HAR systems. I based this framework on my own analysis of previous HAR systems. From my analysis, I defined two usability constructs, namely, manipulability – the ease of handling the device, and comprehensibility – the ease of understanding the presented information. I then explain the unique manipulability and comprehensibility issues that arise from using handheld devices as AR platforms.

In Chapter 3, I explained my usability evaluation framework which explains ideas that can be applied to formative evaluations of HAR. Moreover, I developed the usability scale called the HARUS from my framework. HARUS is a questionnaire for summative evaluation of HAR systems. Using HARUS, non-AR experts can measure the usability, manipulability, and comprehensibility of their HAR

systems with their target users while performing specific tasks. Based on my evaluations, HARUS is valid and reliable.

I proceeded to apply my evaluation framework and usability scale to user studies around learning support. In Chapter 4, I conducted a summative evaluation of a HAR system for situated vocabulary learning called FlipPin. Aside from finding insights in usability, my evaluations showed some possible benefits of using HAR to the memorization and motivation of students. In Chapter 5, I investigated the use of AR X-ray running on handheld devices for near-field distances. This is an interesting use case because it can be applied to HAR systems for educational gaming, museum learning, etc. Based on my experiments, I found that the saliency-based X-ray may be more legible than edge-based for the near-field distances. Moreover, I found that users have varying tolerances to occlusion. Therefore, it is recommended to have some user input to adjust the amount of occlusions.

6.1. Lessons Learned

I reiterate some of the main lessons learned from this thesis as follows:

1. In Chapter 3, I have demonstrated that although manipulability and comprehensibility are related constructs, it is possible for a HAR system to suffer more from manipulability issues than comprehensibility issues, and vice versa. The two constructs sometimes correlate moderately, therefore they should be treated as separate constructs.

2. In Chapter 3, I learned that manipulability and comprehensibility may possibly be the main factors affecting usability of HAR systems; i. e., problems arise because handling the device is difficult most probably due to the new poses required by HAR systems. Moreover, problems arise due to the difficulties in understanding three-dimensional information on a small, two-dimensional screen. By observing only manipulability and comprehensibility, I can already approximate the usability of the HAR system.

3. From Chapter 2 and Chapter 4, I propose that all evaluations of HAR systems for learning support should report the system usability of the HAR

system involved, aside from the benefits claimed. Currently, HAR is an emerging technology and it is highly susceptible to manipulability and comprehensibility issues. We risk comparing an imperfect HAR system to more stable medium of education. Although the HAR systems of today are not yet perfect, we can already see possible improvements in memorization and motivation.

4. From Chapter 5, I was able to identify the important concerns of teachers and students when using AR X-ray. It is important to engage these user groups early in the design phase in order to know the potential problems that may hamper adoption of the HAR system in schools. In particular, it was the teachers who brought up the possible misinformation and difficulties in distinguishing the virtual objects with AR X-ray running on handheld devices.

6.2. Limitations

My experiments evaluating the HARUS questionnaire were focused on using HAR systems for near-field interaction. I think this is the more relevant use case for learning support systems. As such, I do not have experiments in the middle-field and far-field distances. Many HAR applications work in middle-field and far-field distances, such as navigation and tourism. My evaluation framework and usability scale should also be verified in the middle-field and far-field distances.

I proposed ideas and concepts for formative and summative evaluations, but not for guidelines-based evaluation. Guidelines-based evaluation requires design guidelines that take longer time to form. This is because forming such guidelines requires synthesizing accumulated experiences of AR experts. In Chapter 7, I share my efforts towards synthesizing design guidelines for HAR.

6.3. Ongoing Work

In this thesis, I presented my evaluation framework and usability scale which aim to guide formative and summative evaluations of HAR systems. Currently,

it is still difficult to conduct guidelines-based evaluations because there are still limited established design guidelines for HAR. For future work, it's important to gather design guidelines for HAR and HAR systems for learning support. In Section 6.3.1, I discuss my efforts towards establishing design guidelines for HAR.

Aside from creating questionnaires, such as the HARUS, the concepts introduced in my evaluation framework are useful for analyzing sensor logs. As an example, in Section 6.3.2, I explored the use of acceleration data logs in estimating the usability ratings of a HAR system. Acceleration logs may contain information on how easily users can manipulate a HAR system to accomplish a task. The way the device accelerates in different directions may be telling of a HAR system's manipulability.

6.3.1 Toward Design Guidelines for HAR

Developing systems using emerging technology, such as augmented reality, is difficult because there are limited guidelines to inform developers during the design process. In particular, there are no established guidelines for learning support systems based on HAR. To gather such design guidelines, I first summarize existing guidelines for HAR in other fields of application. I then provide my synthesis of these guidelines into five design guidelines. I share my own experience of how I observed these guidelines in developing FlipPin, which is discussed further in Chapter 4. I then propose an additional guideline based on my experience.

6.3.1.1 Background

Designing effective user interfaces using emerging technologies is challenging because there are no existing design guidelines or interaction metaphors [61]. Experienced developers rely on best guesses and intuition which novice developers have yet to develop. In some cases, developers propose completely new ways for users to perceive and interact with information. Thus, there is limited prior experience to inform the developer during the development process. To address this challenge, it is important to gather and synthesize prior experiences into design guidelines.

Augmented reality (AR) is an emerging technology that may be useful for ed-

ucation [11]. In AR, virtual information is presented on the real environment as if it coexists with real objects. It enables many compelling experiences in science education, language learning, history, and culture, etc. Among the many forms of AR, HAR may be the easiest to deploy because of the increasing availability of handheld devices in schools. Although some design guidelines for HAR application exists, these guidelines were formed around more mature application areas of AR, such as navigation and tourism. There are limited design guidelines for developing HAR for learning support.

To gather design guidelines that may be applicable to learning support, I summarize existing guidelines for HAR applications. I then provide my synthesis of these guidelines and explain how I applied these to the design of FlipPin a HAR system for learning new vocabulary. Based on my observations, I suggest that these guidelines are also applicable for learning support. Moreover, I recommend one more guideline for further investigation.

6.3.1.2 Related Work

Gabbard and Swan [61] explain that design guidelines are important to inform the development process. When design guidelines are not available, developers need to conduct user studies to help guide their design. These user studies must be made as general as possible so that the findings could also be applied in other scenarios. Eventually, these individual findings are accumulated into design guidelines and standards. Figure 6.1 shows how user studies help both the formation of design guidelines and development of a particular interface. Similar to Gabbard and Swan, I try to design my user studies with FlipPin in Chapter 4 to have wider generalization.

Gabbard [59] lists a comprehensive collection of design guidelines found in the virtual reality and AR literature from 1987 to 1999. Most of the guidelines focus on AR using head-mounted displays as the presentation device. The guidelines include insights on many aspects of AR systems, such as visual feedbacks, tracking user location and orientation, data gloves and gesture recognition, users and user tasks, object selection and manipulation, etc.

Improvements in handheld devices (camera, processing power, large screen, etc.) and tracking and rendering algorithms have enabled developers to create AR

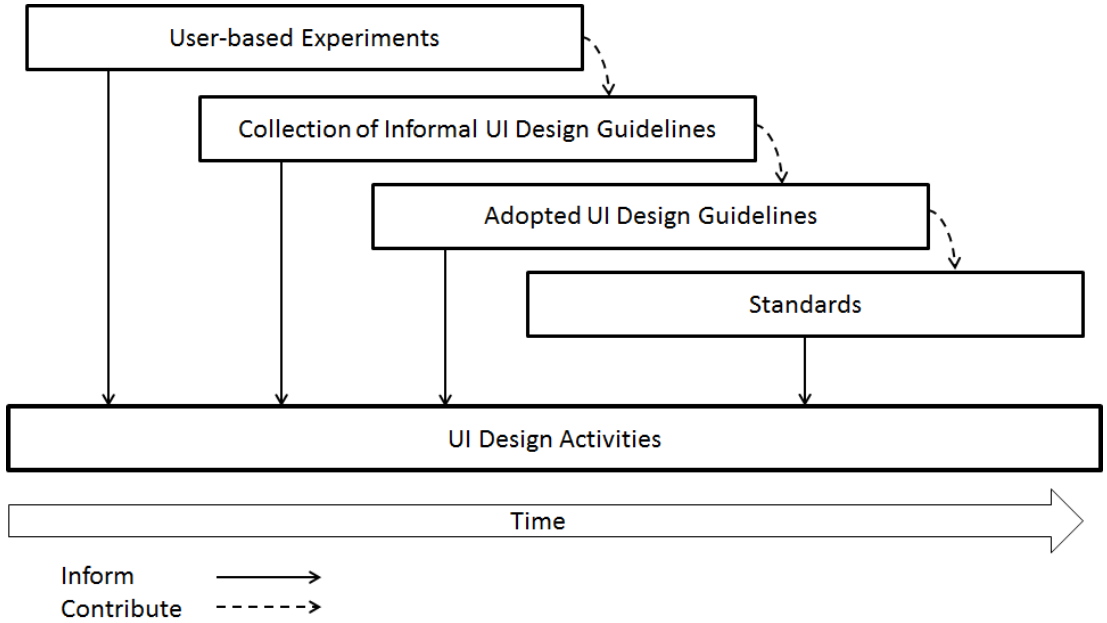

Figure 6.1. Gabbard and Swan's Diagram for the Development of Design Guidelines and Standards for User Interfaces (UI)

applications running on smartphones and tablet computers. Some of the design guidelines from HMD-based AR may apply to HAR. However, HAR also has different usability issues that arise from the use of handheld devices, as discussed in Chapter 3. Some design guidelines for HAR application exist. However, these guidelines were formed around more mature application areas of AR, such as tourism [99], navigation [98] and games [186].

6.3.1.3 Summary of Design Guidelines

Tables 6.1, 6.2 and 6.3 list the design guidelines proposed by Kourouthanassis et al. [99], Ko et al. [98] and Wetzel et al. [186] based on their experiences in making HAR applications for tourism, navigation and gaming, respectively. I found a total of 23 guidelines including 5 for tourism, 6 for navigation, and 12 for games. Although these guidelines are developed around specific types of commercial application, there are several overlaps that may be true for many HAR applications. For example, finding a specific place with the help of HAR

Table 6.1. Design Guidelines for HAR in Tourism

Application	Design Guidelines
Tourism	T1. Provide context-aware content by understanding the users' context using sensing and marker technologies.
	T2. Provide relevant content by allowing the user to personalize, expand, or limit the presented information.
	T3. Protect the user's privacy.
	T4. Provide user feedback on the status of the application.
	T5. Use familiar icons and interaction metaphors.

