Editorial Policy

The Australian Journal of Intelligent Information Processing Systems is an interdisciplinary forum for providing the latest information on research developments and related activities in the design and implementation of intelligent information processing systems.

The areas of interest include, but are not limited to:
- artificial intelligence
- artificial neural networks
- computer science
- fuzzy systems
- virtual reality

The journal both theoretical and application-oriented papers in diverse areas related to intelligent systems. The journal also publishes survey articles, short notes, Ph.D. thesis abstracts and project reports.

Editorial Board

Tom Gedeon  
Editor in Chief  
*Australian National University, Australia*

Yianni Attikiouzel  
*Murdoch University, Australia*

James C. Bezdek  
*University of Western Florida, USA*

Tony Constantinides  
*Imperial College of Science, Technology and Medicine, U.K.*

Peter Eklund  
*University of Wollongong, Australia*

Nikola Kasabov  
*Auckland University of Technology, New Zealand*

László Kóczy  
*Budapest University of Technology & Economics, Hungary*

Takeshi Furuhashi  
*Nagoya University, Japan*

Marimuthu Palaniswami  
*University of Melbourne, Australia*

M. V. Srinivasan  
*University of Queensland, Australia*

A. Venetsanopoulos  
*University of Toronto, Canada*

Responsibility of the content of the papers rests upon the authors and not with the publishers. Data and conclusions developed by the authors are for information only and are not intended for use without independent substantiating investigation on the part of the potential user.

The Australian Journal of Intelligent Information Processing Systems is published quarterly. All correspondence including manuscripts and advertising inquiries should be addressed to the Editor in Chief: Professor Tom Gedeon, Faculty of Engineering and Information Technology, The Australian National University, Canberra ACT 0200, Australia.

Abstracting is permitted with due credit to the Australian Journal of Intelligent Information Processing Systems.

Subscriptions:
Individual rate: A$80.
Institutional rate: A$220 for paper or online subscription.

Individual copies can be purchased at A$15 per single copy.

Reprints of technical articles are available, quantities no less than 50 can be ordered. Orders should be addressed to: AJIIPS, Faculty of Engineering and Information Technology, The Australian National University, Canberra ACT 0200, Australia. Email address: ajiips@cs.anu.edu.au.

1 Online subscription gives additional access to back issues without any fee.
CONTENTS

Designing a Network-Based System for Delivery of Remote Mine Services
Weidong Huang, Craig James, Kazys Stepanas, Eleonora Widzyk-Capehart, Leila Alem, Chris Gunn, Matt Adcock, Kerstin Haustein .............................................................. 1

A Case Study of Collecting Dynamic Social Data: The Pro-Ana Twitter Community
Ian Wood ................................................................................................................................. 13

Comparing eye gaze tracking to reported perceptions of manipulated and unmanipulated digital images
Sabrina B. Caldwell, Tamás D. Gedeon, Richard L. Jones, Martin Henscke ...... 26

Evaluation of e-Readers: A Preliminary Analysis
Ujala Rampaul, Tom Gedeon ................................................................................................. 37

This is a refereed journal.
Abstracted by INSPEC.

Publication Date December 2015
Designing a Network-Based System for Delivery of Remote Mine Services

Weidong Huang¹, Craig James², Kazys Stepanas², Eleonora Widzyk-Capehart², Leila Alem², Chris Gunn², Matt Adcock², Kerstin Haustein²

¹University of Tasmania, Australia
²CSIRO, Australia
{tony.huang, craig.a.james, kazys.stepanas, eleonora.widzyk-capehart, leila.alem, chris.gunn, matt.adcock, kerstin.haustein}@csiro.au

Abstract. There is a great body of work in the areas of tele-assistance/tele-collaboration offering novel and effective ways to improve collaboration between personnel located at a remote mine site and off-site personnel located in major metropolitan areas. Much of this work involves the use of high-bandwidth communications or targeted sensory experiences using large format displays. There are also existing remote access technologies but these suffer from limited functionality (providing text, voice, video or one-way desktop sharing), are often poorly supported in the security-conscious corporate environment and require complicated set up processes. There is currently no singular piece of remote collaboration technology that is suitable for the delivery of high-quality planning and scheduling services to clients at a mining site from a remote operating centre. In response to this issue, as part of a research and technology development effort between CSIRO and a mining engineering firm, we have developed a concept of remote mining engineer (RME), and conducted a functional requirements analysis, for delivering mining engineering services to mine sites remotely. Based on the obtained requirements, a further study was performed to characterise existing technologies and to identify the scope for future work in designing and prototyping a network based system for RME. In this paper, we report on the method and findings of this study.

1 Introduction

Due to the increasing demand for mineral commodities, increasingly complex technologies, machines, and practices are being introduced within mining operations to improve efficiency and productivity [11]. Specialized knowledge and expertise are often required to handle the technical complexities of mining equipment and operations [4, 13]. This leads to a higher demand for skilled and experienced personnel. More specifically, mining engineers play a critical role in on-site mine planning and operations but access to skilled staff willing to work in remote locations is difficult [8, 9]. While mining companies (service requesters) and service providers based in metropolitan centres are able to attract and retain top-level personnel, the need to ensure effective communications between off-site and on-site personnel, requires frequent trips to remote mining locations, which results in high travel burdens and high service costs. Further, after a site visit, the contracted mining engineer continues to engage with on-site personnel via emails or phone calls to ensure the services requested are delivered (such as short term mine planning). The time spent travelling to mine sites and collaborating through phone calls and emails is costly, time consuming and inefficient.

On the one hand, remote communication technologies such as tele-conferencing, desktop sharing and even regular phone and email services can be utilised as a means of communication (e.g., [1]).
However, individually, these technologies do not address the following key challenges in the open cut mining environment:

- **Quality of service** – effective remote communication relies on clear reception of as many cues as possible (text, tone, gesture, facial expressions) to avoid misunderstandings. Current applications of remote technologies do not provide a sufficient mix of these cues at a sufficient level of quality.

- **Low bandwidth** – broadband communications in remote areas is still very poor and the scale of information being shared (voice, text, data) has rapidly out-stripped available bandwidth.

- **High security** – great care must be taken to protect integrity of data and control systems where downtime from malicious intrusions can introduce high production penalties.

- **Usability** – available remote technologies are difficult to set up, configure and maintain, often requiring specialist resources and training.

On the other hand, there is a great body of work in the areas of tele-assistance/tele-collaboration that are specifically designed to improve collaboration between personnel located at a remote mine site and off-site personnel located in major metropolitan areas (e.g., [3, 12, 15]). Much of this work involves the use of high-bandwidth communications or targeted sensory experiences using large format displays (e.g., [4, 5]).

Generally, there is currently no singular piece of remote collaboration technology that is suitable for the delivery of high-quality planning and scheduling services to clients at a mining site from a remote operating centre. In response to this issue, the concept of Remote Mining Engineer (RME) has been investigated through collaboration between CSIRO and a mining engineering firm (service provider) [2]. The ultimate objectives of this collaboration project include:

- Facilitate collaboration between mining engineers inside the service provider and between staff of the service provider and the service requester. These participants are often physically distributed.

- Reduce the need for mining engineers to be present at mine sites without compromising the quality of mine planning and scheduling. This will be achieved by dramatically improving the communication ability between off-site and on-site personnel through technological innovations.

The customised RME system would combine existing (text, voice, visualisation and data sharing) and innovative communication technologies (tele-presence, tele-collaboration, tele-assistance, and immersive environments). These technologies will be combined in a way that improves collaboration and communication over long distances between on-site and off-site personnel, allowing service providers to deliver their mining engineering expertise to remote mine sites from their offices located in major metropolitan centres so that travel costs can be reduced without compromising the quality of the service. More specifically the system will rely on the following technologies:

- **Tele-presence technology** to enable a sense of physical presence of the remote mining engineer within mine site personnel.

- **Collaborative workspace** and **desktop/whiteboard sharing** to enable site personnel to: see what the mining engineer is working on, control the mouse of the remote PC, manipulate the CAD data in collaborative manner, and share ideas using whiteboard to write notes and produce sketches.

- **Communication technologies** (video and audio) to enable information exchange between the remote mine site and the service provider office.

- **Visualisation technology** to share two and three-dimensional mining data.
In the remainder of this paper, we briefly introduce the work we have done for requirement analysis. Then our methods and findings of reviewing existing technologies and recommendations for RME design and prototyping are presented. The paper concludes with a short summary and future work. Part of this work has been reported in a conference as a work-in-progress [16].

2. Background

2.1 Model of Remote Service Delivery

Mining service providers often follow a largely common business model to deliver remote mine services to their clients [2]. This model is illustrated in Figure 1.

![Fig. 1. Business model [16].](image)

As shown in Figure 1, a business case is started by a service request from a client (requester). Upon the receipt of the request, a range of work routines follow on the service provider side. These include project initiation, site visits, internal task assignment, task progress reports and checkups, task collaboration and discussion, document exchange and task reassignment. Depending on the context of the request, the service can be executed by one or more mining engineers either from the same site or from different locations. The communication methods used by people involved include face-to-face, phone calls, one-line video, audio and text discussions, emails, sharing and sending data via physical media.

2.2 Requirement Analysis Method

Understanding user requirements has been widely acknowledged to be critical for the success of any software or system development projects. Our experience with industry related projects show that human factors experts or user experience designers are not often involved at the early stage of the development cycle to elicit and analyse user requirements. They are often called in and/or brought on board after the system has been developed which by then is often too late to address major design issues [14]. To avoid this, we formed a team that consisted of a mix of expertise including user experience designers, technical engineers and human factors experts, and adopted an approach in which observation-based user experience design methods were combined with scenario-based
software design techniques for requirement analysis. More specifically, the user requirements were collected based on the activities described in the following four sections.

**Meeting with the director**
The analysts met with the management of the service provider, at an off-site location, to gather information on expectations, general system requirements and limitations. The meeting also aimed to develop a high-level understanding of typical work scenarios in order to design the on-site observations and focus group sessions to follow. The meeting took approximately three hours.

**Work scenario observation**
Two typical work scenarios were observed on-site at the service provider’s office: a face-to-face task assignment and a task assigned remotely. The analysts observed service provider staff members out of eye-line, making notes and recording audio and video for later comparisons.

**Focus group session**
Seven service provider staff members, including the director, participated in a focus group session once the scenario observations were complete. Experience of the participants varied from one to more than twenty years of working in the mining industry. The group session was conducted in a single meeting room over a period of one and a half hours, with audio recorded for later review.

**Review and confirmation process**
Following completion of the observation and group sessions, the data was processed off-site by the analysts. As necessary, follow-up interviews with individuals from the service provider were conducted over the phone. These interviews helped to clarify and confirm system goals and user requirements.

### 2.3 Results of the Requirement Study

The results of our requirement analysis highlighted three key challenges in developing remote collaboration software for the project, namely:

1. **Bandwidth limitations**
2. **Security concerns**
3. **Usability issues**

Seven user cases were identified as key requirements when implementing a solution for delivering effective services remotely. These user cases were considered as the minimum set of requirements that will allow remote mining engineers and site personnel to gather information, collaborate with each other and deliver the service remotely without compromising the quality of the service. The user cases are:

1. **Ability to communicate through text, audio and video.**
2. **Ability to send a request to use the system and accept / reject it.**
3. **Ability to manage the system and configuration of users.**
4. **Ability to share full screens or application windows.**
5. **Ability to manage different parts of the collaboration tool over different screens.**
6. **Ability to share electronic whiteboards and annotate over applications.**
7. **Ability to transfer data over the network.**
3. **Research Method**

Based on the user requirements obtained, and the limited resources and tight project time frame, it was decided to make most use of existing technologies for design and development of the RME system. To achieve that, we conducted two further studies. The first one was to identify the pros and cons of the client’s daily collaboration practice based on the data obtained. The second one was a scoping study that was intended to identify the best suitable technologies to build a knowledge base for the project team, and to inform the design of the RME system. It should be noted that our scoping study was not a functionality or useability test of those technologies or products.

3.1 **Understand the Current Collaboration Practice**

To be better informed on identification of the best technologies, it is important to have a good understanding of the pros and cons of the client’s daily work practice. This would help analysts to make decisions on which parts of it should continue and which parts could stop or be improved in the new system. The list below includes major limitations that were observed or reported:

1. Engineers often needed to visit the client or mine site.
2. The tool used for video conferencing over the network was not stable: for example, audio was not working.
3. The interface of the tool was not user-friendly when switching between text message, face audio and screen sharing.
4. When discussing over phone or the video conferencing tool, it was difficult to see facial and body languages of the participants.
5. Current facilities and tools did not support hand gestures.
6. For screen sharing, the tool was only able to allow the host side to control the mouse.
7. Files and data were transferred in different ways including CDs, emails and network share, which could cause extra workload.
8. It could be difficult to find specific files and data since they were currently stored in different places, such as, a folder in the local drive, email box or in a network drive. The names of files were also not standardized. There were also temporary files mixed with important files together.
9. There were no history records of conversions and discussions with clients and between engineers, which were important when assigning the task to another person.
10. A number of applications needed to be running at the same time, but there was only limited size of hardware memory (<3GB).

3.2 **Identify Suitable Candidates**

There are a large number of tele-collaboration products on the market that offer a range of services that had the potential to meet some of the requirements for delivering services to remote clients.

A broad product survey was conducted, identifying 56 possible candidates in the market. These 56 candidates were then further examined. This examination was expected to provide a recommendation on an initial system that would satisfy some of the requirements, and highlight scope for extensions or replacements in further implementation stages, in order to satisfy all of the requirements identified.
The scoping study narrowed the candidate list down to 6 likely products, which were then tested against some of the more prominent limitations and user cases required. Out of these tests, 2 candidates were highlighted as products that satisfied some of the limitations and user cases required.

Of these 2 products, Adobe’s Connect Pro [6] and Cisco’s WebEx [7] were highlighted. Connect Pro was felt to be more suitable as it was more stable and handled 3D content better. A detailed summary of this process is included in the next section.

### 3.3 Present Results

The results of the scoping study, a live demonstration of Connect Pro, a recommendation to use Connect Pro in the short term, and design recommendations for the RME system were presented to the project stakeholders during a visit to the mining engineering firm.

This visit also included discussions on future work. It was agreed for CSIRO and the firm to purchase copies of Connect Pro, with the firm to try out using the product between their offices that were located in different cities, and CSIRO to incorporate the product in other strategic research to build up further knowledge.