Table 6.2. Design Guidelines for HAR in Navigation

Application	Design Guidelines
Navigation	N1. Allow users to navigate hidden virtual information by operating the camera.
	N2. Limit the amount of information using the user's context, search result ranking, and/or user input.
	N3. Use familiar icons consistently and provide quick support to clarify icon meanings.
	N4. Allow users to modify the breadth of their search.
	N5. Provide a help menu for HAR features.
	N6. Support operations using only one hand.

is common for all these application areas. Another example would be the use of intuitive icons and menu navigation, which is also applicable to non-AR handheld applications. I summarize the 23 design guidelines into five design guidelines that I think may be applicable to HAR in learning support.

6.3.1.3.1 Present Context-aware Content

AR is essentially a context-aware technology by its definition of presenting virtual objects or digital information onto a real environment (the context). As such, guidelines G3 and G4 in Table 1 emphasize on the purposeful use of the real environment. Moreover, developers should manage the presentation of vir-

Table 6.3. Design Guidelines for HAR in Games

Application	Design Guidelines
Games	G1. Focus on game design by designing the game experience first before deciding the technologies required for implementation.
	G2. Stick to the theme of the game by selecting technologies that are relevant to the time period and ambience.
	G3. Make the user interact with a combination of real and virtual objects.
	G4. Situate the game in meaningful environments, rather than simply placing virtual objects in arbitrary space.
	G5. Keep the interaction simple.
	G6. Allow users to easily share their experience.
	G7. Encourage interaction with other players, non-players and virtual characters.
	G8. Show the real environment by managing the virtual objects to not block the entire view of the real environment.
	G9. Use potential technical problems as game elements, thus part of the gaming experience.
	G10. Adapt, not directly convert, games from other formats to HAR.
	G11. Add meaningful virtual content that contributes to the overall game experience.
	G12. Select the most appropriate tracking method for your target game.

tual elements so that they do not obstruct the view of the real environment, as suggested in G8. Aside from the location as the context, developers can also detect and use other contexts, such as time, user intentions, etc. By considering the context, guidelines T1 and N2 suggests that we can deliver more relevant content.

6.3.1.3.2 Provide Content Controls

Aside from automatically managing content based on the users' context, HAR applications should provide ways for the user to adjust the amount and quality of the content. HAR applications are susceptible to presenting too much information, leading to cluttered screens. To address this, T2, N4, and G8 suggest that applications should allow users to hide, expand or personalize the presented content. For content hidden from the current view, N1 recommends to have hidden content be accessible via camera movement, such as appearing/disappearing depending on where the camera is pointing.

6.3.1.3.3 Preempt Technical Difficulties

As an emerging technology, HAR is susceptible to many perceptual and ergonomic errors. Although AR technology is mature for several applications, AR researchers are still improving its related technologies like tracking, sensor fusion, and graphics rendering. Developers should compensate for this error by providing feedback to users on the current status of the application, suggested by T4 in Table 6.1. For example, T4 recommends informing the user about the loading time of virtual data and if there is tracking instability. In some areas, it is possible to mask technical difficulties, such as including it in the gaming experience, as recommended in G9. G12 recommends choosing the tracking method that would work best for the application. Lastly, N5 suggests having a help menu to assist users with common errors.

6.3.1.3.4 Preserve Intuitive Icons and Menus

Icons and menus still apply for HAR applications. T5 and N3 suggest the use of familiar icons and menu structure, such as those from WIMP interfaces. In

general, G5 suggests keeping the operations simple because we are dealing with a smaller screen compared to desktop computers. For novel icons and menus, T4 and N5 recommend features to assist users in operating the system.

6.3.1.3.5 Promote Social Interactions

Aside from using HAR to support intuitive interaction between the users and the real environment, AR should support interactions among users, and between users and other people, as recommended by G7 in Table 6.3. Moreover, G6 suggests that HAR applications should provide ways for users to easily share their experiences whether face-to-face or through digital means of communication.

6.3.1.4 The FlipPin Application

I developed a HAR application called FlipPin which aims to teach new vocabularies on a real environment. To use FlipPin, users point the handheld device to objects marked by fiducial markers. Then, three-dimensionally registered sprite sheet animations illustrate the action of a verb. Users can hear proper pronunciations by pressing the listen button and read the translation of the target word by pressing the translate button. The application runs on iPad tablet computers and uses the ARToolKit for tracking. For more details, I discussed my design, implementation and user studies further in Chapter 4.

6.3.1.4.1 Designing FlipPin

I tried to observe the five design principles discussed in Section 6.3.1.3 in developing FlipPin. First, I present context-aware content by applying three-dimensionally registered content. For example, in Figure 6.2, I illustrate the music playing ("spielen" is German for "to play") as virtual musical notes emerging from a real CD player.

I provide content controls by rendering the content for the closest fiducial marker only. The content then switches to the next content when the user points the handheld device to a different marker. Moreover, I provide controls for toggling the text panels on and off. I made the text panels transparent to minimize obstruction of the view of the real world, while keeping the texts legible.

Figure 6.2. The FlipPin Interface (left) and the Real Environment (right)

I preempt technical difficulties by using fiducial marker-based tracking instead of point cloud based tracking. Point cloud-based tracking may be unstable to use for this type of scene with many movable individual objects as shown in Figure 6.2, right. Moreover, in this scenario, the fiducial markers points the user to the real objects that are linked with virtual content.

I preserve intuitive icons and menus by using the interface elements of the iPad, such as buttons and labels. Keeping the iPad interface elements allows users to apply their prior knowledge of using the iPad. However, instead of buttons with text labels, graphical icons may be more familiar for the users.

Finally, I promote social interactions by locating the content in a place where people could study and chat with each other. In my user studies with FlipPin, the real environments that I used were an office (Figure 6.2, right) and a refreshment area where people eat snacks. I observed that even after using the HAR system, users would tend to discuss the content related to the objects marked by fiducial markers.

Based on my experience of developing FlipPin, I think that the five design guidelines that I derived from guidelines inspired by other fields of application are also useful for making applications for learning support. Such guidelines are

important to inform developers of HAR applications given the developing nature of AR technology. Aside from these five design guidelines, I propose the following guideline for further investigation:

6.3.1.4.2 Pay Attention to Manipulability

Manipulability refers to the ease of handling the device when operating a HAR application, as discussed in Chapter 3. One of the unique features of HAR is that it expects the user to handle and pose the handheld device in unconventional ways. I recommend limiting the amount of virtual information presented through AR to prevent fatigue. The rest of the information can be presented using more conventional display methods for handheld devices. I also recommend having interactions that allow users to rest before proceeding to the next subtask. For FlipPin, the users pointed the device to real objects in less than 20 seconds. Then, they put the device down and repeat the word to themselves. In addition, N6 in Table 6.2 recommends supporting one-handed operations.

6.3.1.5 Recommendations

Guidelines are important in designing effective HAR applications. However, in learning support, there are limited design guidelines to inform developers. In response, I synthesized five design guidelines based on other researchers' experiences of designing HAR applications for tourism, navigation, and gaming. I then explain how I applied the five guidelines to my own application for learning support, and recommend an additional design guideline. Based on my experience, I think these guidelines are helpful in designing HAR applications for learning support.

In this paper, I offer the six design guidelines for further investigation. For easier memorization, the six design guidelines could be referred to as the Six Ps. The Six Ps are: Present context-aware content, Provide content controls, Preempt technical difficulties, Preserve intuitive icons and menus, Promote social interactions, and Pay attention to manipulability. I hope that these design guidelines would be also helpful to other developers, specifically for those who are beginners in HAR development. Currently, my guidelines focus on usability

and easing cognitive load which is important in learning support. To improve on these guidelines, I plan to compare it with existing guidelines for non-AR handheld applications for learning support. I expect this improvement to grow our understanding of the best practices in the field. We can then modify the Six Ps or add some new guidelines. I also plan to continue developing FlipPin and other learning support systems that use AR. Through user studies, we can find possible improvements on the interface, as well as contribute to the growing knowledge on HAR design, particularly in the field of learning support.

6.3.2 Toward Usability Evaluations from User's Movement

Usability evaluations are important to the development of AR systems. However, conducting large-scale longitudinal studies remains challenging because of the lack of inexpensive but appropriate methods. In response, I propose a method for implicitly estimating usability ratings based on readily available sensor logs. To demonstrate my idea, I explored the use of features of accelerometer data in estimating usability ratings in an annotation task. Results show that my implicit method corresponds with explicit usability ratings at 79% and 84%. These results should be investigated further in other use cases, with other sensor logs.

6.3.2.1 Background

Designing effective AR systems is challenging because there are limited established design guidelines. Often, researchers propose completely new ways of interaction between users and technology. As such, [61] explains that it is necessary to have a usability engineering approach that iteratively applies user studies to inform design. In practice, researchers ask a group of people to use a system. Often, they observe more explicit measures of usability such as errors, timing, etc. through videos, data logging, and expert observers. After using the system, users give their feedback through interviews, and questionnaires. In cross-sectional studies, a user would need to use systems and answer questions multiple times. In effect, conducting user studies requires significant amount of money, time, and manpower.

Lack of resources limits the scale and duration of user studies. In particular, it

is challenging to conduct large-scale longitudinal studies which are necessary for some application areas. For example, in industrial work support, many workers would use the system for weeks before we can observe improvements in collaboration and productivity. In learning support, multiple students need to use the system simultaneously during class hours. Moreover, learning needs to be observed over the duration of month-long courses.

In response, I explore whether in some situations, more implicit user studies can be conducted – specifically when handheld devices are involved. As an example, I estimate the usability of one specific function in AR – text annotation, using one specific sensor – the accelerometer.

6.3.2.2 Proposed Method

Unique usability issues arise when handheld devices, such as smartphones and tablets, are used for AR [181]. One issue that significantly influences the usability of HAR is its manipulability – the ease of handling the HAR system (Chapter 3). HAR requires the user to grip, move, and pose the device in unconventional ways. As such, the device's accelerometer data might contain implicit information on usability [156].

HAR systems for work support, learning support and other application areas may include common functions, such as annotating text, object positioning, etc. To estimate usability for each function, I propose the use of automatically generated sensor data, such as accelerometer data, gyroscope data, etc. that are logged while a user uses each function. To analyze this information, I recommend labelling a small part of this data set with manually gathered usability ratings thereby forming a gold standard. The usability rating from the rest of the data set can then be estimated by comparing it with the gold standard. Using this method, we can minimize the number of times users need to answer questionnaires explicitly.

6.3.2.3 Experiment

I demonstrate my proposed method for the text annotation function of a HAR system. This function introduces unconventional gestures with a tablet. In this scenario, users create a SLAM map by swinging the device from side-to-side.

They then create several three dimensionally-registered labels by pointing the device to the target object, tapping on the screen, and typing the text.

6.3.2.3.1 Platform

I implemented a HAR authoring tool for text annotations on real objects as shown in Figure 6.3. It runs on iPad tablets and it uses SLAM point clouds for tracking. To create the point cloud map, the user needs to move the device from side-to-side. After the system detects enough points, the user can add text annotations onto the scene. While the system is in use, it records accelerations in X, Y and Z directions in the background at 60 logs per second.

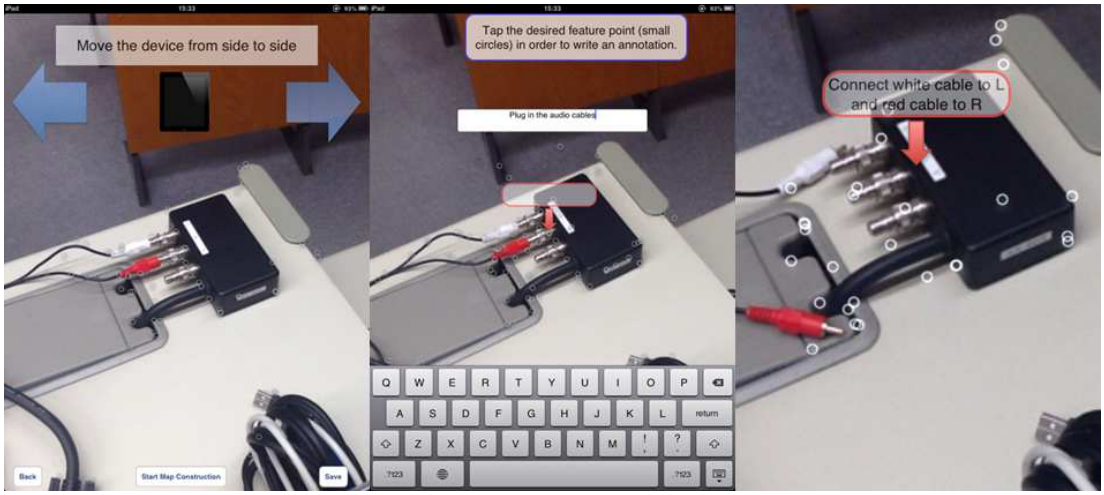

Figure 6.3. Screenshots of AR System for Annotating Text

6.3.2.3.2 Instruments

I use the System Usability Scale (SUS), a valid and reliable usability questionnaire. It aggregates usability ratings into a single score [107]. Based on this score, we can judge if a system is good enough for the target user to accomplish a specific task. In my work, usability refers to how well target users can use a functionality of a system to accomplish a specific task [135].

6.3.2.3.3 Procedures

I recruited 23 voluntary participants (22 to 42 years, 15 male, 8 female). All of them use handheld devices daily. Fourteen have experienced using an AR system at least once. I demonstrated to them how to use the system. I then asked them to add English translations to a Japanese rice cooker and trivia on a Philippine paper bill, as shown in Figure 6.4. I did not set a time limit and the participants could opt out during the experiment. After the task, I asked the participants to answer the SUS.

Figure 6.4. Scenario and Task

6.3.2.3.4 Data Analysis

For each participant's accelerometer log (X, Y, Z), I extracted time and frequency domain feature sets recommended in [147]. Table 6.4 lists the description of the features and the total number of features in the set. I then labeled each feature set as either having "good" or "bad" usability based on the interpretation of the SUS in [107]. For each labeled set of features, I trained J48 decision trees and random trees using WEKA 3.6.[1] I chose decision tree classifiers because they capture my intuition: If the user moves the device too quickly or too slowly, or if they frequently move the device in the wrong direction, they will perceive difficulty. Otherwise, they will find the system easy to use. Finally, I evaluate the tree models using leave-one-out cross validation [36].

I used leave-one-out cross validation because I only had 23 participants. My proposed method is for a big sample size N. For example, I can create a tree

[1] http://www.cs.waikato.ac.nz/ml/weka/

model from the first 22 participants, then use the model to estimate the usability ratings of the 23rd to the Nth participant. Using leave-one-out means that I take 22 participants to build the model, then estimate the usability rating of the remaining person. This is done iteratively wherein all 23 participants assume the role of the remaining person.

Table 6.4. Summary of Time and Frequency Domain Features

Features	Description	N
Mean and SD	Mean and standard deviation	6
Multiple Statistics	Mean, standard deviation, median, 25th and 75th percentile	15
Spectral Energy	Sum of squared FFT coefficients	3
FFT Magnitude	Magnitude of first five components of FFT analysis	15
Combination	Multiple statistics, spectral energy and FFT magnitude	33

6.3.2.4 Results and Discussion

Four of the 23 accelerometer logs were missing due to logging malfunction. As such, I only have 19 samples. Seven rated the system "good," whereas 12 rated "bad." As shown in Table 6.5, the first four feature sets were either moderately lower or higher than 50%. However, combining these time and frequency domain features boosted the accuracy of classification by around 20%. Based only on accelerometer data, I can estimate the usability ratings of users in a simple authoring scenario at a rate of 79–84%. To improve the accuracy, other features of accelerometer data can be explored. Moreover, using device orientation logs and navigation logs may also contribute to better accuracy.

Table 6.5. Correctly Classified Instances (%)

Features	J4.8 Tree	Random Tree
Mean and SD	53%	58%
Multiple Statistics	42%	68%
Spectral Energy	47%	58%
FFT Magnitude	58%	63%
Combination	**79%**	**84%**

6.3.2.5 Recommendations

I demonstrated my proposed method for estimating usability ratings in one specific use case with one specific sensor that is readily available in smartphones and tablets. With this approach, we can inexpensively conduct large-scale longitudinal studies with HAR systems. Such studies are necessary for generating insights on how we can improve AR systems and leverage them more effectively. It is important to further investigate my results in other use cases with other sensor logs. In addition, estimating more challenging classifications offered by the SUS, such as letter grades A to F must be investigated.

Publication List

Book Chapter

1. **Marc Ericson C. Santos**, Mitsuaki Terawaki, Takafumi Taketomi, Goshiro Yamamoto and Hirokazu Kato. Development of Handheld Augmented Reality X-Ray for K-12 Settings, In *Smart Learning Environments, Lecture Notes in Educational Technology*, Maiga Chang and Yanyan Li, Eds. Springer Berlin Heidelberg, pp. 199–220, 2015. (Chapter 5)

Journal Articles

1. **Marc Ericson C. Santos**, Jarkko Polvi, Takafumi Taketomi, Goshiro Yamamoto, Christian Sandor and Hirokazu Kato. Towards Standard Usability Questionnaires for Handheld Augmented Reality, *IEEE Computer Graphics & Applications*, 35(5): 50–59, 2015. (Chapter 3)

2. **Marc Ericson C. Santos**, Angie Chen, Takafumi Taketomi, Goshiro Yamamoto, Jun Miyazaki and Hirokazu Kato. Augmented Reality Learning Experience: Survey of Prototype Design and Evaluation, *IEEE Transactions on Learning Technology*, 7(1): 38–56, 2014. (Chapter 2)

3. Jayzon F. Ty, Ma. Mercedes T. Rodrigo and **Marc Ericson C. Santos**.

A Mobile Authoring Tool for AR Generation Using Images as Annotations, *Philippine Information Technology Journal*, 7(1): 61–70, 2014.

Peer-Reviewed Conference Publications

1. **Marc Ericson C. Santos**, Takafumi Taketomi, Goshiro Yamamoto, Gudrun Klinker, Christian Sandor and Hirokazu Kato. [POSTER] Towards Estimating Usability Ratings of Handheld Augmented Reality Using Accelerometer Data, Poster in *Proceedings of IEEE International Symposium on Mixed and Augmented Reality*, pp. 1–2, in press. (Chapter 7)

2. **Marc Ericson C. Santos**, Arno in Wolde Lübke, Takafumi Taketomi, Goshiro Yamamoto, Ma. Mercedes T. Rodrigo, Christian Sandor and Hirokazu Kato. Evaluating Augmented Reality for Situated Vocabulary Learning, In *Proceedings of APSCE International Conference on Computers in Education*, pp. 701–710, Nara, Japan, December 2014. (Chapter 4)

3. **Marc Ericson C. Santos**, Jayzon F. Ty, Arno in Wolde Lübke, Ma. Mercedes T. Rodrigo, Takafumi Taketomi, Goshiro Yamamoto, Christian Sandor and Hirokazu Kato. Authoring Augmented Reality as Situated Multimedia, Poster in *Proceedings of APSCE International Conference on Computers in Education*, pp.554–556, Nara, Japan, December 2014.

4. **Marc Ericson C. Santos**, Jarkko Polvi, Takafumi Taketomi, Goshiro Yamamoto, Christian Sandor and Hirokazu Kato. A Usability Scale for Handheld Augmented Reality, In *Proceedings of ACM Symposium on Virtual Reality Software and Technology*, pp. 167–176, Edinburgh, United Kingdom, November 2014. (hapter 3)

5. **Marc Ericson C. Santos**, Angie Chen, Mitsuaki Terawaki, Goshiro Yamamoto, Takafumi Taketomi, Jun Miyazaki and Hirokazu Kato. Augmented Reality X-ray Interaction in K-12 Education, In *Proceedings of IEEE International Conference on Advanced Learning Technologies*, pp. 141–145, Beijing, China, July 2013. (Chapter 5)

6. **Marc Ericson C. Santos**, Goshiro Yamamoto, Takafumi Taketomi, Jun Miyazaki and Hirokazu Kato. Authoring Augmented Reality Learning Experiences as Learning Objects, Poster in *Proceedings of IEEE International Conference on Advanced Learning Technologies*, pp. 506–507, Beijing, China, July 2013.