### 4. Results

#### 4.1 Broad Review

The first step was to survey the available collaboration tools and decide on the best one that would immediately be usable for mining engineers to work together across a network. To do this, a list of 56 possible contenders for this project was compiled.

With the wide variety of products it was necessary to narrow the key types of collaboration that would be needed and the relative importance of each of these. These features, in order of importance, were:

1. Audio communication. Without audio, nothing can happen. It was also assumed that the package itself would need to provide the audio connection, since a telephone line may be too expensive or inconvenient.
2. Sharing of snapshots of applications. Being able to take the snapshots of the screen and video feeds is an important function that any tool should support in the context of mining services.
3. Annotation by all participants over those shared snapshots. This will help draw attention, highlight important information and facilitate verbal communications between the collaborators.
4. Sharing of live applications. It was assumed that these applications (such as a mining CAD package called Vulcan) would use OpenGL for 3D display. This is important because several products could share applications but not those using OpenGL.
5. Sharing control of live applications.
6. Video of participants. This will help present visual body languages to the collaborators and increase the sense of co-presence of them.
7. Ability to transfer files.
8. Recording. Recording of conversations in audio, video or text when appropriate will help when there is a need to revisit them at a later stage, for instance a new engineer joins the project.
9. Having a tool to schedule meetings.

Many of these products were eliminated because they did not have video or audio communication, could not share applications, had poor security facilities, could not deal with network firewalls, or otherwise did not sufficiently satisfy enough of the limitations or user cases required by the project.

4.2 Narrow Review

This list of 56 names was then pared down to a shortlist of 8, based on the product descriptions and documentation. These products in the shortlist, including JoinMe, YuuGuu, MeetingPlace, GoToMeeting and Mikogo, TeamViewer, Connect Pro and WebEx, were then downloaded, installed and tested.

The first five were eliminated because they did not have appropriate audio and/or video facilities. The sixth, TeamViewer, was eliminated because it had problems with annotation over OpenGL applications (such as Vulcan). The main problem with sharing OpenGL applications was that the OpenGL refresh would immediately erase or damage any annotations that had been overlaid.

4.3 Use-case Testing

The remaining two products, WebEx and Connect Pro, were investigated in more detail. These two products were installed and tested with a full set of test cases developed based on the user requirements. It was found that these products were very close in the features that were offered. The main difference was in the way they responded when annotation modes were selected:

- Connect Pro took a snapshot of an application and participants drew over the top of this fixed image (see Figure 2).
- WebEx allowed participants to annotate in the same way over a window that could continue to change (such as a 3D animating application).

![Fig. 2. Connect Pro used for sharing and annotation.](image)
The second approach seemed more powerful as it would allow for indicators, such as animating lines or pointers, to continue operating in a scene. However, the live annotation feature in WebEx can sometimes be unreliable with 3D applications, with annotations disappearing in many situations. As a whole, Connect Pro appeared to more preferable than WebEx in meeting our user needs.

4.4 Bandwidth Testing

Connect Pro was then tested against a key project limitation, reduced bandwidth conditions. For this test, a bandwidth-throttling program was installed at one end of a high-bandwidth network to simulate various low-bandwidth conditions.

It was assumed that reliable audio communication was essential for any collaboration, so this was taken as a measure of usability. If the audio became unusable, the condition failed. Tests were repeated for different combinations of features, to find the minimum bandwidth that could sustain the features and still permit understandable bi-directional audio. The results are shown in Table 1.

<table>
<thead>
<tr>
<th>Bandwidth (Kbits/s)</th>
<th>Audio</th>
<th>Video</th>
<th>Sharing</th>
<th>Annotation</th>
<th>Experience</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>Poor</td>
</tr>
<tr>
<td>150</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>Good</td>
</tr>
<tr>
<td>200</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>Good</td>
</tr>
<tr>
<td>250</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Good</td>
</tr>
<tr>
<td>300</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>Good</td>
</tr>
<tr>
<td>350</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>Ok</td>
</tr>
<tr>
<td>400</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Ok</td>
</tr>
<tr>
<td>450</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>Good</td>
</tr>
<tr>
<td>525</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Good</td>
</tr>
</tbody>
</table>

1 – Subjective experience rating; poor, ok, good.
2 – Low refresh frequency OpenGL application (Qt sample).
3 – High refresh frequency OpenGL application (Google Earth).

The shared application used for this test was Google Earth, as it uses OpenGL for its 3D rendering. During the tests, audio was considered unusable if parts of the audio stream were missing or if the latency was so large that conversations were not possible. It was observed that as the bandwidth was limited, the audio latency would increase. This may be due to packet retransmission within the TCP/IP communications mechanism. During two-hours of the bandwidth testing session the overall upload and download data transfer was greater than 1 Gigabyte each way. This may illustrate that the overall throughput allowance must be fairly high regardless of bandwidth.

It should be noted that Connect Pro did not automatically detect bandwidth and adjust any features. These had to be turned on and off by users.

4.5 Results Summary

Connect Pro was chosen for the more explicit style of sharing and annotating over an application. It was clearer that a snapshot view was being annotated upon, and it left the presenter with the ability to interact with other windows on the desktop when annotation is enabled. Also with Connect Pro, there were no irregularities with annotation on any of the windows tested.
5. Design Recommendations

Generally speaking, existing technologies fall into three broad categories:

- Video conferencing
- Remote application and desktop control
- Web conferencing

Connect Pro was recommended as the top-matching product of a very wide field. Its strength highlighted in our study is the coverage it offers in all three of these categories. Although it is a powerful collaboration tool, Connect Pro has several severe shortcomings when contrasted against the project requirements. That said, Connect Pro provides a software development kit and many plugins have been developed that can extend and even replace existing functionality with targeted solutions. This opens up many opportunities for making improvements on current disadvantages.

5.1 Bandwidth and Latency

Bandwidth is generally viewed as an ever-increasing commodity (see Table 2). As such, the attention needed to address bandwidth concerns may often be overlooked. While bandwidth does continue to increase, demand is also set to increase.

<table>
<thead>
<tr>
<th>Table 2. Predicted IP traffic Growth 2014-2016 in Peta Bytes</th>
</tr>
</thead>
<tbody>
<tr>
<td>IP Traffic, 2011-2016 (By Geography (PB per Month))</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>North America</td>
</tr>
<tr>
<td>Western Europe</td>
</tr>
<tr>
<td>Asia Pacific</td>
</tr>
<tr>
<td>Latin America</td>
</tr>
<tr>
<td>Central and Eastern Europe</td>
</tr>
<tr>
<td>Middle East and Africa</td>
</tr>
</tbody>
</table>


In the case of the RME system, mine sites generally have poor internet connections due to security concerns and physical isolation. Even with high bandwidth infrastructure in place, other factors may reduce the available bandwidth and/or introduce latency. Network security may introduce latency, while high concurrent usage can reduce bandwidth – especially on WiFi networks – and mobile internet connections continue to lag behind bandwidth capabilities of fixed infrastructure. In these cases, bandwidth variability is the primary issue.

As such Quality of Service guarantees that cater for bandwidth and latency tolerance should be an essential part of any web-based communication product. At the very least, variable audio/video shaping technologies similar to those implemented by Skype should be included, catering for both low and variable bandwidth situations. Not all products reviewed provide this adaptability, choosing instead to measure available bandwidth only on initialisation.

5.2 Security

Connect Pro uses Secure Sockets Layer encryption to secure communications [6]. However, these connections are routed through central Adobe servers. Self-hosting options exist for enterprise level subscriptions at an additional cost. While this ensures that communications are adequately secured,
some situations may require more direct connections, or the expense of self-hosting may not be justifiable.

Desktop sharing allows participants to see the host’s desktop. Connect Pro supports restricting visibility of individual windows. We view this as an essential security feature. However, some rendering artefacts are visible over hidden windows (garbled graphics).

5.3 Programmatic Enhancements

During testing it was found that while Connect Pro offered a wide variety of features, certain features were missing or needed improvement.

The most technical of these is the ability to annotate over live, 3D windows using a high refresh rate. Connect Pro’s approach of capturing an image of the screen and annotating over this static image provides a very straightforward workaround, but was found to be inconsistent. Often, different clients showed different frames captured from animated windows. In this context, annotations may not make sense to some participants, as they cannot see the intended target image. Annotation over live windows offers a more dynamic experience and allows the simultaneous integration of several features, such as annotation and remote application control.

Below is a list of other desirable features and their justifications:

- **Guaranteed connection**: Connect Pro attempts to maintain a connection in low-bandwidth environments, but was still prone to dropping connections when communications was lost entirely. This required the user to reinitiate a connection and renegotiate a remote connection each time. After initiating a connection, links should continually attempt to re-establish broken connections until the user manually disconnects.

- **Virtual cursor**: each participant controls a virtual mouse pointer. This allows transient gestures and features highlighting without needing to manage annotations.

- **Control request**: where remote application control is allowed, participants may request control of an application using integrated features rather than vocalisation. This forms a more formal and unobtrusive request mechanic.

- **Improved file transfer**: file transfer should not be limited to the current session. That is, file transfers initiated during a session may complete after the session has been terminated. This allows larger file transfer. File transfer must also be integrated with bandwidth and latency tolerance algorithms.

- **Private whiteboards**: private chat features already exist, allowing users to send private text messages. This concept is extended to private whiteboards, allowing sharing of images, documents, drawings and annotations between subsets of participants.

- **User/client database and management**: at the time of evaluation, there was no address book built into Connect Pro. User management and configuration tools allow for better management of repeated connections, meeting scheduling and user privileges. Better still, integration with an external client database maximises the management of clients in a single location.

- **Layout management**: Connect Pro features many internal windows including video, chat, user lists, annotation and presentation, etc. Layout management, restoration and sharing may enhance productivity and familiarity.
6. Conclusion and Future Work

In this paper, we have presented our approach towards the design and implementation of the RME system. This approach makes use of the existing the technologies for knowledge development and for system design. Firstly, 56 candidates were compiled and compared based on their functionalities and application requirements. This resulted in 8 products being identified for further testing in a simulated mining office environment. Lastly, Connect Pro was considered to be the best suitable system to meet our specific user needs, and to inform the design and implementation of the RME system. During the process, further functions were also proposed in addition to what was available in Connect Pro.

For future work, we plan to develop further features based on the obtained user requirements and design recommendations. During the process, end users will be fully involved and their needs will be fully addressed whenever possible. We also plan to experiment and incorporate some additional technologies into the RME system. These include augmented reality, remote gestures, remote fault diagnosis and virtual presence. It is expected that a fully functional RME system will bring following benefits:

- Greater mine planning stability as a result of reduced personnel turnover at mine sites and improved communication capabilities with mining consultants or other external suppliers
- Improved communication tool for general use between head offices and mining sites
- Reduction in subsidized housing or costs due to staff being based in capital cities
- Increased ability to attract mining engineers resulting from the potential to work in city areas
- Increased ability to attract school leavers into mining engineering field due to city based roles
- Successful deployment of the system could lead to the reduction of other service personnel such as environmental and geological staff at the mining site, and enable remote/autonomous mine operation.

Further, it should be noted that although RME has been designed to address challenges posed by unique mining environments and can be used for delivery of remote mining engineering services, its usage goes beyond the mineral industry in any environment where tele-collaboration, tele-presence and virtual technologies could be of value, for example, in the provision of health services remotely.

References


A Case Study of Collecting Dynamic Social Data: The Pro-Ana Twitter Community

Ian Wood
Australian National University, Canberra ACT 0200, Australia, ian.wood@anu.edu.au

Abstract. The study of social processes in on-line social media content is a relatively new and rapidly growing endeavour. Many social media platforms provide a public API which can be used for the targeted collection of data from perceived communities, however existing software for this purpose focusses on a “snapshot” of the community and its communications, and ignores important aspects of its dynamics. We present a data collection system designed to capture tweets and the dynamics of Twitter user profile and friend/follower lists, and an approach to identify a set of tags or keywords that define an on-line community. This approach and system were used to collect a data set spanning more than 3 Christmas periods from the “pro-ana” and eating disorder Twitter community.

Keywords: data set, Twitter, software, adaptive cluster sampling, network dynamics, social networks

1 Introduction

With the exception of one very recent paper [2], data collection for the study of Twitter friend/follower network dynamics has been limited to a small number of network snapshots without finer grained timing [6, 1, 3] and focus on large scale and/or short term effects, without consideration of specific Twitter communities.

Working from the observation that Twitter hash tags can define a community [7, 4], we propose an approach analogous to adaptive cluster sampling [5] to iteratively expand a small selection of hash tags thought to be used by/define a community by examining tags used in tweet samples from queries on the current set of tags. Frequent tags are added if they do not attract many tweets deemed to be from outside the community.

For some time, a phenomenon has existed on the internet that conveys positive messages about anorexia, the so called “pro-ana” movement. In addition to definitively “pro-” anorexia messages, there are other messages relating to eating disorders and connected social phenomena such as the “thin ideal”. In order to study the “pro-ana” phenomenon and it’s social context, we sought to collect “pro-ana” and eating disorder related Tweets. A selection of Twitter hash

\[1\] the authors claim to have precise timing for friend/follower network changes, however they do not describe how they obtained the data
tags used almost exclusively by people with eating disorders was constructed following our approach.

We present a system that captures Twitter friend network dynamics by polling a users’ friends and followers lists and user profile data each time we collect a tweet from them. User profile data is also polled with each tweet. Any changes are stored in a replicated MongoDB database and regular backups are performed. It incorporates several features to ensure reliable operation in the face of extreme tweet rates, outages and failures. The system was used to collect all tweets containing the identified hash tags and associated dynamic user data over a period of spanning three Christmas periods — over 1.2 million tweets, 300 thousand users and 200 thousand images.

In Section 2 I discuss communities defined by hash tags and the approach used to identify tags from the Twitter “pro-ana” community. In Section 3 I discuss the approach for sampling the dynamics of community network and user data. In Section 4 I discuss the software design and technical challenges. In Section 6 I summarise our findings, and finally in Section 7 I talk about possible system enhancements and related research directions.

2 Identifying Communities

2.1 Hash Tags and Communities

Hash tags are used in Twitter and other micro-blogging sites as a way to organise, emphasise and otherwise colour posts. A hash tag is simply a word with the hash character “#” prepended, such as #diet. They allow users to specify aspects of their posts that they consider important and to direct their posts to what they feel is an appropriate audience [7, 4]. It is this second point that we attempt to harness in order to collect the output of a hypothesised community around “pro-ana” (pro-anorexia) and eating disorders.

2.2 Adaptive Sampling for Search Tags

Following the intuition behind adaptive sampling [5], a search query for collecting tweets was selected through an iterative process identifying hash tags used by people with eating disorders. At each step a sample of tweets was collected, then a set of potentially relevant hash tags was selected from frequent tags in that sample.

In the first iteration, a brief study of tweets and Tumblr\(^2\) posts containing #proana revealed related tags #thinspiration #anorexia #bulimia and #pro-mia\(^3\).

For the second iteration, tweets were collected from August 18–22 2012 on these related tags, a total of 1182 tweets. Hash tags were counted in these tweets, and those with more than 3 occurrences were considered. The majority clearly

\(^2\) www.tumblr.com
\(^3\) Short for bulimia nervosa, an eating disorder related to anorexia nervosa.
had much wider usage than by people with eating disorders (e.g.: #diet). A quick manual check (by conducting a twitter search on each tag) indicated that this presumption was correct, and they were discarded. The remaining tags were also manually checked, revealing many that also had wider usage. Such tags were discarded if most of the relevant tweets in which they were used also contained other tags still in the set.

For each tag, a judgement had to be made about the number of irrelevant tweets compared to the number of tweets not collected by other tags in the query. In general the distinction was clear, and no compromise was necessary. However a few, such as #depression and #selfharm were discarded despite identifying a small number of relevant tweets not identified by other tags. It was deemed preferable to maintain a higher degree of data relevance and a smaller query (which stands a smaller risk of saturating the Twitter API limits resulting in lost data). Table 2 lists some typical types of tags that had wider usage with a few examples.

This process to all the remaining tags resulted in a core of 14 tags identified as used almost exclusively by people with eating disorders, as well as capturing nearly all tweets in the sample that were deemed produced by those people. A further 4 days of tweets from September 15–19 2012 were collected on these tags and the analysis repeated, however no extra tags were identified. Table 1 lists the selected tags.

It is interesting to note that searches on the words proana, pro-ana, pro-mia etc. . . (without the #) were also investigated, however a substantial portion of posts from these searches were from people discussing the phenomenon rather than those within it. Also, posts retrieved by these searches which expressed pro-anorexia or were assessed as authored by people with eating disorders could be identified by the use of hash tags. This supports the idea that hash tags can form the focus of a community. Similar results were found for several other non-hash tag terms and phrases.

### 3 Collecting Dynamic Twitter Data

Social psychology is a dynamic process — people enter and leave social groups and groups change and adapt their sense of identity and social norms. It is these dynamics that make our societies what they are, that generate and define human social fabric. In order to study and eventually make predictions about group behaviour, it is essential that we capture the dynamics of the socially meaningful features under study.

---

Table 1. Tags Selected For Search Query

<table>
<thead>
<tr>
<th>#proana</th>
<th>#promia</th>
<th>#anasisters</th>
<th>#bulemia</th>
<th>#bulimic</th>
<th>#ednos</th>
<th>#edproblems</th>
<th>#hipbones</th>
<th>#thingsanataughtme</th>
<th>#thinspiration</th>
<th>#thinspo</th>
<th>#abcdiet</th>
<th>#thighgap</th>
</tr>
</thead>
</table>

---
### Table 2. Non Personal Eating Disorder Tags

<table>
<thead>
<tr>
<th>Description</th>
<th>Tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>diet and weight loss</td>
<td>#diet #weightloss</td>
</tr>
<tr>
<td>body shape and parts</td>
<td>#skinny #thin #supermodelbody #legs #bones</td>
</tr>
<tr>
<td>adjectives and verbs</td>
<td>#perfect #motivation #recovery #perfection</td>
</tr>
<tr>
<td>used to discuss eating disorders</td>
<td>#anorexia #eatingdisorder #eatingdisorders</td>
</tr>
<tr>
<td>fitness and exercise</td>
<td>#fitspiration #fitspo #gym</td>
</tr>
<tr>
<td>depression and it’s symptoms not specific to eating disorders</td>
<td>#selfharm #depression #cutting</td>
</tr>
<tr>
<td>with other meanings irrelevant to eating disorders</td>
<td>#ana #ed #mia</td>
</tr>
</tbody>
</table>

Purchased twitter data as well as tweets collected with the Twitter API’s, contain a snapshot of the tweeting users’ profile information with each tweet, giving some information about user profile changes a user may have made between the collected tweets from that user. However the profile snapshot does not contain the lists of friends and followers of the user, and to the best of my knowledge historical data on changes in the friend/follower network are not commercially available.

Several previous studies of Twitter friend/follower network dynamics have used a small number of network snapshots without finer grained temporal information and collected data for a fixed set of users, without focus on particular communities. One very recent study of note was able to obtain a large data set containing precise timing for friend/follower network changes, however they do not describe how. Their data is sufficiently large that it would contain comprehensive information from many social groups, however it spans just one month.

The Twitter APIs do not provide a feed containing network changes as they happen, however they do provide a REST (representational state transfer) interface for collecting a snapshot of a users’ friends and followers lists. The data collection strategy presented here uses that endpoint to poll a users’ friend/follower lists each time they tweet. In this way the dynamics of the friend/follower network of users active in the data set is recorded in a similar way to the dynamics of their user profiles.

---

4 though user data embedded in tweets can be stale.
5 a recent Quora post claims that the TwitterCounter service provides this for up to one year.
It should be noted that the dynamic data thus collected is not complete. A user who watches tweets in the data set, but themselves tweets rarely, will only be polled on the occasions they tweet — profile changes and follow/unfollow actions made between recorded tweets by users not active in the data are only captured in aggregate (though follow/unfollow actions will be detected if the recipient tweets). The temporal data, then, has an element of sampling error, a systematic lag, which is particularly pronounced with infrequent tweeters. None the less, one might hope that changes to both the user profile and friend/follower lists which are relevant to social processes within the hypothesised pro-ana twitter community will have a high probability of occurring near the time of each tweet to that community (i.e.: each tweet within my data set). We should, however, be careful when analysing behaviour of infrequent tweeters, as there may be a bias in the recorded dynamics of their data.

4 Algorithms and Technical Challenges

The primary design goal is to capture the complete dynamics of this community’s members and interactions. This was not always achieved due to technical issues (which can or have been fixed) and pragmatic concerns (for which strategies can be designed to minimise the impact). Over the course of two years of continuous data collection, a stable system that is robust to twitter outages and sudden increases (by orders of magnitude) in data volume, and efficiently utilises the narrow data bandwidth for collecting friend/follower data has been built.

4.1 Overall Architecture and Fault Tolerance

The system uses a multi-threaded architecture, enabling asynchronous HTTP requests to the various Twitter API end points and media URLs (see Figure 1). Communication between threads is achieved with thread-safe queues. A tweet collection thread regularly polls the search/tweets REST API. Each tweet is parsed and information passed to the other data collection threads: user id’s to the threads for user profiles, friends list and followers list, and media URLs and ids to a media download thread. There is also a backup thread that initiates regular database backups.

Twitter has two modes for accessing its REST APIs: with user authentication and with application only authentication. With application only authentication, you cannot perform tasks on behalf of a user (which we don’t want to do anyway), but you are given separate API rate limits. For a data collection application such as this, using both forms of authentication essentially doubles your rate limit — in the case of polling friend and follower lists, this is significant. Thus four threads were used for collecting friends/follower information (each of friends or followers with each of user and application only authentication). In principle, multiple ‘applications’ could be registered, multiplying the available data rate, but that would contravene Twitters terms of service.
Fig. 1. Overall Architecture. Arrows indicate data passed between threads.
On restarting after system shut down, an initialisation thread reconstructs the data collection queues by scanning the tweet and user databases for partial and out of date data (e.g.: friend or follower lists that have not been polled since the last tweet from a user). All tweets since the last collected tweet are requested, however twitter does not guarantee that some will not be missed. Our experience indicates that for this query, tweets are usually accessible for several days, though this is probably not the case during heightened tweet rates.

For improved stability, the main thread monitors the other threads, restarting them if they crash and coordinating system shut-down when requested. To help debug frozen threads, the main thread also responds to operating system QUIT signals, dumping a stack trace of each thread. As a further precaution, a unix cron script (which is run regularly by the operating system) monitors the system and restarts the system if an hour passes without activity (measured by log file modification time — note that even if no tweets are collected, log messages are generated and activity is recognised).

Data is stored in a replicated MongoDB instance. MongoDB was chosen due to its easy deployment, easily modified schema, easy replication and because the native format of stored data is JSON, the same as what is returned by the Twitter APIs. Three main collections are maintained: tweets, user profiles and media meta-data (see Table 3). In addition there is a redundant collection containing tweets collected by the streaming API to test the relative reliability vs. the REST API (see Section 4.4). The Nectar research cloud was used to house database replicas and for reliable storages of database backups.

<table>
<thead>
<tr>
<th>Collection</th>
<th>Twitter Data</th>
<th>Added Meta-Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>tweets</td>
<td>tweet data</td>
<td>– how and when it was collected</td>
</tr>
<tr>
<td>entities</td>
<td>media entity data</td>
<td>– tweets that contain the entity</td>
</tr>
<tr>
<td></td>
<td></td>
<td>– a history of any changes to it’s data</td>
</tr>
<tr>
<td></td>
<td></td>
<td>– download attempts/success</td>
</tr>
<tr>
<td>user profiles</td>
<td>user profile data</td>
<td>– history of profile changes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>– friends and followers lists</td>
</tr>
<tr>
<td></td>
<td></td>
<td>– a history of friend/follower list changes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>– when and how the data was last polled</td>
</tr>
<tr>
<td></td>
<td></td>
<td>– number of tweets collected from this user</td>
</tr>
</tbody>
</table>

Table 3. Database Collections

4.2 Polling friend/follower lists — the main bottleneck

Tweet rates approximately follow a power law distribution, and the pro-ana/eating disorder query is no exception. Tweet rates remain at a low level most of the time, however occasionally, the rate increases by an order of magnitude, and

---

6 http://nectar.org.au/research-cloud
very occasionally by many orders of magnitude as very popular tweets are re-tweeted many times. During these “retweet storms”, Twitter API data rates can be exceeded by the desired data collection strategy.

Four public Twitter REST API endpoints were used to collect data: tweets by query string, user profiles, user friend lists and user follower lists. All these API’s have data rate limits\(^7\) which were on occasion met. Collection of tweets for the pro-ana and eating disorder community query fell behind during the highest tweet rates, however due to the ability of the tweet search endpoint to retrieve past tweets, the pro-ana/eating disorder query did not apparently lose data as a result.

The user profile endpoint can poll 100 users per query with 180 queries per 15 minutes. This high rate quickly caught up with even the most extreme “re-tweet storms”, and it was sufficient to prioritise users whose previously stored data was oldest (unseen users first, ties resolved by user id).

The API endpoints for friends and followers lists poll only one user per query and 15 queries per 15 minutes. Also, each query returns a maximum of 5000 friend/follower ids — occasional users with millions of followers require hundreds of queries. Frequent ‘moderate’ re-tweet storms often took days to clear the queue, and extreme events could take weeks. This substantial delay was considered unacceptable.

Investigation of the re-tweet storms indicated that the majority of tweeting users had no other tweets in our data, especially for the more extreme events. Thus a strategy was implemented where users with at least one other tweet in our data were given priority. Of those, the user whose friend/follower data was oldest (i.e.: least recently polled) was given priority. With this strategy, more frequent tweeters were quickly re-polled, while the queue of less interesting one-tweet users can take many days to eventually clear. A newly seen user who tweets again before having been polled is moved to the front of the higher priority queue (‘never’ is considered least recent). In an attempt to get a snapshot of at least some one-tweet users friend/follower lists at the time they tweet, the most recently seen first-tweet users are polled first. In both priority schemes, rare ties are resolved by lexical order of user names.

\[
\begin{array}{|c|c|}
\hline
\text{User seen before? Which user first?} & \text{Priority} \\
\hline
\text{First Priority} & \text{Repeat User} & \text{Least recently polled} \\
\hline
\text{Second Priority} & \text{New User} & \text{LIFO queue} \\
\hline
\end{array}
\]

\textbf{Table 4.} friend/follower lists polling priority scheme

\subsection*{4.3 Image Collection}

Many tweets, and especially tweets in this data, contain images. Twitter includes media URLs in tweet meta-data, and assigns a unique id to each image. When

\(^7\) https://dev.twitter.com/rest/public/rate-limits
a tweet containing images is collected, it’s meta-data is stored in the database including a link to the tweet. If that twitter image id has not been downloaded yet, it is downloaded and stored as a file on disk. A simple cron script is used to backup the stored images to the servers running the database replicas.

4.4 Other Technical Challenges

During the first year of data collection, Twitter announced that it was making significant changes to it’s APIs, and especially to rate limits and the ways they are reported and applied. The system attempts to utilise it’s rate limits as fully as possible without exceeding them (which can prompt Twitter to block the application for a time), so the API changes required substantial adjustment to the rate limit monitoring logic. There were also a few changes to the meta-data for tweets and users. This did not directly require changes to program logic, however in order to keep database consistency, some logic was added to update old format records.

At the time of initial development, Tweepy\textsuperscript{8} was chosen for access to the Twitter APIs. Unfortunately, at that time Tweepy did not have support for application only authentication. Twython\textsuperscript{9} did however, and since both packages present the Twitter API in a similar way, it was not difficult to add threads that utilised this capability.

Tweets can be directed to a recipient twitter user. Collecting the friend/follower lists of recipients was also attempted, however it soon became evident that recipients were frequently celebrities with millions of followers, causing extra burden on the already stretched follower API endpoint. Users of interest that are part of the pro-ana/eating disorder community would be tweeting regularly, and we would be polling their friends and followers lists regularly anyway, so it was decided that polling tweet recipients should be abandoned.

To test the relative reliability of the Twitter streaming and tweet search APIs, a separate process received tweets via the streaming API and stored them in an extra database collection. This collection was monitored by the main program, and any extra tweets were copied to the main tweet collection. We found that no tweets would have been lost without the streaming API data, so this part of the system is unnecessary.

During system development, a bug in the python http library and difficulties coordinating thread locks were identified from stack traces generated by the main thread in response to Unix QUIT signals. Early in development, data loss was avoided by automatically restarting the system via an hourly Unix cron script when further bugs triggered by infrequent combinations caused the system to crash. The system has now been running continuously for over a year without any of these problems.