7. **Marc Ericson C. Santos**, Goshiro Yamamoto, Mitsuaki Terawaki, Jun Miyazaki, Takafumi Taketomi and Hirokazu Kato. Towards Participatory Design for Contextual Visualization in Education Using Augmented Reality X-ray. Poster in *Proceedings of ACM Augmented Human International Conference*, p. 241, Stuttgart, Germany, March 2013. (Chapter 5)

Other Conference and Workshop Publications or Presentations

1. **Marc Ericson C. Santos**, Takafumi Taketomi, Goshiro Yamamoto, Ma. Mercedes T. Rodrigo, Christian Sandor and Hirokazu Kato. Toward Guidelines for Designing Handheld Augmented Reality in Learning Support, In *Workshop Proceedings of APSCE International Conference on Computers in Education*, pp. 1–6, in press. (Chapter 7)

2. Resty Collado and **Marc Ericson C. Santos**. Design of a Handheld-based Motion Graphing Application for Physics Classes, In *Workshop Proceedings of APSCE International Conference on Computers in Education*, pp. 1–10, in press.

3. Shalika Pathirathna, Liem Hoang, **Marc Ericson C. Santos**, Sirawat Pitaksarit, Goshiro Yamamoto, Takafumi Taketomi, Christian Sandor and Hirokazu Kato. A Learning Support System using Camera Tracing for High School Physics Experiments. In *Proceedings of IPSJ Special Interest Group on Entertainment Computing Workshop*, pp. 1–4, Izumo, Japan, June 2015. (Chapter 7)

4. Jayzon F. Ty, Ma. Mercedes T. Rodrigo and **Marc Ericson C. Santos**. A Mobile Authoring Tool for Augmented Reality Content Generation Us-

ing Images as Annotations. In *Proceedings of CSP Philippine Computing Science Congress*, pp. 251–259, Davao, Philippines, March 2014.

5. **Marc Ericson C. Santos**, Angie Chen, Takafumi Taketomi, Goshiro Yamamoto, Jun Miyazaki and Hirokazu Kato. Inherent Advantages of Augmented Reality for K-12 Education: A Review of Augmented Reality Learning Experiences, In *USB Proceedings of Korea-Japan Workshop on Mixed Reality*, Okinawa, Japan, April 2013. (Chapter 2)

6. Angie Chen, **Marc Ericson C. Santos**, Takafumi Taketomi, Goshiro Yamamoto, Jun Miyazaki and Hirokazu Kato. An Augmented Reality Application Supporting Observation in Experiential Learning Theory. In *USB Proceedings of Korea-Japan Workshop on Mixed Reality*, Okinawa, Japan, April 2013.

Acknowledgements

More than three years ago, Prof. Hirokazu Kato asked me, "What is your dream?" As a young Filipino, I was already living my dream of doing graduate studies under a top researcher. When I look back to the past three years, I can't help but be grateful to Kato-sensei because he encouraged me to dream further. Kato-sensei adheres to the policy of "not doing research for research sake." We should always strive to apply our talents to what is important. During my studies under Kato-sensei, I have witnessed many researchers trying to solve relevant problems. I would also like to thank Kato-sensei for his patience in guiding me with my research, despite my shortcomings. Someday, I hope I can also serve many people with my skills just like Kato-sensei.

I also indebted to my other mentors who guided me throughout my doctoral studies. I would like to thank Assoc. Prof. Christian Sandor for the amazing energy he brought to the lab. I improved a lot in conducting and communicating my research thanks to him. It was refreshing to have Chris-sensei in the lab because of his experiences from other labs in various countries like Germany and Australia. I would like to thank Asst. Prof. Goshiro Yamamoto and Asst. Prof. Takafumi Taketomi for supporting me beyond their duty. I am thankful to Yamamoto-sensei for being accommodating, especially during the last few months of my dissertation writing. My dissertation improved a lot because of our discussions about the comments of the other thesis committee members. I

am also thankful to Taketomi-sensei for his support, such as the time he helped me write funding applications in Japanese, and the time he accompanied me to a conference in Beijing. I would also like to thank my thesis committee members and co-supervisors, Prof. Naokazu Yokoya and Prof. Ma. Mercedes T. Rodrigo. Thank you for being patient in reviewing this dissertation and for giving me advice. Thank you Ma'am Didith for setting a good example of a researcher in the Philippine context. Lastly, I would like to thank Prof. Jun Miyazaki and Asst. Prof. Yuki Uranishi for their support when I was still starting this research, and Prof. Gudrun Klinker for her comments on my ongoing work.

My stay at the Interactive Media Design Lab have been enriching because of all the people I have met through this lab. I am thankful to the IMD lab members for providing a supportive environment for me to grow and to enjoy research life. To all the past and present IMD lab members, thank you for all the challenging times and fun moments we shared. I am thankful to my sempais, Jaakko Hyry, Henry Chu, Marina Oikawa, Igor de Souza Almeida, Atsushi Keyaki, Yuichiro Fujimoto and Muhammad Zeeshan Asghar for all their suggestions for my research. I am thankful to my co-authors, Igor, Arno in Wolde Lübke and Jarkko Polvi, for sharing their expertise on the studies we collaborated on. Arno also helped me in many ways like teaching me programming and helping me give an AR workshop in the Philippines. I would also like to thank Dan Liem Hoang, Sirawat Pitaksarit and Shalika Pathirathna for working on our educational AR application which we deployed in the Philippines.

I am grateful to my friends for keeping me sane during this period of intense work. I would like to thank:

Maricris Marimon, David Joseph Tan and Timothy Ong for putting up with my rants and for being there after all these years.

the Filipino sempais, Erlyn Manguilimotan, Gemalyn Abrajano, Ramon Francisco Mejia, Jovilyn Fajardo and Mary-Clare Dy for helping me adjust to Japan and NAIST.

Jane Louie Zamora and Gian Carlo Mayuga for our conversations over wine.

Jimson Ngeo and Lorlynn Mateo for taking care of me during my first months in Japan, and during other tough times.

Michael Joseph Tan and Lorenzo Fabian Dayrit for being my gym buddies and helping me stay healthy.

Marius Georgescu and Gabriela Georgescu for always inviting me to go out of NAIST. Marius also fixed my laptop after it crashed, partly because of writing this dissertation.

Jennifer Damasco Ty, Christian Deus Cayao and Chittaphone Phonharath for sharing the crazy job-hunting phase with me.

Enzo, Gian, Jen, and Jason Paul Cruz for helping proofread this dissertation.

Hattori-sensei, Keiko Inoue, Naoyuki Ishiga and Tsuyoshi Aida for teaching me about Japanese language and culture allowing me to enjoy Japan more.

the Filipino community, Get Drunk Friday, international students, and friends I play sports with, for all the times we enjoyed together.

Finally, I would like to thank my family, relatives and friends in the Philippines for their love and support. Every time I go home to the Philippines, I feel renewed to tackle my research. I would like to thank my parents, Emmanuel Santos and Melissa Santos, for understanding somehow my choice to leave home to pursue graduate studies. I would also like to thank my brothers, Emmanuel Santos Jr. and Raphael Santos for letting me talk about my research and helping me at times. EJ helped me with crafting my questionnaires and accessing some research articles. Rap drew the sprite sheet animations in Chapter 4 and edited videos for me so I can explain my research better.

When I look back to all these years, I could not have accomplished my dream of finishing my graduate studies without the help of many people. I am eternally indebted to them – my mentors, family and friends. More importantly, because of my experiences during my graduate studies under Kato-sensei and the IMD lab, I am able to dream a better vision for my life. Although there are many things in life that I am unsure of, I feel inspired and I hope that someday I could exemplify a life of service and excellence.

Bibliography

[1] Advanced learning technologies journal list. http://celstec.org/content/advanced-learning-technologies-journal-list. Accessed: August 2012.

[2] Intrinsic motivation inventory. http://www.selfdeterminationtheory.org/questionnaires/10-questionnaires/50. Accessed: August 2012.

[3] Zooburst: create your own interactive 3d pop-up books. http://www.zooburst.com/. Accessed: October 2012.

[4] Khalil Al-Mekhlafi, Xianpei Hu, and Ziguang Zheng. An approach to context-aware mobile chinese language learning for foreign students. In *Proceedings of International Conference on Mobile Business*, pages 340–346, 2009.

[5] Maryam Alavi. Computer-mediated collaborative learning: An empirical evaluation. *MIS quarterly*, 18(2):159–174, 1994.

[6] Leonard Annetta and Stephen Bronack. *Serious Educational Game Assessment: Practical Methods and Models for Educational Games, Simulations and Virtual Worlds*. Amsterdam, The Netherlands. Sense Publishers, 2011.

[7] Theodoros N. Arvanitis, Argeroula Petrou, James F. Knight, Stavros Savas, Sofoklis Sotiriou, Michael Gargalakos, and Elpida Gialouri. Human factors

and qualitative pedagogical evaluation of a mobile augmented reality system for science education used by learners with physical disabilities. *Personal and Ubiquitous Computing*, 13(3):243–250, 2009.

[8] Kikuo Asai, Hideaki Kobayashi, and Tomotsugu Kondo. Augmented instructions – a fusion of augmented reality and printed learning materials. In *Proceedings of IEEE International Conference on Advanced Learning Technologies*, pages 213–215, July 2005.

[9] Ben Avery, Christian Sandor, and Bruce H. Thomas. Improving spatial perception for augmented reality x-ray vision. In *Proceedings of IEEE Virtual Reality Conference*, pages 79–82, March 2009.

[10] Ronald T. Azuma. A survey of augmented reality. *Presence-Teleoperators and Virtual Environments*, 6(4):355–385, 1997.

[11] Jorge Bacca, Silvia Baldiris, Ramon Fabregat, Sabine Graf, and Kinshuk. Augmented reality trends in education: A systematic review of research and applications. *Journal of Educational Technology & Society*, 17(4):133–149, 2014.

[12] Michael Bajura, Henry Fuchs, and Ryutarou Ohbuchi. Merging virtual objects with the real world: Seeing ultrasound imagery within the patient. In *Proceedings of ACM SIGGRAPH Conference on Computer Graphics and Interactive Techniques*, volume 26, pages 203–210, 1992.

[13] Aaron Bangor, Philip T. Kortum, and James T. Miller. An empirical evaluation of the system usability scale. *International Journal of Human-Computer Interaction*, 24(6):574–594, 2008.

[14] Sylvie Barma and Sylvie Daniel. Mind your game, game your mind! mobile gaming for co-constructing knowledge. In *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications*, pages 324–334, 2011.

[15] Jennifer S. Beaudin, Stephen S. Intille, Emmanuel Munguia Tapia, Randy Rockinson, and Margaret E. Morris. Context-sensitive microlearning of

foreign language vocabulary on a mobile device. In *Ambient Intelligence*, pages 55–72. Springer, 2007.

[16] Mark Billinghurst and Andreas Dünser. Augmented reality in the classroom. *Computer*, 45(7):56–63, July 2012.

[17] Joe Blalock and Jacob Carringer. Augmented reality applications for environmental designers. In *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications*, pages 2757–2762, 2006.

[18] Tobias Blum, Valerie Kleeberger, Christoph Bichlmeier, and Nassir Navab. mirracle: An augmented reality magic mirror system for anatomy education. In *Proceedings of IEEE Virtual Reality Conference*, pages 115–116, March 2012.

[19] Jennifer M. Brill and Yeonjeong Park. Facilitating engaged learning in the interaction age taking a pedagogically-disciplined approach to innovation with emergent technologies. *International Journal of Teaching and Learning in Higher Education*, 20(1):70–78, 2008.

[20] John Seely Brown, Allan Collins, and Paul Duguid. Situated cognition and the culture of learning. *Educational Researcher*, 18(1):32–42, 1989.

[21] Erin Peters Burton, Wendy Frazier, Leonard Annetta, Richard Lamb, Rebecca Cheng, and Margaret Chmiel. Modeling augmented reality games with preservice elementary and secondary science teachers. *Journal of Technology and Teacher Education*, 19(3):303–329, 2011.

[22] Su Cai and Qian Song. Ar-based remote video learning system. In *Proceedings of IEEE International Conference on Wireless, Mobile and Ubiquitous Technology in Education*, pages 322–324, March 2012.

[23] Julie Carmigniani, Borko Furht, Marco Anisetti, Paolo Ceravolo, Ernesto Damiani, and Misa Ivkovic. Augmented reality technologies, systems and applications. *Multimedia Tools and Applications*, 51(1):341–377, 2011.

[24] George Chang, Patricia Morreale, and Padmavathi Medicherla. Applications of augmented reality systems in education. In *Proceedings of Society*

for Information Technology & Teacher Education International Conference, number 1, pages 1380–1385, 2010.

[25] Wen-Chih Chang, Te-Hua Wang, Freya H. Lin, and Hsuan-Che Yang. Game-based learning with ubiquitous technologies. *IEEE Internet Computing*, 13(4):26–33, 2009.

[26] William Chang, Qing Tan, and Fang Wei Tao. Multi-object oriented augmented reality for location-based adaptive mobile learning. In *Proceedings of IEEE International Conference on Advanced Learning Technologies*, pages 450–451, July 2010.

[27] Yuan-Jen Chang, Chin-Hsing Chen, Wen-Tzeng Huang, and Wei-Shiun Huang. Investigating students' perceived satisfaction, behavioral intention, and effectiveness of english learning using augmented reality. In *Proceedings of IEEE International Conference on Multimedia and Expo*, pages 1–6, July 2011.

[28] Chih-Ming Chen and Yen-Nung Tsai. Interactive augmented reality system for enhancing library instruction in elementary schools. *Computers & Education*, 2012.

[29] I-Jung Chen and Jung-Chuan Yen. Hypertext annotation: Effects of presentation formats and learner proficiency on reading comprehension and vocabulary learning in foreign languages. *Computers & Education*, 63:416–423, 2013.

[30] Tzung-shi Chen, Cheng-Sian Chang, Jeng-Shian Lin, and Hui-Ling Yu. Context-aware writing in ubiquitous learning environments. *Research and Practice in Technology Enhanced Learning*, 4(1):61–82, 2009.

[31] Yu-Chien Chen. A study of comparing the use of augmented reality and physical models in chemistry education. In *Proceedings of ACM International Conference on Virtual Reality Continuum and Its Applications*, pages 369–372, 2006.

[32] Christopher Coffin, Svetlin Bostandjiev, James Ford, and Tobias Hollerer. Enhancing classroom and distance learning through augmented reality. In *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications*, number 1, pages 1140–1147, 2010.

[33] Anna Grasielle Dionisio Correa, Irene Karaguillas Ficheman, Marilena do Nascimento, and Roseli de Deus Lopes. Computer assisted music therapy: A case study of an augmented reality musical system for children with cerebral palsy rehabilitation. In *IEEE International Conference on Advanced Learning Technologies*, pages 218–220, July 2009.

[34] Lee J. Cronbach and Paul E. Meehl. Construct validity in psychological tests. *Psychological Bulletin*, 52(4):281–302, 1955.

[35] Sebastien Cuendet, Quentin Bonnard, Son Do-Lenh, and Pierre Dillenbourg. Designing augmented reality for the classroom. *Computers & Education*, 68:557–569, 2013.

[36] Anthony Dalton and Gearóid OLaighin. Comparing supervised learning techniques on the task of physical activity recognition. *IEEE Journal of Biomedical and Health Informatics*, 17(1):46–52, 2013.

[37] Mattias Davidsson, David Johansson, and Katrin Lindwall. Exploring the use of augmented reality to support science education in secondary schools. In *Proceedings of International Conference on Wireless, Mobile and Ubiquitous Technology in Education*, pages 218–220, March 2012.

[38] Fred D. Davis and Viswanath Venkatesh. A critical assessment of potential measurement biases in the technology acceptance model: three experiments. *International Journal of Human-Computer Studies*, 45(1):19–45, 1996.

[39] David Dearman and Khai Truong. Evaluating the implicit acquisition of second language vocabulary using a live wallpaper. In *Proceedings of ACM SIGCHI Conference on Human Factors in Computing Systems*, pages 1391–1400, 2012.

[40] Chris Dede. The evolution of distance education: Emerging technologies and distributed learning. *American Journal of Distance Education*, 10(2):4–36, 1996.

[41] Chris Dede. Emerging technologies, ubiquitous learning, and educational transformation. In *Towards Ubiquitous Learning*, pages 1–8. Springer, 2011.

[42] Frank N. Dempster. Effects of variable encoding and spaced presentations on vocabulary learning. *Journal of Educational Psychology*, 79(2):162–170, 1987.

[43] Arindam Dey, Andrew Cunningham, and Christian Sandor. Evaluating depth perception of photorealistic mixed reality visualizations for occluded objects in outdoor environments. In *Proceedings of Symposium on 3D User Interfaces*, pages 127–128, 2010.

[44] Arindam Dey, Graeme Jarvis, Christian Sandor, and Gerhard Reitmayr. Tablet versus phone: Depth perception in handheld augmented reality. In *Proceedings of IEEE International Symposium on Mixed and Augmented Reality*, pages 187–196, 2012.

[45] Arindam Dey, Graeme Jarvis, Christian Sandor, Ariawan Kusumo Wibowo, and Ville-Veikko Mattila. An evaluation of augmented reality x-ray vision for outdoor navigation. In *Proceedings of International Conference on Artificial Reality and Telexistence*, pages 187–196, 2011.

[46] Arindam Dey and Christian Sandor. Lessons learned: Evaluating visualizations for occluded objects in handheld augmented reality. *International Journal of Human-Computer Studies*, 72(10–11):704–716, 2014.

[47] Ángela Di Serio, María Blanca Ibáñez, and Carlos Delgado Kloos. Impact of an augmented reality system on students' motivation for a visual art course. *Computers & Education*, 68:586–596, 2013.

[48] Matt Dunleavy and Chris Dede. Augmented reality teaching and learning. In *Handbook of research on educational communications and technology*, pages 735–745. Springer, 2014.

[49] Andreas Dünser, Mark Billinghurst, James Wen, Ville Lehtinen, and Antti Nurminen. Exploring the use of handheld ar for outdoor navigation. *Computers & Graphics*, 36(8):1084–1095, 2012.

[50] Andreas Dünser, Karin Steinbügl, Hannes Kaufmann, and Judith Glück. Virtual and augmented reality as spatial ability training tools. In *Proceedings of ACM SIGCHI New Zealand Chapter International Conference on Computer-Human Interaction: Design-Centered HCI*, pages 125–132, 2006.

[51] Darren Edge, Elly Searle, Kevin Chiu, Jing Zhao, and James A. Landay. Micromandarin: mobile language learning in context. In *Proceedings of ACM SIGCHI Conference on Human Factors in Computing Systems*, pages 3169–3178, 2011.

[52] Neven A. M. El Sayed, Hala H. Zayed, and Mohamed I. Sharawy. Arsc: Augmented reality student card. *Computers & Education*, 56(4):1045–1061, 2011.

[53] Carrie Demmans Epp. Mobile adaptive communication support for vocabulary acquisition. In *Proceedings of Artificial Intelligence in Education*, pages 876–879, 2013.

[54] Morten Fjeld, Daniel Hobi, Lukas Winterthaler, Benedikt Voegtli, and Patrick Juchli. Teaching electronegativity and dipole moment in a tui. In *Proceedings of IEEE International Conference on Advanced Learning Technologies*, pages 792–794, August 2004.

[55] Floyd J. Fowler and Carol Cosenza. *International Handbook of Survey Methodology*, chapter Writing Effective Questions, pages 136–159. Taylor & Francis, 2008.

[56] Yuichiro Fujimoto, Goshiro Yamamoto, Jun Miyazaki, and Hirokazu Kato. Relation between location of information displayed by augmented reality and user's memorization. In *Proceedings of ACM Augmented Human International Conference*, pages 7:1–7:8, 2012.

[57] Yuichiro Fujimoto, Goshiro Yamamoto, Takafumi Taketomi, Jun Miyazaki, and Hirokazu Kato. Relation between displaying features of augmented reality and user's memorization. *Transactions of the Virtual Reality Society of Japan*, 18(1):81–91, 2013.

[58] Thomas A. Furness. The super cockpit and its human factors challenges. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 30, pages 48–52. Sage, 1986.

[59] Joseph L. Gabbard. Researching usability design and evaluation guidelines for augmented reality (ar) systems, 2001.