\textsuperscript{8} http://www.tweepy.org/
\textsuperscript{9} http://twython.readthedocs.org/en/latest/
5 Summary of Pro-Ana Data

As of 30 January 2015, the data contained 1,283,875 tweets, 296,483 users and 307,723 image ids. There were 1,616,199 follow events, 1,616,188 unfollow events and 1,655,280 user profile changes. Hash tag usage followed a typical power law distribution (Figure 3), as did the number of followers and friends (Figure 2) though follower and friend counts were not correlated. The number of tweets per user, both in our data and overall also follow a power law.

![Figure 2](image-url). Number of followers and friends of users (one mark per user).

![Figure 3](image-url). Tag frequencies (converted to lower case, one mark per tag).

Hash tags related to “thinspiration” (typically images of people, mostly women) dominate the data, with 73% of tweets. Retweets and images also account for a significant portion, with 57% of collected tweets retweets and 71%
containing images. Thinspiration tweets account for 89% of the images, 80% of the retweets contain images and 76% of retweets contain thinspiration tags.

6 Conclusion

Capturing data on dynamic aspects of social media communities is important for the study of online social behaviour. Systems designed to capture data from Twitter and other social media typically lack the ability to capture important dynamics, such as changes in the social network. We have constructed a system that captures these dynamics from Twitter communities that can be identified by their use of hash tags or other search terms. The system is very robust to the bursty nature of tweet streams, network problems and other difficulties associated with online data collection.

We present an approach similar to adaptive sampling to identify hash tags relevant to a community. With this approach, we identified a set of tags that is used almost exclusively by the Twitter “pro-ana” and eating disorder community and used our system to collect a nearly unbroken record of tweets, user and network data from that community covering three Christmas periods: over 1.2 million tweets, 300 thousand users and 200 thousand images.

7 Further Work

For a longitudinal study of an online community, it may be appropriate to revise the tweet search query on a regular basis (say, each month or quarter), as tag usage may change over time, with new tags adopted by the community and old tags losing favour.

Identification of users of interest (e.g.: apparent members of a social group), perhaps through analysis of the friend/follower, re-tweet and/or user mention networks, could be used to prioritise and enable collection of extra data from those users. For example, regular polling of friend/follower data and user profiles in periods of spare API bandwidth and/or at the expense of timely data collection from less interesting users.

The search/tweets REST API is not saturated by the current data collection approach, and substantially more tweets could at times be collected. Tracking tweets by all individuals in the data would quickly become intractable, however it could be valuable to collect more or all tweets by identified interesting users. A simpler strategy of collecting all tweets from users for a certain time since their last tweet could also be valuable.

The final search query from tags in Table 1 represents a balance between a wide net and saturating the twitter API limits. The choice was made to keep the query small in order to maintain a high degree of relevance in the data at the expense of not collecting a small number of relevant tweets. It is an interesting feature of the “pro-ana” phenomena on Twitter that the choice was quite clear, and that little compromise was needed. It would be of interest to investigate...
other potential Twitter communities to see if their boundaries can be so clearly delineated.

In Section 3 we mentioned the systematic temporal sampling error inherent in the data collection process. A valuable addition to research into social media as a metric for social processes would be the study of this and related sampling errors (e.g.: self selection bias).

Since the creation of this software, Twitter has introduced several new API endpoints that could be integrated into the collection strategies to improve the resolution and fidelity of dynamic, particularly friend/follower network, data. Of particular interest is the friendships/show Twitter API endpoint, which returns information about the relationship between two twitter users and has a high rate limit of 180 calls per 15 minutes. Given a technique to regularly identify of users of particular interest, their user relationships could be polled more frequently. Another Twitter feature that may be of interest is lists. Users can create and join lists, and use their list membership(s) to filter the tweets that appear in their Twitter feeds or manually view list tweets. List membership of users in the pro-ana data follows a typical power law distribution, with about 40% of users members of some list.

Acknowledgement.

Many thanks to Dr Henry Gardner and Dr Richard L. Jones for assistance preparing this paper.

References

8 Appendix—Thread Locks and Events

The system has several resources that are shared between threads: the friends list threads, followers list threads and user data thread all write to the database collection containing user profile information. There is also an object for tracking rate limit status that is shared by all data collection threads. Locks are required to prevent threads from making changes simultaneously and potentially losing or corrupting data.

MongoDB and its Python API are thread-safe, and so long as the same record is not processed simultaneously, we are ok. In order to prevent excessive wait times as other threads process their data, a collection of locks was implemented, keyed by Twitter user id’s. In this way, a thread must wait only when another thread is processing the same user id. The lock collection creates a new Lock object each time a new user id lock is requested, and so grows steadily. If we were dealing with millions of user id’s, this could be a problem, however for this system it was considered acceptable and no strategy was implemented to remove old, long unused Locks.

The system has two methods relating to rate limit status: one to poll the rate limit status and update the internal record, the other to simply query the internal record, but with an option to update it also. Due to the at times inconsistent rate limit reports from Twitter, a number of heuristics are applied to detect and abandon incorrect reports. To do this, the update method needs to make calls to the query method. Since these methods call each other and potentially make changes to the internal record, a re-entrant thread lock is used (a re-entrant lock can be acquired by a thread multiple times). Note that query method calls from within the polling method never request an update, so infinite recursion is avoided.

Further communication between threads is implemented with events. An event is an object shared between threads which can be set, cleared, read or a thread can wait until the event is set (with an optional time limit). Events were used for two purposes: to request threads to wind up their activities and close down when the system is shutting down; and to inform data collection threads when there is new data in their respective queues.
Comparing eye gaze tracking to reported perceptions of manipulated and unmanipulated digital images

Sabrina B. Caldwell, Tamás D. Gedeon, Richard L. Jones, Martin Henschke
Research School of Computer Science, The Australian National University
Acton ACT 0200 Australia

Abstract
To investigate human perceptions of image manipulation at both the conscious and non-conscious levels, we compared participants’ verbal reporting of image manipulation to data recordings of their eye movements while viewing 36 images of varying manipulation levels. To further understand subjects’ ability to use image comparison tools to aid in manipulation detection, variants of a trial ‘image packaging’ software provided two levels of image comparison support tools.

Keywords: Eye gaze; manipulated images; detect image manipulation

Introduction
Increasingly, we encounter our information about the world in image form. At the same time, the ability of humans to manipulate images is greater than at any time previously in history. While there is research on the use of manipulated images in advertising there is very little understanding of the effects of ubiquitous photo manipulation such as is rife in social media and family photos.

As a step in understanding this phenomenon, this experiment uses both eye gaze tracking and verbal questioning to compare what subjects see (as represented by their eye gaze tracking results) and what they perceive (as represented by their question responses) when provided with both standalone images and images that have been packaged with additional assistive information.

This experiment investigates peoples’ ability to see manipulations in images, and seeks to identify whether providing additional comparison images along with the presentation image enables participants to identify manipulations in images more accurately and/or quickly.

In addition, the experiment attempts to determine how subjects interpret images in relation to any manipulations they contain.

1 For example effect of airbrushed models on teen body image [Grabe et al 2008]
Materials and Methods

Twelve volunteer participants undertook eye gaze tracking and verbal questioning as they viewed images of photographs ranging from unmanipulated to strongly manipulated. The participants’ mean age was 32.7 (SE 11.1) years.

Facelab 5.0.2 by Seeing Machines was used to track eye gaze with two infra-red (IR) cameras and a single IR light emitter pod centrally located in front of and below the monitor displaying the images. Eyeworks v3.8, also by Seeing Machines, was used for experiment delivery, recording and analysis.

Subjects were shown three sets of 12 images each comprising 3 unmanipulated images and 9 images manipulated by splicing in or erasing elements of varying sizes from the scene (examples Figure 2). The first set comprised standalone images. The second and third sets were presented in a mobile, self-contained image format (MSCI), a trial ‘image packaging’ software currently under development. The second set presented bundled images in which a presentation image was accompanied by the original image for comparison. This configuration was repeated in the third set but also accompanied by a ‘differences map’ image that highlighted any changed pixels.

Images within sets were varied in order using Latin square randomisation to avoid any ordering bias. In each subsequent set some images were repeated in the new format to identify how much assistance subjects needed to identify manipulations.

Subjects were assigned to one of two sub-groups, those who had pre-existing familiarity with image manipulation issues through exposure to the authors’ research (informed), and those who did not (uninformed). In some cases (4 subjects) an additional set of eight images were employed after the common part of the experiment, to further test whether focused exposure to image manipulation predisposes subjects to identify manipulated images with increased accuracy.

At the same time as their eye gaze was being tracked, participants were asked a short set of questions relevant to each phase of the experiment and their responses recorded. These questions targeted their perception of any manipulations that might appear in the images they viewed, and their interpretations of the images.

Finally, participants completed a short survey and responded to open-ended questions about their attitudes towards image manipulation.

Figure 1: Image manipulation identification accuracy n=12

Accuracy in identifying manipulated images unaided

<table>
<thead>
<tr>
<th></th>
<th>Informed</th>
<th>Uninformed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stand-alone Images</td>
<td>43.1%</td>
<td>31.9%</td>
</tr>
<tr>
<td>With Comparison Image</td>
<td>97.2%</td>
<td>81.9%</td>
</tr>
<tr>
<td>With Comparison Image and Difference Map</td>
<td>97.2%</td>
<td>97.2%</td>
</tr>
<tr>
<td>Set</td>
<td>Roses copy/move</td>
<td>Pier retouch</td>
</tr>
<tr>
<td>--------------</td>
<td>--------------------------------------------------------------------------------</td>
<td>------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Set 1 - Standalone image</td>
<td>![Image of roses]</td>
<td>![Image of pier]</td>
</tr>
<tr>
<td>Set 2 - Comparison original</td>
<td><img src="image1" alt="Comparison original image" /></td>
<td><img src="image2" alt="Comparison original image" /></td>
</tr>
<tr>
<td>Set 3 - Comparison original and difference map</td>
<td><img src="image3" alt="Comparison original and difference map" /></td>
<td><img src="image4" alt="Comparison original and difference map" /></td>
</tr>
</tbody>
</table>

Figure 2: Examples of images shown to subjects in Experiment A
<table>
<thead>
<tr>
<th>Coins copy/move</th>
<th>Man on beach retouch</th>
<th>Helicopter copy/move</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not shown in set 1.</td>
<td>Not shown in set 1.</td>
<td>Not shown in set 1.</td>
</tr>
</tbody>
</table>

Figure 2 (continued)
Results

Overall, the ability of participants to verbally identify manipulated images with no assistance was weak, only 37.5% accuracy on average (Figure 1). This was despite eye gaze data indicating that subjects had looked directly at the manipulated areas with greater intensity than would be predicted by the area of the manipulated regions of the image (Table 1). Informed subjects were more likely to report image manipulations than uninformed subjects (43.1% vs 31.9%).

<table>
<thead>
<tr>
<th>Regions of manipulation: % of participant views in relation to area of manipulated regions in Set 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>% of gaze used in manipulated region(s)</td>
</tr>
<tr>
<td>-----------------------------------------------</td>
</tr>
<tr>
<td>Area of manipulated region(s) in pixels</td>
</tr>
<tr>
<td>Area of image in pixels</td>
</tr>
<tr>
<td>Manipulated region % area of image</td>
</tr>
<tr>
<td>% of views in relation to % of area</td>
</tr>
</tbody>
</table>

Table 1: Comparing ratio of participant views to area size of manipulated regions

When given a comparison original image, participants’ success rate at identifying manipulations more than doubled to 85.4%, although often they either could not say what had been changed or else misreported or under reported what had been changed. Eye gaze data indicated that participants gaze traversed the two images to identify and compare areas of difference. Again, there was a difference in accuracy between informed and uninformed subjects (97.2% vs 81.9%).

When also given a difference map in set 3 of the experiment, participants’ success rate in identifying manipulated images increased to 97.2% for both groups, and the difference in quality of perception of manipulations increased. Eye gaze data indicated that participants’ gaze referred to the difference map as an aid in locating manipulated areas presentation (manipulated) images as compared with the original.

Using the difference map, participants could identify 90.1% of specific manipulations in
detail. When offered a comparison map but not a difference map, participants accurately identified the specific manipulations only 68.1% of the time. Further, incorrect alternative explanations for the effects of manipulations were given, for example the insertion of three additional roses into an image was described by three participants as increased colour contrast.

Moving from the second set to the third set also increased speed of identification of manipulated images, with the time from image appearance to the decision point reducing from 12.4 to 5.3 seconds (standard deviations of mean in seconds were 5.16 and 2.99 respectively). In both cases this compares favourably with the time required for potential identification of manipulated images in Set 1 in which participants were presented with standalone images, 45.7 seconds (SD 7.1).

There was little correlation between the size of the manipulation and the accuracy rate (Figure 3). Comparing eye gaze tracking to question responses indicated that subjects’ eye gaze fixated on regions of manipulation in images even when they did not report the image as altered. Overall, while participants only identified 37.5% of the nine manipulated images in Set 1 (standalone images), their eye gaze rested in the regions of manipulation up to 11 times as often as the area the manipulations occupied would predict (Table 1).

As an example, the image of Australian Prime Minister (1996-2007) John Howard with a spliced image of Queen Elizabeth II from Set 1 standalone images was not verbally reported as manipulated by any participant despite the eye gaze ‘heat map’ demonstrating that the region of the spliced image was viewed intensively at an average 27% of eye gaze within area of the photograph and over 4 times more frequently than would be seen by area alone (Figure 4).

In responding to the survey questions, all participants stated that they cared about photo credibility (average 8.7 out of 10 indicating that on average all the subjects care a lot), societal implications of photo manipulation (average 9.1 out of 10 indicating this significantly matters), and the lack of photo authentication solutions available (average 8.25 out of 10 indicating concerned to very concerned).
Responses to the question “How easy is it for you to tell if a digital photograph you are looking at has been manipulated?” yielded answers clustered around the midway point indicating a middle ground between difficult and easy (average 4.9 out of 10).

In post experiment open-ended questioning, 10 (83%) of participants responded that they use one or more photo editing software packages including Adobe Photoshop, Microsoft Paint, Fireworks, Instagram, Gimp and Adobe Illustrator. The uses to which these software systems are put include cropping, red-eye reduction, colour adjustment, light adjustment, ‘fun filters’ in the case of Instagram, and making collages.

In response to the question of how they identified manipulations when viewing photographs, the strategies identified by participants were “searching for sharp edges,” “when things don’t look right, like one face on another person’s body,” “lighting effects,” “if the dimensions are wrong,” “if things are too perfect,” and “shadows going the wrong way.” Some participants (25%) stated they did not know how to identify manipulations in images.