[60] Joseph L. Gabbard, Deborah Hix, and J. Edward Swan. User-centered design and evaluation of virtual environments. *IEEE Computer Graphics and Applications*, 19(6):51–59, 1999.

[61] Joseph L. Gabbard and J. Edward Swan. Usability engineering for augmented reality: Employing user-based studies to inform design. *IEEE Transactions on Visualization and Computer Graphics*, 14(3):513–525, 2008.

[62] Joseph L. Gabbard, J. Edward Swan, Deborah Hix, Si-Jung Kim, and Greg Fitch. Active text drawing styles for outdoor augmented reality: A user-based study and design implications. In *Proceedings of IEEE Virtual Reality Conference*, pages 35–42, March 2007.

[63] Raphael Grasset, Tobias Langlotz, Denis Kalkofen, Markus Tatzgern, and Dieter Schmalstieg. Image-driven view management for augmented reality browsers. In *Proceedings of IEEE International Symposium on Mixed and Augmented Reality*, pages 177–186, 2012.

[64] Jens Grubert, Tobias Langlotz, and Raphael Grasset. Augmented reality browser survey. Technical report, Institute for Computer Graphics and Vision, University of Technology Graz, 2011.

[65] Jian Gu, Nai Li, and H. Been-Lirn Duh. A remote mobile collaborative ar system for learning in physics. In *Proceedings of IEEE Virtual Reality Conference*, pages 257–258, March 2011.

[66] Anders Henrysson, Mark Billinghurst, and Mark Ollila. Virtual object manipulation using a mobile phone. In *Proceedings of International Conference on Augmented Tele-existence*, pages 164–171, 2005.

[67] Kuei-Fang Hsiao. Using augmented reality for students health-case of combining educational learning with standard fitness. *Multimedia Tools and Applications*, 64(2):407–421, 2013.

[68] Kuei-Fang Hsiao and Habib F. Rashvand. Body language and augmented reality learning environment. In *Proceedings of IEEE International Conference on Multimedia and Ubiquitous Engineering*, pages 246–250, 2011.

[69] Hsin-Chun Hsieh, Chih-Ming Chen, and Chin-Ming Hong. Context-aware ubiquitous english learning in a campus environment. In *Proceedings of IEEE International Conference on Advanced Learning Technologies*, pages 351–353, 2007.

[70] Feng Huang, Yu Zhou, Yao Yu, Ziqiang Wang, and Sidan Du. Piano ar: A markerless augmented reality based piano teaching system. In *Proceedings of International Conference on Intelligent Human-Machine Systems and Cybernetics*, volume 2, pages 47–52, August 2011.

[71] Wenhao Huang, Wenyeh Huang, Heidi Diefes Dux, and Peter K. Imbrie. A preliminary validation of attention, relevance, confidence and satisfaction model-based instructional material motivational survey in a computer-based tutorial setting. *British Journal of Educational Technology*, 37(2):243–259, 2006.

[72] Maria Blanca Ibanez, Angela Di Serio, Diego Villaran, and Carlos Delgado Kloos. Experimenting with electromagnetism using augmented reality: Impact on flow student experience and educational effectiveness. *Computers & Education*, 71:1–13, 2014.

[73] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, 1998.

[74] Takahiro Iwata, Tetsuo Yamabe, and Tatsuo Nakajima. Augmented reality go: Extending traditional game play with interactive self-learning support. In *Proceedings of International Conference on Embedded and Real-Time Computing Systems and Applications*, pages 105–114, August 2011.

[75] Hyung-Keun Jee, Sukhyun Lim, JinYoung Youn, and Junsuk Lee. An immersive authoring tool for augmented reality-based e-learning applications. In *Proceedings of International Conference on Information Science and Applications*, pages 1–5, April 2011.

[76] Hyung-Keun Jee, Sukhyun Lim, Jinyoung Youn, and Junsuk Lee. An augmented reality-based authoring tool for e-learning applications. *Multimedia Tools and Applications*, 68(2):225–235, 2014.

[77] Tai Fook Lim Jerry and Cheng Chi En Aaron. The impact of augmented reality software with inquiry-based learning on students' learning of kinematics graph. In *Proceedings of International Conference on Education Technology and Computer*, volume 2, pages V2–1–V2–5, June 2010.

[78] Larry Johnson, Samantha Adams, and Malcolm Cummins. The 2005 horizon report. 2005.

[79] Larry Johnson, Samantha Adams, and Malcolm Cummins. The 2006 horizon report. 2006.

[80] Sam Joseph, Kim Binsted, and Dan Suthers. Photostudy: Vocabulary learning and collaboration on fixed & mobile devices. In *Proceedings of IEEE International Workshop on Wireless and Mobile Technologies in Education*, pages 1–5, 2005.

[81] Samuel R. H. Joseph and Maria Uther. Mobile devices for language learning: Multimedia approaches. *Research and Practice in Technology Enhanced Learning*, 4(1):7–32, 2009.

[82] Carmen Juan, Francesca Beatrice, and Juan Cano. An augmented reality system for learning the interior of the human body. In *Proceedings of IEEE International Conference on Advanced Learning Technologies*, pages 186–188, July 2008.

[83] Carmen Juan, Edith Llop, Francisco Abad, and Javier Lluch. Learning words using augmented reality. In *Proceedings of International Conference on Advanced Learning Technologies*, pages 422–426, July 2010.

[84] Carmen Juan, Giacomo Toffetti, Francisco Abad, and Juan Cano. Tangible cubes used as the user interface in an augmented reality game for edutainment. In *Proceedings of International Conference on Advanced Learning Technologies*, pages 599–603, July 2010.

[85] Denis Kalkofen, Erick Mendez, and Dieter Schmalstieg. Interactive focus and context visualization for augmented reality. In *Proceedings of IEEE International Symposium on Mixed and Augmented Reality*, pages 191–201, November 2007.

[86] Denis Kalkofen, Erick Mendez, and Dieter Schmalstieg. Comprehensible visualization for augmented reality. *IEEE Transactions on Visualization and Computer Graphics*, 15(2):193–204, 2009.

[87] Denis Kalkofen, Eduardo Veas, Stephanie Zollmann, Martin Steinberger, and Dieter Schmalstieg. Adaptive ghosted views for augmented reality. In *Proceedings of IEEE International Symposium on Mixed and Augmented Reality*, pages 1–9, October 2013.

[88] Amy M. Kamarainen, Shari Metcalf, Tina Grotzer, Allison Browne, Diana Mazzuca, M. Shane Tutwiler, and Chris Dede. Ecomobile: Integrating augmented reality and probeware with environmental education field trips. *Computers & Education*, 68:545–556, 2013.

[89] Hirokazu Kato and Mark Billinghurst. Marker tracking and hmd calibration for a video-based augmented reality conferencing system. In *Proceedings of IEEE International Workshop on Augmented Reality*, pages 85–94, 1999.

[90] Hannes Kaufmann. Construct3d: An augmented reality application for mathematics and geometry education. In *Proceedings of ACM International Conference on Multimedia*, pages 656–657, 2002.

[91] Hannes Kaufmann and Dieter Schmalstieg. Designing immersive virtual reality for geometry education. In *Proceedings of IEEE Virtual Reality Conference*, pages 51–58, March 2006.

[92] Hannes Kaufmann, Dieter Schmalstieg, and Michael Wagner. Construct3d: A virtual reality application for mathematics and geometry education. *Education and Information Technologies*, 5(4):263–276, 2000.

[93] Jens Keil, Michael Zöllner, Mario Becker, Folker Wientapper, Timo Engelke, and Harald Wuest. The house of olbrich – an augmented reality tour through architectural history. In *Proceedings of IEEE International Symposium on Mixed and Augmented Reality*, pages 15–18, October 2011.

[94] John M. Keller. Development and use of the arcs model of instructional design. *Journal of Instructional Development*, 10(3):2–10, 1987.

[95] Lucinda Kerawalla, Rosemary Luckin, Simon Seljeflot, and Adrian Woolard. making it real: exploring the potential of augmented reality for teaching primary school science. *Virtual Reality*, 10(3):163–174, 2006.

[96] Eric Klopfer and Josh Sheldon. Augmenting your own reality: Student authoring of science-based augmented reality games. *New Directions for Youth Development*, 2010(128):85–94, 2010.

[97] Eric Klopfer and Kurt Squire. Environmental detectivesthe development of an augmented reality platform for environmental simulations. *Educational Technology Research and Development*, 56(2):203–228, 2008.

[98] Sang Min Ko, Won Suk Chang, and Yong Gu Ji. Usability principles for augmented reality applications in a smartphone environment. *International Journal of Human-Computer Interaction*, 29(8):501–515, 2013.

[99] Panos E. Kourouthanassis, Costas Boletsis, and George Lekakos. Demystifying the design of mobile augmented reality applications. *Multimedia Tools and Applications*, 74(3):1045–1066, 2013.

[100] Jon A. Krosnick and Stanley Presser. *Handbook of Survey Research*, chapter Question and Questionnaire Design, pages 263–313. Emerald Group Publishing Limited, 2010.

[101] Ernst Kruijff, J. Edward Swan, and Steven Feiner. Perceptual issues in augmented reality revisited. In *Proceedings of IEEE International Symposium on Mixed and Augmented Reality*, pages 3–12, Octobers 2010.

[102] Stan Kurkovsky, Ranjana Koshy, Vivian Novak, and Peter Szul. Current issues in handheld augmented reality. In *Proceedings of International Conference on Communications and Information Technology*, pages 68–72, June 2012.

[103] Fusako Kusunoki, Masanori Sugimoto, and Hiromichi Hashizume. Symphony-q: a support system for learning music in collaborative learning. In *Proceedings of International Conference on Systems, Man and Cybernetics*, volume 4, pages 1–6, October 2002.

[104] Tobias Langlotz, Stefan Mooslechner, Stefanie Zollmann, Claus Degendorfer, Gerhard Reitmayr, and Dieter Schmalstieg. Sketching up the world: in situ authoring for mobile augmented reality. *Personal and Ubiquitous Computing*, 16(6):623–630, 2012.

[105] Gun A. Lee, Andreas Dünser, Seungwon Kim, and Mark Billinghurst. Cityviewar: A mobile outdoor ar application for city visualization. In *Proceedings of IEEE International Symposium on Mixed and Augmented Reality*, pages 57–64, November 2012.

[106] Kangdon Lee. Augmented reality in education and training. *TechTrends*, 56(2):13–21, 2012.