When asked what they look for in a photograph, with suggestions of aesthetics, meaning, or representations of reality, participants responded overall that they looked first for aesthetics, then meaning, then representations of reality.

**Discussion**

Participants brought a diverse range of understanding of photographic images to the viewing exercise. This understanding often informed their detection of image manipulation. Most significantly, participants who were aware that the research involved photo credibility were more successful in identifying manipulations (Figure 1), which may have resulted from them viewing the photographs with a more critical eye (that is, perhaps they were more likely to consciously pay attention to the results of their non-conscious identification of changed regions of the photograph).

Knowledge and life experiences generally played a role in participants understanding the meaning of the photographs they viewed. For example, of the 12 participants, only 1 articulated the connection between the image of Queen Elizabeth spliced into the photograph of then Prime Minister John Howard in Figure 4 above (John Howard’s well known monarchist views on whether Australia should be a Republic). Three subjects were unable to identify John Howard at all and focussed on the Queen or the media aspects of the image.
It was expected that manipulations of larger sizes would be spotted more readily than manipulations of smaller sizes, this was not the case. There was no obvious correlation between the size of the manipulation and the accuracy rate (Figure 3). In an image used in all three sets, in which it appears three people are about to jump over a pier rail next to a sign that reads “JUMPING FROM PIER” (from which the word ‘NO’ was erased from the image) the overall size of the manipulation was only 0.5% yet the accuracy rating of participants identifying the manipulation was 25%. At the same time, an image of missiles in which 13% of the image was an additional spliced image had a 0% success rate of manipulation identification.

Given the tendency of participants to rationalise elements in images (discussed below), it may be that a more defining characteristic of more easily discernible manipulations is their saliency, i.e. the extent to which the elements added to or removed from an image contribute to the understanding of that image. This suggests that further experimentation to tease out the differences between apparent size and saliency impact may be useful.

While all participants stated at a level of 8.25 or over out of 10 that they cared about photo credibility, societal implications of photo manipulation, and the lack of photo authentication solutions, their verbal exposition when discussing the meaning of the photographs presented to them indicated that they were more likely to justify the oddness of the image than to question it.

This was true even when the subject of image manipulation had been discussed moments earlier. Short-term increased awareness of image manipulation issues (as represented by participation in the experimental study) appeared to have little effect in conditioning participants to look at photographs critically. Four participants who had been assigned to the untrained group prior to the experiment participation were also asked to view 8 additional images, 2 of which were unmanipulated and 6 of which were manipulated similarly to the common part of the experiment. Three
responded “no” in response to the query “Do you believe that any of these eight photos has been faked?” The 4th subject responded yes but could not identify more than one faked photograph from the 6 presented.

For example, although common sense would dictate that the photograph in Figure 5 had been manipulated, participants explained it away with justifications such as “maybe the car was warmer to sleep on than the snow,” “relates to the use of leather in cars,” or an inspirational message of unknown origin “don’t think that anything is impossible.”

This matches the uncertainty of responses to the survey question about ways to identify photograph manipulations, as well the verbal exposition and eye gaze data in which participants often used words indicating uneasiness with a photograph as they looked at manipulated elements in images, such as one participant commenting “that shouldn’t be up there” as her eye gaze rested on the cow in the image at Figure 5.

It was noted that a ‘hiding’ effect occurred when additional, less obvious alterations were included in an image. In Figure 6 an image of a field of coins was presented to subjects for the first time in Set 2 of the experiment wherein subjects were offered both a presentation image and the original for comparison. Although most (11 of 12) noted that a coin had been added to the original, only one participant noted that another coin in the image had been rotated 180 degrees. This was despite eye gaze tracking identifying that subjects looked at the rotated coin more (2.3% of area within photo) than the added coin (1.6% within photo).

![Figure 7: Subjects seek second manipulation when offered a difference map](image)

When the coins photograph was presented again in the third phase of the experiment in which they were offered a difference map identifying changed pixels (Figure 7), all subjects used an eye gaze strategy that compared the pixels demarcated in the difference map as changed to
find and identify the rotated coin in the presentation image by comparison with the original image.

In some cases the use of the difference map in Set 3 enabled participants to confirm previously identified differences noted in Set 2, somewhat like an answer key. In other cases the difference map provided participants with information that enabled them to identify that there were differences or additional differences in images they had previously passed as unchanged. Overall, using the image configuration in Set 3 provided the greatest level of accuracy (97.2% manipulated images identified, 90.1% of all manipulated regions identified) and speed (5.3 seconds on average).

It may be that one reason subjects are more likely to explain than to question manipulated images is that they want to believe they can spot fakes and therefore seek alternative explanations for unlikely elements in the images. Conversely, it may also be the case that subjects feel they can’t spot fakes and therefore focus on and rationalise the meanings of the images.

These rationalisations can be quite unexpected, as in the case of a participant attempting to explain the juxtaposition of a trio of sea anemones and a frozen pond (Figure 8). In this case the participant focussed on the anemones being sea creatures and reinterpreted the snow on the edging stones as salt.

It is worth noting that the co-existence of high levels of concern about photo manipulation and the tendency to justify rather than identify manipulated images is dichotomous. It may be that photographs have not yet shed their cachet of being representations of reality; subjects are conditioned to look at images as ‘real’. This would also be useful to examine in future experiments.

**Summary**

Comparing eye gaze tracking to question responses reveals that subjects may see more of the changes in manipulated images than they consciously report.

It is not necessarily the case that larger manipulations are more easily seen. The saliency of the manipulation may influence the identification success rate, and this bears further investigation.

It appears that it may be that when an image has an obvious manipulation, other lower profile manipulations may not be consciously identified even when they are viewed by the eye gaze.
Subjects who were aware that this research involves manipulated images (the informed group) looked more critically for possible manipulations and performed better when presented with both standalone image and comparison images than those who were uninformed.

Exposing participants to additional standalone images subsequent to the experiment proper did not result in improved performance comparable to the ‘informed’ cohort in identifying manipulations.

There is a dichotomy between the high levels of concern expressed about photo manipulation and the tendency of participants to explain away manipulations instead of identifying them.

The two levels of MSCI image bundling assisted subjects in perceiving image manipulations more accurately and quickly.

**Attributions**

All photographs and photo alterations by Sabrina Caldwell with the following exceptions:

Queen Elizabeth II (photograph) used in John Howard photograph.

Cow on BMW. Originally tweeted by Surrey Roads Police. Prior origin unknown. Source: www.flickr.com/photos/96057563@NO2/9430346171/. Accessed 17 January 2014. Work is assumed to be copyrighted and used under the fair use provisions of copyright law.

**References**


Henderson, John M. Human gaze control during real-world scene perception. TRENDS in Cognitive Sciences Vol. 7 No. 11 November 2003


Seeing Machines. seeingmachines.com

Evaluation of e-Readers: A Preliminary Analysis

Ujala Rampaul and Tom Gedeon

Research School of Computer Science
Australian National University
{u5051329 | tom.gedeon}@anu.edu.au

Abstract. Evaluation of new consumer goods, particularly electronic, are often done from the perspective of analysing the latest models, comparing their advantages and disadvantages with respect to price. This style of evaluation is often performed by one of a few product experts on a wide range of features that may not be applicable to each user. This study instead used a scenario-based approach to evaluate a number of e-readers, mimics a user interested in a new product or technology with a limited budget. The objective is to evaluate from a purchasing perspective the quality and usability of e-readers available within that budget range. The e-readers evaluated were operated in multiple ways, which suggests that the interface design should cater for users with different levels of experience with technology. The results of a large user study with over 70 participants, shows that the popular brands do not necessarily produce the best products in terms of e-readers. We found that navigation within books to be the most significant differentiator between the eReaders in our scenario based evaluation process.

1 Introduction

In 2007, Amazon launched the Kindle in the United States and it sold out in under 6 hours (Patel, 2007). Prior to the Kindle launch there was not a huge demand for e-readers as earlier releases were the Rocket eBook in 1998, the Sony Libriè in 2004 and Sony Reader in 2006. These earlier devices were unsuccessful, perhaps because they were expensive, had technological limitations and lacked available content. The Kindle was different because it was affordable, consumers could purchase a wide variety of e-books via the wireless data connection and it was comparatively easy to use.

E-readers with e-ink screens simulate the experience of reading a paper book and are not multifunctional devices. The advantage of an e-reader is that it is easier to read even in direct sunlight, does not consume a lot of battery, is lighter in weight, permits undisturbed reading of an e-book and the eyes do not fatigue compared to devices with backlit LCD displays.

In the United States, the cost of e-readers has continued to fall, making them more accessible to purchase. Data from Pew Research Centre over a three-year period from May 2010 to January 2014 indicate e-reader ownership has grown over 20%, see Figure 1 on ownership growth of e-readers and tablets in the United States. Overall, 50% of Americans now have a dedicated handheld device – either a tablet computer like an iPad, or an e-reader such as a Kindle or Nook for reading e-content, and \( \frac{1}{3} \) of US adults now own an eReader (Zickuhr and Rainie, 2014).

![Figure 1: eReader ownership in the US](image-url)
The motivation of this project was to evaluate from a purchasing perspective the quality of e-readers that could be purchased within a limited budget. Shoppers of e-readers and e-reading devices are often influenced by advertising, reviews conducted by computer and consumer sites such as PC Authority and Choice, as well as opinions of friends and families. The commercial reviews compare the performance, features, file formats and specifications of the latest e-reader models against the market price to identify the best value for money.

To obtain e-readers for this study a budget based on the cost of a new paper book ($25) was established and multiplied to set the budget range of approximately $50-75 (excluding shipping costs) for the purchase of an e-reader. Establishing a budget reduced the range of e-readers that could be obtained i.e. either, earlier, e-reader models with a higher specification or later models with a relatively low specification. We argue that if we focused on the quality of the devices available within this budget range, then the differences in the e-reader model and specifications were not limiting factors. The experiment requires that a realistic range of e-readers of a high quality could be obtained from eBay and other auction sites for an effective comparison.

The following nine e-readers were selected and mostly purchased from within Australia except for the Sony and jetBook, at a lower price range than originally estimated being $35 -$70. This includes a selection of e-readers having buttons only, touch screen only or a combination of both.

<table>
<thead>
<tr>
<th>e-reader</th>
<th>Cost in AUD$</th>
<th>Study device label</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amazon Kindle 4G</td>
<td>57.85</td>
<td>C</td>
</tr>
<tr>
<td>Barnes &amp; Noble Nook</td>
<td>69.90</td>
<td>A</td>
</tr>
<tr>
<td>Ectaco jetBook</td>
<td>61.28</td>
<td>G</td>
</tr>
<tr>
<td>Elonex 621EB</td>
<td>41.00</td>
<td>F</td>
</tr>
<tr>
<td>EZReader EZ601</td>
<td>55.00</td>
<td>I</td>
</tr>
<tr>
<td>iRiver Story HD</td>
<td>52.00</td>
<td>E</td>
</tr>
<tr>
<td>Kobo Touch</td>
<td>60.00</td>
<td>B</td>
</tr>
<tr>
<td>Pandigital Novel</td>
<td>36.85</td>
<td>H</td>
</tr>
<tr>
<td>Sony PRS-600</td>
<td>44.88</td>
<td>D</td>
</tr>
</tbody>
</table>

Table 1. Cost of second-hand e-readers

2 Background

The rise of e-books and e-readers are part of a larger story about the shift from printed to digital material. Early research focused on academic contributions towards education where students used e-readers in the classrooms to read textbooks.

There have been numerous approaches taken when evaluating both ebooks and ebook readers in projects such as Superbook and Electronic Books On-screen interface (EBONI). (Gibson & Gibb, 2011).

Other studies included university and public libraries investigating strategies to support the borrowing of e-books and e-readers. “In particular, many academic libraries since begun pilot projects using a variety of different reader devices to investigate the possibilities for simplifying and innovating the content and related services they offer to users in light of the technologies that are available. Countless libraries have experimented with offering lending programs for their devices. Many libraries have undertaken these initiatives on their own whereas others have partnered with either Sony electronics or Amazon to design and carry out their projects” (Behler & Lush, 2010).
To establish design guidelines for e-readers (Pearson, Buchanan, & Thimbleby, 2010) conducted a study to determine the usability concerns with user interface of the e-readers using human computer interaction (HCI) principles. The experiment was carried out using the Kindle 2, Sony PRS 600 and Sony PRS 300. These three e-readers had similar screen type (e-Ink technology), screen size (six inches) and resolution (600 x 800). The Sony PRS 600 was a combination of touch and button device whereas the Kindle 2 and Sony PRS 300 were button devices.

Based on the guidelines and principles for ergonomics, consistency, completeness, page numbering, book marks, annotation and magnification, the Kindle was found to be better than Sony PRS-300 and PRS-600. The positioning of the buttons in the Kindle made it easier to turn pages for both left and right-handed users. The Kindle had a full QWERTY keyboard which supported web browsing and location identifiers in relation to the file instead of page numbers. The availability of full annotation and zooming was only available on the Sony PRS-600. User feedback on the three e-readers was leveraged from Amazon’s online customer review and weighted on a Likert scale rating. This could have attributed for the higher ratings for the Kindle, as it is produced by Amazon.

The study by Gibson and Gibb (2011) evaluated four second-generation e-reading devices, namely Sony PSRS505; Bookeen Cybook 3g; iRex iLiad and Asus Eee PC 105 HA, with 33 participants. Their evaluation measured weight, quality of display, size and robustness of the devices based on a five point Likert scale rating.

On overall impressions and functionality, the Asus Eee netbook was the best, followed by Sony PRS 505, iLiad and lastly Cybook. The Asus Eee netbook was the preferred choice because participants were familiar with the design and layout compared to the dedicated e-readers (Gibson & Gibb, 2011).

Both the Sony PRS505 and Booken Cybook were similar in screen size (six inches) and resolution (600 x 800 pixels) whereas the iRex iLiad had a larger screen size (eight inches) and a larger resolution (768 x 1024 pixels). All three devices had screens with e-Ink technology. The Asus Eee PC 105 HA, a netbook had a LCD screen, a much larger screen size (ten inches) as well as a large resolution (600x1024 pixels) and weighed seven times more than the Booken Cybook, the lightest (175g) device in the trial. This meant the refresh rate for turning the page were faster in the netbook compared to the e-readers using e-Ink technology. This was an unequal comparison between three dedicated e-readers and a netbook.

Usability was favourable towards the e-readers for lightness and portability, readability and ease of use. Participants commented, “the screen was not wearing on the eyes” (iLiad) and that it was a “straightforward operation” (Sony), and that “the non-glare screen made the text as easy to read as ink” (iLiad) (Gibson & Gibb, 2011).

Based on reported sales and market share in the United States, Richardson & Mahmood (2012) studied five leading e-readers namely Kindle 3G; iPad 1G; Nook; Kobo N647; and Sony PRS950. Their objective was to identify the advantages and disadvantages of e-readers and “to compare and contrast the most popular (i.e. bestselling) devices against a comprehensive, if not exhaustive, set of technical specifications as well as qualitative judgment of users” (Richardson & Mahmood, 2012).

The results from Richardson and Mahmood’s study showed that the Kindle was the most popular device, though some participants commented on poor navigation. Some 47% of the participants that undertook this study owned a Kindle, accounting for a probably biased popularity of the Kindle. There was no evidence for the technical comparison nor did they elaborate what it consisted of. The comparison between the iPad and the e-readers was probably unjust as the tablet is a multi-function device, whereas e-readers are designed to read and purchase e-books.

The NextMedia eReading Project conducted a qualitative and quantitative study on reading behaviours and patterns on e-readers including the effects of age groups. The objective of the main project was to encourage Finnish consumers to adopt and grow an e-reading community. This eReading Project consisted of multiple smaller projects that focused on different aspect of the study such as usability of e-reading devices, types of content, emotional and cognitive reactions.

Heikkilä (2011) usability study of e-reading devices was based on a scenario, which involved the participants recording their experiences of opening and using a newly purchased e-reading device.
A number of pages from a paper book was the benchmark for their evaluations. The same content being available on the e-reading devices. Over a week, seventeen participants recorded their reading times, places they had read and their experiences on using the e-readers. The time-based assessment tasks included opening the device; finding a book; finding a specific spot in the book and changing the font size. Heikkilä used the following e-readers in his study Kindle (included both Kindle 2 and Kindle 3); Sony PRS600; iPad 1G; BeBook Neo; Booken Cybook Opus; Elonex eBook and Samsung galaxy.

The outcome of the usability study was the creation of a conceptual model of an e-reader labelled Acola. The Acola was a combination of touch sensitive and gesture savvy pad on an e-ink device. The pad would handle page turning and skimming. Swiping to the left would turn the page forward, swiping to the right would move one page backwards. Swiping fast two times successively, would turn two pages, three times: three pages etc. Swiping with two fingers simultaneously would move you between chapters. A menu and an OK button could be situated in the upper and lower border of the pad (Heikkilä, 2011).

There is some similarity between Heikkilä’s (2011) study and this study in the use of a scenario for the evaluation of the e-reading devices. It is interesting that his study evaluated two tablets and four e-readers to create their ideal e-reader.

Siegenthaler, et al., (2012) conducted a study to with the following reading devices Sony PRS505, Sony PRS600 and an Apple iPad, to determine effects of touch screen technology. Twelve participants tested the three devices sequentially within a session based on a set of time related tasks such as open a book, open a specific page within the book, highlight a sentence, find a highlighted sentence, delete a highlighted sentence, change the orientation of the layout and increase the font size. The participants rated the above tasks on the navigation, design, handiness and handling based on a Likert scale rating.

The results showed “that e-reading devices with touch screens correlate with better navigation rating” and touch screen technology also has its advantages in terms of a higher intuitiveness and flexibility for adaptations of the navigation (e.g., due to firmware updates) compared to devices with static buttons. (Siegenthaler, Bochud, Wurtz, Schmid, & Bergamin, 2012). The evaluation of handiness and handling was ambiguous, for example the related questions were “How handy do you rate the reading device?” and “how easy was it for you to handle the reading device?” There was no clarification to what this meant and the evaluation questions were vague.

We earlier identified the differences between an e-reader, which is solely for reading e-books whereas a tablet is a multifunction device making this evaluation an unequal comparison.

Siegenthaler, et al, (2012) argued, “E-ink technology has low power consumption, thereby increasing battery life and allowing for a more lightweight device. Another advantage is that e-ink devices can be used outside without glare being a big issue. However, e-ink screens have some disadvantages, most of them are black and white and the pages do not refresh as quickly as devices with an LCD screen”.

The disadvantage of tablets is that they cost considerably more than e-readers, the LCD screens are susceptible to glare, they do not provide a comfortable reading experience, are heavier and the battery does not last as long.

E-ink technology is designed to emulate printed books. Independent studies on the reading behaviours measured by eye tracking found no difference between a paper book and a dedicated e-reader (Nielsen, Kindle 2 Usability Review, 2009) and (Siegenthaler, Wurtz, & Groner, Improving the usability of e-book readers, 2010).

Similarly, analysis of the reading behaviours between a dedicated e-reader and an e-reading device (tablet) found no significant difference as measured by eye speed (Nielsen, iPad and Kindle Reading Speeds, 2010) and (Siegenthaler, Wyss, Schmid, & Wurtz, 2012). However, the iPad “actually scores slightly higher in user satisfaction” (Nielsen, iPad and Kindle Reading Speeds, 2010).

Please see Figure 2 which maps the reading devices that are common to this study and to other studies identified. Most studies shared one common e-reader, for example Richard & Mahmood
(2010) share the jetBook and Pearson et al., (2010) share the Sony PRS-600. The exception is
Heikkilä (2011) where there two common e-readers namely the Sony PRS-600 and the Elonex.

Please see Figure 3 which maps the tasks that are common to this study and to other studies
identified. All studies share at least one common task, with Siegenthaler et al, (2010) and (2012)
evaluating two similar tasks namely, open a document (open a book) and increase font size. The
other exception is Heikkilä (2011), where four similar tasks were evaluated that is turn the power on
(open device), navigate to first document (find a particular book), increase font size (change font
size) and navigate to specific section within the book (find a specific place within the book). All
studies have commented on readability.

No previous studies have been conducted to evaluate from a purchasing perspective the quality
and usability of e-readers purchased within a budget of up-to $70. This study proposes to conduct
the evaluation of the e-readers based on a scenario with two experiments.

1. Evaluation of nine e-readers that are button interface devices only, touch screen only and a
   combination of both to determine which device is the best
2. Evaluation of e-readers with touch screens

3 Methodology

The Human Research Ethics Committee at the Australian National University (ANU) approved
this study. Participants from the ANU College of Engineering and Computer Science were recruited
for the experiments.

The participants evaluated the devices in pairs so one could act as the scribe and observe, while
the other operated the device. On completing the evaluation of the first device, the participants
swapped roles to evaluate the next device. This grouping encouraged discussion amongst
themselves on their observations and interactions with the e-readers. In total, 72 senior students (3rd
year HCI course participants, Honours, Masters and PhD) took part in the evaluations. Please see
Table 1 for the list of 9 devices evaluated. Each device was allocated an alphabetical letter based on
the arrival sequence for easy identification in the randomisation matrix.

The randomisation was similar to a Latin square where two sets of the e-readers were labelled
from A to I and arranged in such a way that no row contained the same letter twice. The pairs were
balanced, so that the number of times a letter came first or second was the same. There were 36
pairs of e-readers for 72 participants to evaluate, which meant each device was evaluated four times
as a first device and then four times as a second device. Table 2 shows an example of the device ‘A’
evaluated eight times with the other devices.

<table>
<thead>
<tr>
<th>BA</th>
<th>BH</th>
<th>HE</th>
<th>IF</th>
</tr>
</thead>
<tbody>
<tr>
<td>GE</td>
<td>CA</td>
<td>BI</td>
<td>GC</td>
</tr>
<tr>
<td>DC</td>
<td>IE</td>
<td>DA</td>
<td>BG</td>
</tr>
<tr>
<td>DI</td>
<td>EC</td>
<td>CH</td>
<td>EA</td>
</tr>
<tr>
<td>AF</td>
<td>HD</td>
<td>CF</td>
<td>HF</td>
</tr>
<tr>
<td>CB</td>
<td>AG</td>
<td>GD</td>
<td>CI</td>
</tr>
<tr>
<td>GF</td>
<td>DB</td>
<td>AH</td>
<td>ED</td>
</tr>
<tr>
<td>FD</td>
<td>IH</td>
<td>EB</td>
<td>AI</td>
</tr>
<tr>
<td>IG</td>
<td>FE</td>
<td>HG</td>
<td>FB</td>
</tr>
</tbody>
</table>

Table 2. Matrix of the e-reader randomisation pair for device A.

The evaluation commenced when each participant pair were given the first e-reader from a
randomised pair, and a questionnaire. The participants received an overview of the study and its
objectives. After performing the tasks on the first e-reader, the tasks were repeated on the second e-reader.

The questionnaire included a scenario, the evaluation tasks based on Likert-like numerical scale, open-ended questions to gain an insight on improving the device and comments after each device evaluated. It also gathered optional personal details including the name, gender, age, year of study, familiarity with e-readers and follow-up contact details for purposes of clarification or removal from the experiment in adherence with The Human Research Ethics Committee guidelines of the Australian National University. Lastly, the questionnaire requested the subjective opinions and reasons on whether one of the two devices was preferred.

The evaluations were designed around the same scenario to provide context to the experiment. The scenario was based on a user finding a gift received a couple months ago. Included in the box was a note to register the device prior to use. The participant had to register the device without the assistance of the user manual, which was based on a series of tasks, to be accomplished with the assistance of a fellow student, who had similarly received an e-reader as a gift. For both experiment, the participants had to assume the battery on the e-readers were fully charged, as it was not a requirement of the scenario.

The tasks was measured on a Likert-like summative scale rating ranging from the most positive to the least positive:

- 5 = Very good
- 4 = Good
- 3 = Ok
- 2 = Bad
- 1 = Very bad

The tasks were
1. To turn the device on
2. Navigate to a document (called “Somedevice XY User Guide”)
3. Open the document
4. Increase the font size
5. Navigate to specific section to find the ‘model number’
6. Navigate to a second document (“Dates-2013” and use the model number).
7. Assess readability on the screen

Explicit instructions were given to not use the recent history, continue reading or the ‘date’ search in the e-readers. Most of the e-readers provided such functionality, which manufacturers seem to consider of such high utility that in general it is not possible to turn off these modes, which makes a repeated experiment such as ours impossible without this instruction to not use such features.

Additional subjective open-ended questions were requested, on likes and dislikes about the e-readers; previous use or ownership of an e-reader and the model and if they were regular users of a smartphone; tablet or laptop. These questions were to gain a better understanding into the user’s experience with mobile devices.

4 Results

Statistical analysis was performed using F-statistics based on a repeated measures ANOVA for the within factor of e-readers. A critical p < 0.05 was used for statistical significance in all analyses. Each of the following sections report the ANOVA results for that question based on the usability tasks measured on a Likert-like numerical scale rating from 1 (very bad) to 5 (very good) on the e-readers.
Figure 1. Map of common reading devices between this study and previous study.

Pearson et al., (2010)
- Kindle 2
- Sony PRS-600
- Sony PRS-300

Heikkilä (2011)
- Kindle 2&3
- Sony PRS-600
- BeBook Neo
- Booken Cybook Opus
- Elonex eBook
- Samsung galaxy tablet
- iPad 1G

ANU Experiment 1
- A. Barnes & Noble Nook
- B. Kobo Touch
- C. Amazon Kindle 4G
- D. Sony PRS-600
- E. iRiver Story HD
- F. Elonex eBook
- G. Eclaco jetBook
- H. Pandigital Novel
- I. EZReader EZ601

Richardson & Mahmood (2012)
- Sony PRS-900
- iPad 1G
- Kobo N647
- Nook
- Kindle 3G

Siegenthaler et al (2012)
- Sony PRS-600
- Sony PRS-505
- iPad 1G

Gibson & Gibb (2011)
- Sony PRS-505
- Booken Cybook
- iRex iLiad
- Asus Eee PAC105HA

Siegenthaler et al (2010)
- Sony PRS-505
- Booken Cybook Gen
- iRex iLiad
- BeBook
- Eclaco jetBook
Figure 2. Map of tasks between this study and previous studies (all comment on readability)

**Person et al., (2010)**
- bookmarking
- annotation
- page turning/navigation
- magnification (zoom)

**Heikkilä (2011)**
- open device
- find a particular book
- find a specific place in a book
- change font size
- page orientation

**Nielsen (2010)**
- navigation
- readability

**Gibson & Gibb (2011)**
- portability
- navigation
- screen glare
- zoom

**ANU experiment 1**
1. turn power on
2. navigate to 1st document
3. open document
4. increase font size
5. navigate to specific section
6. navigate to 2nd document
7. readability

**Richardson & Mahmood (2012)**
- portability
- navigation
- readability

**Siegenthaler et al. (2012)**
- open book
- open page
- font size
- highlight sentence
- find highlighted sentence
- delete highlighted sentence
- page orientation

**Siegenthaler et al., (2010)**
- open a book
- increase font size
- page orientation
- open an audio file
- open a picture
- readability

*Indicates links to other studies*
4.1 Turning on the e-reader

There were no significant differences found in the responses to the question on powering on the e-readers. Table 3 shows the results and ranks for the e-readers. The jetBook, Pandigital and the Kobo are ranked at the top. There are four e-readers (i.e. Nook, iRiver, Elonex and EZReader) ranked at the middle in equal fourth. The Sony and the Kindle were ranked at the bottom, however there were no significant differences found in the e-readers for turning them on.

<table>
<thead>
<tr>
<th>e-Reader</th>
<th>Mean</th>
<th>SD</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>jetBook</td>
<td>4.1</td>
<td>0.83</td>
<td>1</td>
</tr>
<tr>
<td>Pandigital</td>
<td>4.0</td>
<td>0.76</td>
<td>2</td>
</tr>
<tr>
<td>Kobo</td>
<td>3.9</td>
<td>1.17</td>
<td>3</td>
</tr>
<tr>
<td>Nook</td>
<td>3.8</td>
<td>1.28</td>
<td>=4</td>
</tr>
<tr>
<td>iRiver</td>
<td>3.8</td>
<td>1.28</td>
<td>=4</td>
</tr>
<tr>
<td>Elonex</td>
<td>3.8</td>
<td>0.71</td>
<td>=4</td>
</tr>
<tr>
<td>EZReader</td>
<td>3.8</td>
<td>1.28</td>
<td>=4</td>
</tr>
<tr>
<td>Sony</td>
<td>3.4</td>
<td>1.06</td>
<td>8</td>
</tr>
<tr>
<td>Kindle</td>
<td>3.1</td>
<td>0.83</td>
<td>9</td>
</tr>
</tbody>
</table>

Table 3. Mean Ranking and Standard Deviation (SD) for turning the e-readers on

4.2 Finding the first document

There were no significant differences found in response to the question on finding the first document. Table 4 shows the results and ranks for the e-readers. Sony and Kobo ranked in the top for finding a document whereas the iRiver, Pandigital, jetBook, and Nook were ranked in the middle. EZReader and Kindle were ranked at the bottom but there was no significant difference found in all e-readers. Note that each eReader was pre-loaded with over 70 books downloaded from the Project Gutenberg (2013) website of books now in the public domain. The selection was eclectic, consisting of popular fiction (e.g. Burroughs (1917)), mythology (Werner, 1922), craft (de Dillmont, 1886), history (Gibbon, 1776-1789) and classical philosophy (Plato, 350 BC).