[107] James R. Lewis and Jeff Sauro. The factor structure of the system usability scale. In *Human Centered Design*, pages 94–103. Springer, 2009.

[108] Nai Li, Leanne Chang, Yuan Xun Gu, and Henry Been-Lirn Duh. Influences of ar-supported simulation on learning effectiveness in face-to-face collaborative learning for physics. In *Proceedings of IEEE International Conference on Advanced Learning Technologies*, pages 320–322, July 2011.

[109] Qi-Ming Li, Yi-Min Chen, De-Yi Ma, Chen Huang, Sheng Xu, Ren-Miao Lu, Yan Liu, and Wang Xi-Chen. Design and implementation of a chinese character teaching system based on augmented reality interaction technology. In *Proceedings of IEEE International Conference on Computer Science and Automation Engineering*, pages 322–326, June 2011.

[110] Yiqun Li, Aiyuan Guo, Jimmy Aaddison Lee, Yan Gao, and Yii Leong Ling. Visual interactive and location activated mobile learning. In *Proceedings of International Conference on Wireless, Mobile and Ubiquitous Technology in Education*, pages 235–237, March 2012.

[111] Chih Cheng Lin and Hsien Sheng Hsiao. The effects of multimedia annotations via pda on efl learners vocabulary learning. In *Proceedings of APSCE International Conference on Computers in Education*, 2011.

[112] Chih Cheng Lin and Ying Chieh Wu. The effects of different presentation modes of multimedia annocations on sentential listening comprehension. In *Proceedings of APSCE International Conference on Computers in Education*, 2013.

[113] Chih Cheng Lin and Ya Chuan Yu. Efl learners cognitive load of learning vocabulary on mobile phones. In *Proceedings of APSCE International Conference on Computers in Education*, 2012.

[114] Chiu-Pin Lin, Shelley Chwu-Ching Young, and Hui-Chun Hung. The game-based constructive learning environment to increase english vocabulary acquisition: Implementing a wireless crossword fan-tan game (wicfg) as an example. In *Proceedings of IEEE International Conference on Wireless, Mobile, and Ubiquitous Technology in Education*, pages 205–207, 2008.

[115] Pei-Hsun Emma Liu and Ming-Kuan Tsai. Using augmented-reality-based mobile learning material in efl english composition: An exploratory case study. *British Journal of Educational Technology*, 44(1):E1–E4, 2013.

[116] Tsung-Yu Liu. A context-aware ubiquitous learning environment for language listening and speaking. *Journal of Computer Assisted Learning*, 25(6):515–527, 2009.

[117] Tsung-Yu Liu, Tan-Hsu Tan, and Yu-Ling Chu. 2d barcode and augmented reality supported english learning system. In *Proceedings of International Conference on Computer and Information Science*, pages 5–10, July 2007.

[118] Wei Liu, Adrian David Cheok, Charissa Lim Mei-Ling, and Yin-Leng Theng. Mixed reality classroom: learning from entertainment. In *Proceedings of International Conference on Digital Interactive Media in Entertainment and Arts*, pages 65–72, 2007.

[119] Mark A. Livingston. Evaluating human factors in augmented reality systems. *IEEE Computer Graphics and Applications*, 25(6):6–9, 2005.

[120] Mark A Livingston, Arindam Dey, Christian Sandor, and Bruce H Thomas. Pursuit of x-ray vision for augmented reality. In *Human Factors in Augmented Reality Environments*, pages 67–107. Springer, 2013.

[121] Dorit Maor. A teacher professional development program on using a constructivist multimedia learning environment. *Learning Environments Research*, 2(3):307–330, 1999.

[122] George Margetis, Panagiotis Koutlemanis, Xenophon Zabulis, Margherita Antona, and Constantine Stephanidis. A smart environment for augmented learning through physical books. In *Proceedings of IEEE International Conference on Multimedia and Expo*, pages 1–6, July 2011.

[123] Sergio Martin, Gabriel Diaz, Elio Sancristobal, Rosario Gil, Manuel Castro, and Juan Peire. New technology trends in education: Seven years of forecasts and convergence. *Computers & Education*, 57(3):1893–1906, 2011.

[124] Jorge Martín-Gutiérrez, José Luís Saorín, Manuel Contero, Mariano Alcañiz, David C Pérez-López, and Mario Ortega. Design and validation of an augmented book for spatial abilities development in engineering students. *Computers & Graphics*, 34(1):77–91, 2010.

[125] Shinya Matsutomo, Takenori Miyauchi, So Noguchi, and Hideo Yamashita. Real-time visualization system of magnetic field utilizing augmented reality technology for education. *IEEE Transactions on Magnetics*, 48(2):531–534, February 2012.

[126] Richard E. Mayer. Cognitive theory of multimedia learning. In Richard E. Mayer, editor, *Cambridge Handbook of Multimedia learning*. Cambridge University Press, 2005.

[127] Richard E. Mayer. *Multimedia Learning*. Cambridge University Press, 2009.

[128] Samuel Messick. Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9):741–749, 1995.

[129] Samuel Messick. Validity of test interpretation and use. Technical report, Educational Testing Service, 1998.

[130] Paul Milgram and Fumio Kishino. A taxonomy of mixed reality visual displays. *IEICE Transactions on Information and Systems*, 77(12):1321–1329, 1994.

[131] Hiroyuki Mitsuhara, Yoneo Yano, and Toshiyuki Moriyama. Paper-top interface for supporting note-taking and its preliminary experiment. In *Proceedings of IEEE International Conference on Systems Man and Cybernetics*, pages 3456–3462, 2010.

[132] Yoichi Motokawa and Hideo Saito. Support system for guitar playing using augmented reality display. In *Proceedings of IEEE International Symposium on Mixed and Augmented Reality*, pages 243–244, October 2006.

[133] Alessandro Mulloni, Hartmut Seichter, and Dieter Schmalstieg. Handheld augmented reality indoor navigation with activity-based instructions. In *Proceedings of ACM International Conference on Human Computer Interaction with Mobile Devices and Services*, pages 211–220, 2011.

[134] Alessandro Mulloni, Hartmut Seichter, and Dieter Schmalstieg. User experiences with augmented reality aided navigation on phones. In *Proceedings of IEEE International Symposium on Mixed and Augmented Reality*, pages 229–230, October 2011.

[135] Jakob Nielsen. *Usability Engineering*. Elsevier, 1994.

[136] Jakob Nielsen and Robert L. Mack. *Usability Inspection Methods*. Wiley & Sons, Inc. New York, 1994.

[137] Salaheddin Odeh, Shatha Abu Shanab, Mahasen Anabtawi, and Rami Hodrob. Remote augmented reality engineering labs. In *Proceedings of IEEE Global Engineering Education Conference*, pages 1–6, April 2012.

[138] Hiroaki Ogata, Mengmeng Li, Bin Hou, Noriko Uosaki, and Moushir M. El-Bishouty. Scroll: Supporting to share and reuse ubiquitous learning log in the context of language learning. *Research and Practice in Technology Enhanced Learning*, 6(2):69–82, 2011.

[139] Hiroaki Ogata, Toru Misumi, Tsuyoshi Matsuka, Moushir M. El-Bishouty, and Yoneo Yano. A framework for capturing, sharing and comparing learning experiences in a ubiquitous learning environment. *Research and Practice in Technology Enhanced Learning*, 3(3):297–312, 2008.

[140] Hiroaki Ogata and Yoneo Yano. Context-aware support for computer-supported ubiquitous learning. In *Proceedings of IEEE International Workshop on Wireless and Mobile Technologies in Education*, pages 27–34, 2004.

[141] Hiroaki Ogata, Chengiu Yin, Moushir M El-Bishouty, and Yoneo Yano. Computer supported ubiquitous learning environment for vocabulary learning. *International Journal of Learning Technology*, 5(1):5–24, 2010.

[142] Thomas Olsson and Markus Salo. Online user survey on current mobile augmented reality applications. In *Proceedings of IEEE International Symposium on Mixed and Augmented Reality*, pages 75–84, October 2011.

[143] Patrick O'Shea, Chris Dede, and Rebecca Mitchell. A study of the efficacy of an augmented reality curriculum. In *Proceedings of Society for Information Technology & Teacher Education International Conference*, pages 1510–1514, 2009.

[144] Kiran Pala, Anil Kumar Singh, and Suryakanth V. Gangashetty. Games for academic vocabulary learning through a virtual environment. In *Proceedings of IEEE International Conference on Asian Language Processing*, pages 295–298, 2011.

[145] Sobah Abbas Petersen, Jan-Kristian Markiewicz, and Sondre Skaug Bjornebekk. Personalized and contextualized language learning: Choose when, where and what. *Research and Practice in Technology Enhanced Learning*, 4(1):33–60, 2009.

[146] Stephen D. Peterson, Magnus Axholt, Matthew Cooper, and Stephen R. Ellis. Visual clutter management in augmented reality: Effects of three label separation methods on spatial judgments. In *Proceedings of IEEE Symposium on 3D User Interfaces*, pages 111–118, March 2009.

[147] Stephen J. Preece, John Yannis Goulermas, Laurence P. J. Kenney, and David Howard. A comparison of feature extraction methods for the classification of dynamic activities from accelerometer data. *IEEE Transactions on Biomedical Engineering*, 56(3):871–879, 2009.

[148] Jochen Prümper and Michael Anft. Die evaluation von software auf grundlage des entwurfs zur internationalen ergonomie-norm iso 9241 teil 10 als beitrag zur partizipativen systemgestaltung-ein fallbeispiel. In *Proceedings of Software-Ergonomie*, volume 93, pages 145–156, 1993.

[149] Whitney Queensberry. Using the 5es to understand users. http://wqusability.com/articles/getting-started.html. Accessed: August 2012.

[150] Rama B. Radhakrishna. Tips for developing and testing questionnaires/instruments. *Journal of Extension*, 45:1–4, 2007.

[151] Abu Saleh Md. Mahfujur Rahman, Jongeun Cha, and Abdulmotaleb El Saddik. Authoring edutainment content through video annotations and 3d model augmentation. In *Proceedings of IEEE International Conference on Virtual Environments, Human-Computer Interfaces and Measurements Systems*, pages 370–374, May 2009.

[152] Gerhard Reitmayr and Dieter Schmalstieg. Collaborative augmented reality for outdoor navigation and information browsing. In *Proceedings of Symposium on Location Based Services and TeleCartography*, pages 31–41, 2004.