<table>
<thead>
<tr>
<th>e-Reader</th>
<th>Mean</th>
<th>SD</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sony</td>
<td>4.0</td>
<td>0.53</td>
<td>1</td>
</tr>
<tr>
<td>Kobo</td>
<td>3.5</td>
<td>1.41</td>
<td>2</td>
</tr>
<tr>
<td>iRiver</td>
<td>3.3</td>
<td>1.16</td>
<td>=3</td>
</tr>
<tr>
<td>Pandigital</td>
<td>3.3</td>
<td>1.49</td>
<td>=3</td>
</tr>
<tr>
<td>jetBook</td>
<td>3.0</td>
<td>1.31</td>
<td>5</td>
</tr>
<tr>
<td>Nook</td>
<td>2.9</td>
<td>1.46</td>
<td>6</td>
</tr>
<tr>
<td>Elonex</td>
<td>2.5</td>
<td>0.76</td>
<td>7</td>
</tr>
<tr>
<td>EZReader</td>
<td>2.4</td>
<td>1.60</td>
<td>8</td>
</tr>
<tr>
<td>Kindle</td>
<td>2.3</td>
<td>0.89</td>
<td>9</td>
</tr>
</tbody>
</table>

Table 4. Mean Ranking and Standard Deviation (SD) for finding the document
4.3 Opening the document

An analysis of variance showed that the responses to the question on opening the document was significantly different between the e-readers, $F(8,63) = 3.21$, $p = 0.004$. Table 5 shows the results and rankings for opening the document. The Kobo, Kindle and the iRiver were ranked at the top. The Sony and Elonex were ranked equal fourth. The jetBook and Pandigital ranked equal sixth. The EZReader and Nook rank at the bottom. This shows there is a significant difference in the e-readers to perform the task of opening a document.

<table>
<thead>
<tr>
<th>e-Reader</th>
<th>Mean</th>
<th>SD</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kobo</td>
<td>4.6</td>
<td>0.52</td>
<td>1</td>
</tr>
<tr>
<td>Kindle</td>
<td>4.5</td>
<td>0.93</td>
<td>2</td>
</tr>
<tr>
<td>iRiver</td>
<td>4.5</td>
<td>0.76</td>
<td>2</td>
</tr>
<tr>
<td>Sony</td>
<td>4.4</td>
<td>0.74</td>
<td>4</td>
</tr>
<tr>
<td>Elonex</td>
<td>4.4</td>
<td>0.74</td>
<td>4</td>
</tr>
<tr>
<td>jetBook</td>
<td>4.3</td>
<td>1.39</td>
<td>6</td>
</tr>
<tr>
<td>Pandigital</td>
<td>4.3</td>
<td>0.89</td>
<td>6</td>
</tr>
<tr>
<td>EZReader</td>
<td>3.5</td>
<td>1.20</td>
<td>8</td>
</tr>
<tr>
<td>Nook</td>
<td>2.8</td>
<td>1.16</td>
<td>9</td>
</tr>
</tbody>
</table>

Table 5. Mean Ranking and Standard Deviation (SD) for opening the document: **significant**

4.4 Increasing the font size

An analysis of variance showed that increasing the font size was significantly different between the e-readers, $F(8,63) = 2.25$, $p = 0.03$. Table 6 shows the results and rankings on the e-readers for increasing the font size. The Pandigital, jetBook and iRiver were ranked at the top three positions. The Kindle and Elonex were ranked equal fourth. The Nook was ranked at sixth position. The Kobo and EZReader were ranked at the bottom two positions as equal eighth. This shows a significant difference was found in the e-readers to perform the task of increasing the font size.

<table>
<thead>
<tr>
<th>e-Reader</th>
<th>Mean</th>
<th>SD</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pandigital</td>
<td>4.6</td>
<td>0.52</td>
<td>1</td>
</tr>
<tr>
<td>jetBook</td>
<td>4.1</td>
<td>1.13</td>
<td>2</td>
</tr>
<tr>
<td>iRiver</td>
<td>4.0</td>
<td>1.07</td>
<td>3</td>
</tr>
<tr>
<td>Kindle</td>
<td>3.9</td>
<td>1.46</td>
<td>4</td>
</tr>
<tr>
<td>Elonex</td>
<td>3.9</td>
<td>1.13</td>
<td>4</td>
</tr>
<tr>
<td>Nook</td>
<td>3.4</td>
<td>1.06</td>
<td>6</td>
</tr>
<tr>
<td>Sony</td>
<td>3.3</td>
<td>0.71</td>
<td>7</td>
</tr>
<tr>
<td>Kobo</td>
<td>2.8</td>
<td>1.75</td>
<td>8</td>
</tr>
<tr>
<td>EZReader</td>
<td>2.8</td>
<td>1.49</td>
<td>8</td>
</tr>
</tbody>
</table>

Table 6. Mean Ranking and Standard Deviation (SD) for increasing font size: **significant**
4.5 Navigate to specific section for find the ‘model number’

An analysis of variance showed that navigating to a specific section within the document was significantly different between the devices, $F(8,63) = 3.43, p = 0.002$. Table 7 shows the results and rankings on the e-readers for navigating to a specific section within the document. The jetBook, Pandigital and the Kindle ranked at the top. The Sony, iRiver and the Kobo ranked in the middle and the Elonex, Nook and EZReader ranked in the bottom positions. This shows a significant difference in the e-readers for navigating to specific section within a document.

<table>
<thead>
<tr>
<th>e-Reader</th>
<th>Mean</th>
<th>SD</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>jetBook</td>
<td>4.0</td>
<td>0.93</td>
<td>1</td>
</tr>
<tr>
<td>Pandigital</td>
<td>3.6</td>
<td>1.19</td>
<td>2</td>
</tr>
<tr>
<td>Kindle</td>
<td>3.5</td>
<td>1.07</td>
<td>3</td>
</tr>
<tr>
<td>Sony</td>
<td>3.4</td>
<td>1.41</td>
<td>4</td>
</tr>
<tr>
<td>iRiver</td>
<td>3.3</td>
<td>1.28</td>
<td>5</td>
</tr>
<tr>
<td>Kobo</td>
<td>2.8</td>
<td>1.39</td>
<td>6</td>
</tr>
<tr>
<td>Elonex</td>
<td>2.6</td>
<td>1.30</td>
<td>7</td>
</tr>
<tr>
<td>Nook</td>
<td>1.9</td>
<td>1.46</td>
<td>8</td>
</tr>
<tr>
<td>EZReader</td>
<td>1.6</td>
<td>0.92</td>
<td>9</td>
</tr>
</tbody>
</table>

Table 7. Mean Ranking and Standard Deviation (SD) for navigation to section 3: significant

4.6 Navigate to a second document

There was no significant differences found in navigating to a second document with the e-readers. Table 8 shows the results and ranks for navigating to a second document. The iRiver, jetBook and the Sony ranked in the top three positions. The Pandigital, Elonex and the Kobo ranked at the middle and the Nook, Kindle and EZReader ranked at the bottom three positions. Navigating to a second document was not significantly different in the e-readers.

<table>
<thead>
<tr>
<th>e-Reader</th>
<th>Mean</th>
<th>SD</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>iRiver</td>
<td>4.0</td>
<td>0.93</td>
<td>1</td>
</tr>
<tr>
<td>jetBook</td>
<td>3.9</td>
<td>1.81</td>
<td>2</td>
</tr>
<tr>
<td>Sony</td>
<td>3.8</td>
<td>1.04</td>
<td>3</td>
</tr>
<tr>
<td>Pandigital</td>
<td>3.6</td>
<td>0.92</td>
<td>4</td>
</tr>
<tr>
<td>Elonex</td>
<td>3.4</td>
<td>1.69</td>
<td>5</td>
</tr>
<tr>
<td>Kobo</td>
<td>3.1</td>
<td>1.64</td>
<td>6</td>
</tr>
<tr>
<td>Nook</td>
<td>2.9</td>
<td>1.46</td>
<td>7</td>
</tr>
<tr>
<td>Kindle</td>
<td>2.6</td>
<td>1.41</td>
<td>8</td>
</tr>
<tr>
<td>EZReader</td>
<td>2.5</td>
<td>1.31</td>
<td>9</td>
</tr>
</tbody>
</table>

Table 8. Mean Ranking and Standard Deviation (SD) for navigation to second document

4.7 Readability on the screen

There were no significant differences found in response on assessing the readability. Table 9 shows the results and ranks for readability. The iRiver was ranked the highest, followed by Kobo, Kindle
and jetBook in second position and the Pandigital in the fifth. The Nook and EZReader were both ranked equal sixth. The Elonex and Sony ranked lower but no significant difference were found between the e-readers for readability.

<table>
<thead>
<tr>
<th>e-Reader</th>
<th>Mean</th>
<th>SD</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>iRiver</td>
<td>4.5</td>
<td>0.76</td>
<td>1</td>
</tr>
<tr>
<td>Kobo</td>
<td>4.0</td>
<td>0.53</td>
<td>=2</td>
</tr>
<tr>
<td>Kindle</td>
<td>4.0</td>
<td>0.93</td>
<td>=2</td>
</tr>
<tr>
<td>Jetbook</td>
<td>4.0</td>
<td>0.93</td>
<td>2</td>
</tr>
<tr>
<td>Pandigital</td>
<td>3.9</td>
<td>0.35</td>
<td>5</td>
</tr>
<tr>
<td>Nook</td>
<td>3.8</td>
<td>0.71</td>
<td>=6</td>
</tr>
<tr>
<td>EZReader</td>
<td>3.8</td>
<td>1.28</td>
<td>=6</td>
</tr>
<tr>
<td>Elonex</td>
<td>3.5</td>
<td>1.20</td>
<td>8</td>
</tr>
<tr>
<td>Sony</td>
<td>3.3</td>
<td>1.28</td>
<td>9</td>
</tr>
</tbody>
</table>

Table 9. Mean Ranking and Standard Deviation (SD) for readability

4.7 E-Reader preference

For the question, “Did you like one of the two devices better?” an analysis of variance showed that the responses were significantly different between the e-readers, $F(8,54) = 2.33, p = 0.03$. Table 10 shows the results and rankings of the participants’ preference when comparing e-readers. The Kindle and iRiver ranked equal first, followed by Sony, jetBook and Pandigital in equal third position. The Elonex was ranked sixth, Kobo seventh and both Nook and EZReader ranked at the bottom. This indicates a significant difference in the e-readers that participants’ preferred the most.

<table>
<thead>
<tr>
<th>e-Reader</th>
<th>Mean</th>
<th>SD</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kindle</td>
<td>0.6</td>
<td>0.52</td>
<td>1</td>
</tr>
<tr>
<td>iRiver</td>
<td>0.6</td>
<td>0.52</td>
<td>1</td>
</tr>
<tr>
<td>Sony</td>
<td>0.5</td>
<td>0.53</td>
<td>3</td>
</tr>
<tr>
<td>jetBook</td>
<td>0.5</td>
<td>0.53</td>
<td>3</td>
</tr>
<tr>
<td>Pandigital</td>
<td>0.5</td>
<td>0.53</td>
<td>3</td>
</tr>
<tr>
<td>Elonex</td>
<td>0.4</td>
<td>0.52</td>
<td>6</td>
</tr>
<tr>
<td>Kobo</td>
<td>0.3</td>
<td>0.46</td>
<td>7</td>
</tr>
<tr>
<td>Nook</td>
<td>0.0</td>
<td>0.00</td>
<td>=8</td>
</tr>
<tr>
<td>EZReader</td>
<td>0.0</td>
<td>0.00</td>
<td>=8</td>
</tr>
</tbody>
</table>

Table 10. Mean Ranking and Standard Deviation (SD) for e-reader preference: significant

4.7 Pairwise comparison on e-readers with significant differences

Pairwise comparison is commonly used to estimate preference values of finite alternatives with respect to a given criterion. This is part of the model structure of the analytical hierarchy process, a widely used multi-criteria decision-making methodology. The main difficulty is to reconcile the
inevitable inconsistency of the pairwise comparison matrix elicited from the decision makers in real-world applications (Choo & William, 2004).

The elimination of designs or candidates can change the tabulated rankings of those designs or candidates that remain under consideration. The determination of which design is “best” or which candidate is “preferred most” may well be sensitive to the set of designs considered (Dym, Wood, & Scott, 2002).

The overall ranking we used based on the qualitative and quantitative mean rankings of the questions about e-readers that indicated a significant difference as shown in 4.3 opening a document, 4.4 increasing the font size, 4.5 navigating to specific section within the document and 4.8 e-reader preference. Table 11 shows the overall ranking of e-readers where significant difference was noted.

<table>
<thead>
<tr>
<th>e-Reader</th>
<th>Qualitative ranking</th>
<th>Quantitative ranking</th>
<th>Overall ranking</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kindle</td>
<td>=1</td>
<td>3</td>
<td>=1</td>
</tr>
<tr>
<td>Pandigital</td>
<td>=3</td>
<td>1</td>
<td>=1</td>
</tr>
<tr>
<td>iRiver</td>
<td>=1</td>
<td>4</td>
<td>=3</td>
</tr>
<tr>
<td>jetBook</td>
<td>=3</td>
<td>2</td>
<td>=3</td>
</tr>
<tr>
<td>Sony</td>
<td>=3</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Elonex</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>Kobo</td>
<td>7</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Nook</td>
<td>=8</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>EZReader</td>
<td>=8</td>
<td>9</td>
<td>9</td>
</tr>
</tbody>
</table>

Table 11. Qualitative, quantitative and overall ranking

The qualitative rankings were based on the number of yes responses to the question on whether one of the two devices was better. The quantitative rankings were based on Likert-like numerical scale rating on the individual tasks and participants’ responses in questions where a statistically significant difference was found in the responses.

The qualitative and quantitative rankings are similar, within one rank for the majority of e-readers. This suggests that our rankings can be relied on, as the two forms of data are largely consistent. The exceptions are the Kindle (1st in qualitative ranking and 3rd in quantitative ranking); Pandigital (3rd in qualitative ranking and 1st in quantitative ranking); iRiver (1st in qualitative ranking and 4th in quantitative ranking) and Sony (3rd in qualitative ranking and 5th in quantitative ranking). These differences are discussed below.

5 Discussion

Interaction with e-readers should be simple, effective and a pleasant experience as it is emulating a paper book.

The power ON/OFF buttons on the Sony and the Kindle was not easily noticeable. On the Sony, three similar buttons of the same shape and size were located at the top edge of the device and the label was not distinct due to the reflective surface and the angle at which the device was commonly held. A comment on the Sony was “doesn’t stand out from the other buttons on the same area”. The power switch on the Kindle was positioned on the lower edge of the device and the button was not labelled. Some of the comments on the Kindle power button were “button hard to find at the bottom - unusual position”; “hard to find switch button”; and “hard to locate the button, button un-labeled”. The ANOVA results showed no significant differences in the e-readers in response to the question for turning the e-readers on.

Some devices such as Sony, Elonex, jetBook, iRiver, Pandigital and EZReader stored the recently read e-books in the current reading or recent history folder. This extremely useful feature allows the reader to go straight to their e-book without being distracted with cumbersome
navigation and searching. In the experiments there was only one of each e-reader, when an e-reader was available for the next pair to evaluate there was not sufficient time to refresh the reading history folder, as it generally requires a full factory reset. It was a compromise to leave the documents within the folder especially if alternate methods for finding the document were supported. The recent history was not an obvious feature to all participants. Users were instructed that if the e-reader ‘accidentally’ started in the target document, then they should back out of it and navigate to it.

Typing a search query in the Kindle was cumbersome and required too much effort to complete the full query as it involved navigating the keyboard via the 5-way controller. Prevention of keyboard slips and incorrect spelling was not supported and frustrating to the user. An autocomplete feature would be a suitable design solution to combat the complexity in typing and error management.

Buttons are designed to afford pressing to trigger a visible change on screen. The search button on the EZReader was unresponsive and disconcerting as the users were unsure how to continue. If a button is displayed it must be associated to an action otherwise it leads to user frustration and a waste of valuable real estate on the interface. Although the participants encountered some difficulty with some of the e-readers on finding the document, the ANOVA results show there was no significant difference in the e-readers.

Opening a document was found to be either intuitive on some e-readers or challenging on other e-readers. This is also reflected in the ANOVA results, which showed a significant difference in the e-readers to support opening a document. For example, tapping on the document title opened the document on the Kobo, Sony and Pandigital. The iRiver had two “enter” buttons, one situated above the keyboard and the other within the keyboard for easy selection. The center buttons on the multi-controllers of the Kindle, Elonex and the jetBook behaved as the “enter” key.

Comments on the EZReader for opening a document were “hard to find”, “tried the arrows keys first”, and “a bit slow”. The “enter” button on the Nook touch panel was not identifiable because it did not meet a user’s concept of a button and it was not labeled. A comment on the Nook for opening the document was “open button is not obvious, task a bit to find. Fact screen is not touch is not obvious”. The “enter” key could not be distinguished if it was a radio button or a decorative element.

The ANOVA results for increasing the font size showed a significant difference in the e-readers to accomplish this task. The Pandigital supported alternate methods to change the font size for example, within an opened document from previous task, zoom icons (both plus and minus) were available on the bottom right corner of the screen and selecting the menu button displayed a toolbar with additional access buttons such as “My Library; Dictionary; TOC; Bookmarks; Go To; Font size and Next” on the top of the screen.

The font size adjust button was visible on the multi-controller of the jetBook, the keyboard on the iRiver and a stepper on the Elonex. The option to change the font size in the Kindle was accessible via the “menu” button.

Comments on the Nook on changing the font size were “menu a bit tricky to scroll through” and “not intuitive to find”. The Kobo has a multi-touch screen which meant each area of the screen was reserved for a function, for example tapping in the middle or top right of the screen turned a page forward and tapping on the top left of the screen turned the page backwards. A tap to the bottom of the screen was essential to unfold the font size icon button and a comment was “hidden toolbar, obvious icon”.

In the EZReader a “font” option was displayed via the “option” button but this was ambiguous as only the font type could be changed and not the font size. Comments on changing the font size on the EZReader were “no idea where to navigate”; “had to select font first, that wasn't any good so we kept looking through other menus” and “can't find the setting”. Each e-reader had its own methodology to accomplish the task of changing the font size which accounts for the difference identified in the analysis of variance calculation.
In the Kobo the “table of contents’ options was available via the concealed toolbar, similar to accessing the font size option. Comments on the Elonex were “first tried to use ‘explore’ button and press keyword ‘model number’, no response came out. Easy to find when using ‘content’ button”. The ‘content’ button was accessible via the ‘menu’ button.

There were no visual cues for a search or table of contents feature in the Nook and user’s found it to confusing. Comments on navigation in the Nook were “couldn't figure out how to scroll the screen”, “not found” and kept pressing the wrong button (left button sent pages forward)”. Similar to the Nook there were no visual cues in the EZReader to gain access to the specific section within the document. Comments on the EZReader to navigate to the section within the document were “cannot find it”, “we cannot find it”, “needed 15% hint for go to page”, “it's impossible to find that paragraph” and “can't find the document”. These variations for navigating in the e-readers to a specific section within a document were also identified in the analysis variance results.

Navigating and opening a second document reinforced the learning experience from the previous task for finding and opening a document and no significant differences were found in the ANOVA results.

There was no significant difference for readability. Almost all the e-readers had e-Ink screen types. Pandigital had SiPiX and jetBook had a LCD screen type, but these did not affect the readability. SiPiX technology uses white particles suspended in black ink while e-Ink technology has both black and white particles suspended in a clear fluid. A comment on Sony was “screen is dull and not enough backlight for contrast” but it was not significantly different compared to the other e-readers.

A significant difference in the analysis of variance results were found when asked which e-reader was preferred from the pair evaluated. This was then used to conduct a comparison to determine which e-readers was overall significantly better, based on counting the number of times each device was considered better than the other device.

There were four e-readers for which their qualitative rankings were more than one-step different in the quantitative rankings. The Kindle, Sony and the iRiver were all higher by two to three steps and the Pandigital was two steps lower. This could be due to the Pandigital being an unknown brand compared to the Kindle and Sony as popular brands, though this does not explain the iRiver which showed the highest difference between the qualitative (1st) and quantitative (4th). It is possible that the brand recognition is the primary answer for the Kindle and Sony. For the iRiver, it appears that the overall interface is visually appealing to the users to account for the high qualitative rankings but the functional aspects and the aesthetics of the interface are not in complete harmony to meet the user’s expectation.

Comments on the iRiver and EZReader were that the keyboards were redundant; the buttons were too small and overall not very useful. A full list of participants’ comments is available on request. Here is a summary of the comments:

1) The e-readers were either non-intuitive or intuitive
2) Some e-readers were found to be user friendly and easy to use
3) Expectations were for faster and more responsive feedback
4) E-readers with un-labelled buttons and use of icons were confusing
5) The lack of contrast colour of buttons and interface was deceptive
6) Preference for a device with touch screen interface

A “grounded theory” approach was considered to interpret the qualitative data but the comments were found to be too short and insufficient to code and to form concepts of the code, for a meaningful analysis. Grounded theory is a systematic generation of theory from systematic research (Corbin & Strauss, 2008).

In this study, three e-readers, the Pandigital, the Sony and the Kobo used Touch technology. The Pandigital and the Sony are also button-operated interfaces, while the Kobo was touch only. The
qualitative rankings for these devices were Pandigital (3rd) and Sony (3rd) and quantitative rankings were Pandigital (1st) and Sony (5th). The Kobo is consistent with both its qualitative and quantitative ranking at 7th. It is clear from these tasks highlighted in 4.9 pairwise comparison, that button operated interfaces are consistently better, that is the three devices which had touch interfaces are distributed among the top, middle and bottom. Further, the only full touch e-reader (Kobo) is among the bottom three.

It is worth noting that most new devices available on the market in 2015 are mostly touch devices, and over 60% of the participants commented that they either preferred a device with touch screen interface, or would improve the device by adding a touch screen interface.

In experiment 2, the preliminary analysis of variance calculation showed a significant difference in the e-readers only for navigating and opening a second document. This could be attributed to the instruction on to not use the ‘date’ search. That is, mentioning ‘search’ could have predisposed the participants to use the search feature to find the documents. The second document could be found only by the title search. Content search was not possible because the second document was a very short document, with very little content to search on.

The search and table of contents features has not been investigated in previous studies, most likely because it is not a feature associated with paper books. However, on an electronic device a search feature is critical to finding information quickly. The search feature should be investigated in the future.

The overall ranking is based on the qualitative and quantitative mean rankings of the e-readers with a significant difference from the experiment 1 (Q3-Q5) and experiment 2 (Q6) to understand the difference between the sets.

6 Conclusion

The ANOVA results showed significant differences between the e-readers for the tasks of opening the document, increasing the font size, navigating to a specific section within the document, and user preference on the e-reader from the device pair evaluated.

In this study the tasks was measured on a Likert-like numerical scale rating the e-readers which was used to calculate the Mean, Standard Deviation and Analysis of Variance (ANOVA). From ANOVA results, a further comparison on the qualitative and quantitative rankings were used to determine which e-readers were overall better.

In the overall ranking the Kindle, Pandigital, iRiver and jetBook were in the top four positions. The Kindle, iRiver and jetBook are button-operated e-readers whereas the Pandigital is combination of buttons and touch. E-Readers with buttons are intuitive to use. Users know when they have selected a physical button whereas tapping a button on a touchscreen is not always instantaneous due to the speed of response and the user may continue to tap the screen.

This study concluded an analysis on nine e-readers. The next stage of this project is to extend the evaluation for touch based e-readers as they seem to be dominating the current market for e-readers, requiring more data collection and further analysis. Exploring the direction for further studies could involve:

- Updating the scenario to evaluate tasks for continuous reading
- Evaluating the e-readers by concealing the brand names
- Evaluating the usability of an e-reader to read one-handed whilst drinking, coffee, holding an umbrella, standing on a bus or train and so on
- Evaluating the search and table of contents features in e-readers
- Investigate different styles of documents (e.g. Rho & Gedeon, 2000) and readability
- Directly investigate reading behaviour with e-readers using eye gaze (e.g. Vo et al, 2010)
References

Gibbon, E. (1776-1789): The History of the Decline and Fall of the Roman Empire, volumes 1-6, Strahan & Cadell, London.
Plato (circa 350 BC): Euthyphro, publisher unknown.
Call for Papers for publication in 2016

The Australian Journal of Intelligent Information Processing Systems is now calling for articles for Volume 14 Issue 3 onwards, for publication in 2015.

Intelligent information processing systems is an exiting research field since information will be a key issue for future society. Research in Bio-inspired Computing, Computational Intelligence, Knowledge Engineering, and Ontological Studies are major technologies that will enhance the advances of Intelligent Information Processing Systems. The Australian Journal of Intelligent Information Processing Systems is an interdisciplinary forum for providing the latest research developments.

The theoretical and application areas of interest include, but are not limited to:

- Artificial Neural Networks
- Fuzzy Systems
- Evolutionary Computing
- Cluster Analysis
- Statistical Learning
- Ontological Studies
- Virtual Reality
- Information Retrieval
- Data Mining and Knowledge Discovery
- Human Centred Computing

The journal publishes both theoretical and application oriented papers in diverse area related to intelligent systems.

Professor Tom Gedeon  
Australian Journal of Intelligent Information Processing Systems  
Research School of Computer Science  
The Australian National University  
108 North Road, Acton 2601, Australia  
Phone: +612 612 51052  
Fax: +612 612 50010  
Email: ajiips@cs.anu.edu.au  
Web: http://cs.anu.edu.au/ajiips
Information for Authors

The Australian Journal of Intelligent Information Systems (AJIIPS) is published approximately quarterly. The aim of AJIIPS is to publish papers describing theory, methods, and techniques, applications, and tutorial presentations in a range of areas relating to computer engineering and computer science. AJIIPS publishes full papers, short notes, and survey articles. Specific areas include but are not limited to:

- Artificial Neural Networks
- Fuzzy Systems
- Evolutionary Computing
- Cluster Analysis
- Statistical Learning
- Ontological Studies
- Virtual Reality
- Information Retrieval
- Data Mining and Knowledge Discovery
- Agent Driven Systems

All the papers will be reviewed by at least two reviewers and one of the editors. Decisions as to the suitability of the papers will be made by the Editor in Chief.

Papers should be clearly presented, consistent with giving proper description of primary contribution. Theory papers should be based on clear, formal foundations; methods and techniques should indicate the novelty and advantage of the technique; applications papers should present appropriate results and evaluation of performance; tutorial papers should be clearly presented to a reader with a general background in areas of interest to the journal but no prior background in the specific topic.
Australian Journal of Intelligent Information Processing Systems

Subscription Types

<table>
<thead>
<tr>
<th>Subscription type</th>
<th>Format</th>
<th>Duration</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paper subscription</td>
<td>Print</td>
<td>4 issues</td>
<td>220.00 (AU$)</td>
</tr>
<tr>
<td>Online access to back issues</td>
<td>Online</td>
<td>1 year</td>
<td>220.00 (AU$)</td>
</tr>
</tbody>
</table>

Subscriptions Contact:

Assistant to the Editor in Chief
Australian Journal of Intelligent Information Processing Systems
Research School of Computer Science
The Australian National University
108 North Road, Acton 2601, Australia
Phone: +612 6125 1052 59664
Fax: +612 6125 0010
Email: ajiips@cs.anu.edu.au

Subscription Order Form

Australian Journal of Intelligent Information Processing Systems
ISSN 1321-2133

<table>
<thead>
<tr>
<th>Name</th>
<th>Address</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Format</th>
<th>Duration</th>
<th>From (issue/year): .......... To (issue/year): ..........</th>
</tr>
</thead>
</table>

I enclosed my cheque for AU$:AU$ : ......................................

....................

(Your signature)