[153] Jun Rekimoto. A magnifying glass approach to augmented reality systems. *Presence*, 6(4):399–412, 1997.

[154] Rafael Alves Roberto and Veronica Teichrieb. Arblocks: A projective augmented reality platform for educational activities. In *Proceedings of IEEE Virtual Reality Conference*, pages 175–176, March 2012.

[155] Young Sam Ryu and Tonya L. Smith-Jackson. Reliability and validity of the mobile phone usability questionnaire (mpuq). *Journal of Usability Studies*, 2:39–53, 2006.

[156] Alireza Sahami Shirazi, Niels Henze, Tilman Dingler, Kai Kunze, and Albrecht Schmidt. Upright or sideways?: Analysis of smartphone postures in the wild. In *Proceedings of the International Conference on Human-computer Interaction with Mobile Devices and Services*, pages 362–371, 2013.

[157] Patrick Salamin, Tej Tadi, Olaf Blanke, Frederic Vexo, and Daniel Thalmann. Quantifying effects of exposure to the third and first-person perspectives in virtual-reality-based training. *IEEE Transactions on Learning Technologies*, 3(3):272–276, 2010.

[158] Christian Sandor, Andrew Cunningham, Arindam Dey, and Ville-Veikko Mattila. An augmented reality x-ray system based on visual saliency. In *Proceedings of IEEE International Symposium on Mixed and Augmented Reality*, pages 27–36, 2010.

[159] Christian Sandor, Arindam Dey, Andrew Cunningham, Sebastian Barbier, Ulrich Eck, Donald Urquhart, Michael R. Marner, Graeme Jarvis, and Sang Rhee. Egocentric space-distorting visualizations for rapid environment exploration in mobile mixed reality. In *Proceedings of IEEE Virtual Reality Conference*, pages 47–50, 2010.

[160] Jeff Sauro. Measuring usability with the system usability scale (sus), 2011.

[161] Gerhard Schall, Erick Mendez, Ernst Kruijff, Eduardo Veas, Sebastian Junghanns, Bernhard Reitinger, and Dieter Schmalstieg. Handheld aug-

mented reality for underground infrastructure visualization. *Personal and Ubiquitous Computing*, 13(4):281–291, 2009.

[162] Richard F. Schmid, Robert M. Bernard, Eugene Borokhovski, Rana Tamim, Philip C. Abrami, Anne Wade, Michael A. Surkes, and Gretchen Lowerison. Technologys effect on achievement in higher education: a stage i meta-analysis of classroom applications. *Journal of Cmputing in Higher Education*, 21(2):95–109, 2009.

[163] Gerhard Schwabe and Christoph Göth. Mobile learning with a mobile game: design and motivational effects. *Journal of Computer Assisted Learning*, 21(3):204–216, 2005.

[164] Kristopher Scott and Rachid Benlamri. Context-aware services for smart learning spaces. *IEEE Transactions on Learning Technologies*, 3(3):214–227, 2010.

[165] Angela Di Serio, Maria Blanca Ibanez, and Carlos Delgado Kloos. Impact of an augmented reality system on students' motivation for a visual art course. *Computers & Education*, 68:586–596, 2013.

[166] Brett E. Shelton and Nicholas R. Hedley. Exploring a cognitive basis for learning spatial relationships with augmented reality. *Technology, Instruction, Cognition and Learning*, 1(4):323–357, 2004.

[167] Luca Simeone and Salvatore Iaconesi. Toys++ ar embodied agents as tools to learn by building. In *Proceedings of IEEE International Conference on Advanced Learning Technologies*, pages 649–650, July 2010.

[168] Luca Simeone and Salvatore Iaconesi. Anthropological conversations: Augmented reality enhanced artifacts to foster education in cultural anthropology. In *Proceedings of IEEE International Conference on Advanced Learning Technologies*, pages 126–128, July 2011.

[169] Aw Kien Sin and Halimah Badioze Zaman. Live solar system (lss): Evaluation of an augmented reality book-based educational tool. In *Proceedings of International Symposium in Information Technology*, volume 1, pages 1–6, June 2010.

[170] Sofoklis Sotiriou, Stamatina Anastopoulou, Sherman Rosenfeld, Osnat Aharoni, Avi Hofstein, Franz Bogner, Heie Sturm, and Kay Hoeksema. Visualizing the invisible: The connect approach for teaching science. In *Proceedings of IEEE International Conference on Advanced Learning Technologies*, pages 1084–1086, July 2006.

[171] Marcus Specht, Stefaan Ternier, and Wolfgang Greller. Mobile augmented reality for learning: a case study. *Journal Of The Research Center For Educational Technology*, 7(1):117–127, 2011.

[172] Stuart P. Stenton, Richard Hull, Patric M. Goddi, Josephine E. Reid, Ben J. C. Clayton, Tom J. Melamed, and Susie Wee. Mediascapes: Context-aware multimedia experiences. *IEEE MultiMedia*, 14(3):98–105, July–September 2007.

[173] Desi Dwistratanti Sumadio and Dayang Rohaya Awang Rambli. Preliminary evaluation on user acceptance of the augmented reality use for education. In *Proceedings of International Conference on Computer Engineering and Applications*, volume 2, pages 461–465, March 2010.

[174] J. Edward Swan and Joseph L. Gabbard. Perceptual and ergonomic issues in mobile augmented reality for urban operations, 2003.

[175] Stella Sylaiou, Katerina Mania, Athanasis Karoulis, and Martin White. Exploring the relationship between presence and enjoyment in a virtual museum. *International Journal of Human-Computer Studies*, 68(5):243–253, 2010.

[176] Rana M. Tamim, Robert M. Bernard, Eugene Borokhovski, Philip C. Abrami, and Richard F. Schmid. What forty years of research says about the impact of technology on learning a second-order meta-analysis and validation study. *Review of Educational Research*, 81(1):4–28, 2011.

[177] Wernhuar Tarng and Kuo-Liang Ou. A study of campus butterfly ecology learning system based on augmented reality and mobile learning. In *Proceedings of IEEE International Conference on Wireless, Mobile and Ubiquitous Technology in Education*, pages 62–66, March 2012.

[178] Thomas Tullis and William Albert. *Measuring the User Experience: Collecting, Analyzing, and Presenting Usability Metrics*. Morgan Kaufmann, 2008.

[179] Poonsri Vate-U-Lan. Augmented reality 3d pop-up children book: Instructional design for hybrid learning. In *Proceedings of IEEE International Conference on e-Learning in Industrial Electronics*, pages 95–100, November 2011.

[180] Eduardo Veas, Raphael Grasset, Ioan Ferencik, Thomas Grünewald, and Dieter Schmalstieg. Mobile augmented reality for environmental monitoring. *Personal and Ubiquitous Computing*, 17(7):1515–1531, 2013.

[181] Eduardo Veas and Ernst Kruijff. Vesp'r: Design and evaluation of a handheld ar device. In *Proceedings of IEEE International Symposium on Mixed and Augmented Reality*, pages 43–52, 2008.

[182] Eduardo Veas and Ernst Kruijff. Handheld devices for mobile augmented reality. In *Proceedings of International Conference on Mobile and Ubiquitous Multimedia*, page 3, 2010.

[183] Dirk Walther. *Interactions of visual attention and object recognition: computational modeling, algorithms, and psychophysics*. PhD thesis, California Institute of Technology, February 2006.

[184] Yuan Wang, Tobias Langlotz, Mark Billinghurst, and Tim Bell. An authoring tool for mobile phone ar environments. In *Proceedings of New Zealand Computer Science Research Student Conference*, volume 9, pages 1–4, 2009.

[185] Stuart Webb. The effects of repetition on vocabulary knowledge. *Applied Linguistics*, 28(1):46–65, 2007.

[186] Richard Wetzel, Rod McCall, Anne-Kathrin Braun, and Wolfgang Broll. Guidelines for designing augmented reality games. In *Proceedings of Conference on Future Play: Research, Play, Share*, pages 173–180. ACM, 2008.

[187] Sean White and Steven Feiner. Sitelens: Situated visualization techniques for urban site visits. In *Proceedings of ACM SIGCHI Conference on Human Factors in Computing Systems*, pages 1117–1120. ACM, 2009.

[188] Lung-Hsiang Wong and Chee-Kit Looi. Mobile-assisted vocabulary learning in real-life setting for primary school students: Two case studies. In *Proceedings of IEEE International Conference on Wireless, Mobile and Ubiquitous Technologies in Education*, pages 88–95, 2010.

[189] Chun-Cheng Wu, Chih-Wei Chang, Baw-Jhiune Liu, and Gwo-Dong Chen. Improving vocabulary acquisition by designing a storytelling robot. In *Proceedings of IEEE International Conference on Advanced Learning Technologies*, pages 498–500, 2008.

[190] Hsin-Kai Wu, Silvia Wen-Yu Lee, Hsin-Yi Chang, and Jyh-Chong Liang. Current status, opportunities and challenges of augmented reality in education. *Computers & Education*, 62:41–49, 2013.

[191] Tetsuo Yamabe, Hiroshi Asuma, Sumire Kiyono, and Tatsuo Nakajima. Feedback design in augmented musical instruments: A case study with an ar drum kit. In *Proceedings of International Conference on Embedded and Real-Time Computing Systems and Applications*, volume 2, pages 126–129, August 2011.

[192] Fang Chuan Ou Yang. Using personalized vls on mobile english vocabulary learning. In *Proceedings of IEEE International Conference on Wireless, Mobile and Ubiquitous Technology in Education*, pages 232–234, 2012.

[193] Mau-Tsuen Yang and Wan-Che Liao. Computer-assisted culture learning in an online augmented reality environment based on free-hand gesture interaction. *IEEE Transactions on Learning Technologies*, 7(2):107–117, April 2014.

[194] Norziha Megat Mohd. Zainuddin, Halimah Badioze Zaman, and Azlina Ahmad. A participatory design in developing prototype an augmented reality book for deaf students. In *Proceedings of International Conference on Computer Research and Development*, pages 400–404, May 2010.

[195] Stephanie Zollmann, Denis Kalkofen, Erick Mendez, and Gerhard Reitmayr. Image-based ghostings for single layer occlusions in augmented reality. In *Proceedings of IEEE International Symposium on Mixed and Augmented Reality*, pages 19–26, 2010.