Utrecht University

Graduate School of Natural Sciences

Faculty of Science

Department of Information and Computing Sciences

Programme: MSc Game and Media Technology

Master Thesis

"HiStory" Leveraging human perception for effective performance in video retrieval tasks on mobile devices

Dimitrios Paris Darzentas - 3611183

Supervisor: Dr. W.O. Hürst

2012

Table of Contents

Table of figures

Abstract

The objective of this MSc project was to explore the use of mobile platforms, such as touch screen equipped smart phones and tablets, for video retrieval applications and to produce and evaluate an innovative prototype.

The MSc project built upon the research undertaken for a preceding (experimentation) project. That work validated the viability of complex thumbnail based interfaces for mobile platforms. It provided a strong foundation for the technical parameters such as thumbnail size, quantity, and layout. In addition, interesting findings in the domain of human perceptual abilities and how they pertain to the field of video retrieval were also documented. Naturally, this work also led to a number of new research questions which formed the basis of the MSc thesis.

More specifically, the MSc thesis treated two interrelated tasks. The first was to perform a more extensive study and analysis on the effects of physical user interaction, in this case touch gestures, on the performance of video retrieval related tasks; a question that was directly derived from the findings of the experimentation project. The second was to utilise these findings, along with the previous findings in order to design, implement, test, and evaluate a thumbnail based video browser for mobile platforms. The aim of this browser was to take full advantage of the interaction characteristics of the platform and what we have learnt of the abilities of human perception. The browser, called "HiStory" (Hierarchical Storyboard), provided the users with a seamless and dynamic overview of the content of a video while giving them the option to control the granularity of the layout therefore allowing for effective Known-Item-Search task solving.

With regard to the evaluation of the browser, its effectiveness was measured by conducting a user study, which gathered quantitative data on accuracy and speed, and qualitative data in the form of live feedback and questionnaires. Also taken into account were the evaluation methods stipulated by the international academic community involved in the field of video retrieval and Human-Computer-Interaction principles.

The resulting data from the evaluation led to a number of interesting conclusions which support both the viability of scrolling grid interfaces and the effectiveness of alternative interfaces such as HiStory. Between meaningful quantitative results and very positive qualitative feedback, a justified avenue for further research has been created that will supported future endeavours in this most promising, and rewarding, field of study.

Acknowledgements

I would like to thank my supervisor Dr. Wolfgang Hürst for his guidance, support and infectious enthusiasm for research.

I would also like to thank all the users who participated in the experiments. Their positive feedback and encouragement was instrumental in seeing this project through.

Finally I thank my family for their love and support.

1 Introduction

In the digital age, the rapidity of technological change and development is well-documented, and in recent years we are witnessing a phenomenon whose exponential growth has affected a multitude of disciplines. This phenomenon is digital content: its creation, distribution - and most importantly for this work - its consumption.

The combination of an increasingly technology literate population, affordable and accessible hardware and software and -a recognised key driver- the human need to create, share and socialise (Odlyzko, 2001) have led to the creation of massive amounts of digital information. This includes content of all types, especially in text-based media. However, in this MSc thesis the focus is on video based content, and ways to enable searching and retrieving information that is created in that medium.

In terms of video content, professionally produced items (such as films, shows, documentaries, etc.) have been steadily increasing in quantity even before the digital age due to their enormous popularity. Filmmaking has been available to the general public with amateur movie making equipment (cine-cameras) for more than half a century and has also been popular since that time, reflecting the limitlessness of human creativity. Yet, back in the days of celluloid film, video content had a physical existence, severely limiting its transferability and longevity, and consequently limiting its distribution and by extension the ability of film creators to share it.

The advent of digital media drastically changed this situation. The continuing decrease in the effort and cost for the creation of content with personal video cameras and then cameras integrated in to mobile devices was one aspect. But the final barrier between creators and audience was decisively dismantled with the advent of internet based digital video content distributions, such as YouTube[1], Vimeo [2] and others. Distribution has become not just easy, but practical and powerful.

Moreover, it is also a focus of activity with whole social trends revolving solely around distribution methods. Already in 2007, a study of user behaviours in video consumption on mobile devices found that it was far more than the "simplistic notion of viewing to kill time wherever you may be". Instead they noted behaviours with strong social implications, such as sharing and exchanging videos, and searching out particular known segments to view over and over and to distribute to others so they can view them too (O'Hara et al., 2007).

Finally, in a circular effect, the increase in easy channels of distribution leads to an even greater prevalence of amateur content, directly available to the viewers. As a result, this has led to the established concept of, and real need for, browsing in relation to video content. Thus floodgates have opened, the content is there, it is increasing by the minute and easily accessible, - or is it?

When there are large amounts of content of any kind, efficient ways to browse that content are required. From simple tables of contents for documents to elaborate library card indexes for collections of documents, starting with the *pinakes* of Kallimachus, systems have been created to accompany information for the sole purpose of enabling humans to find what they are looking for, rather than having to exhaustively search through all the content. This kind of metadata approach

[1] www.youtube.com
[2] www.vimeo.com

(indexes and catalogues) can work fairly well for text based content. However, the task is more complicated when non-textual content is considered. Even putting aside the difficult and diverse technical aspects, complex multimedia content like video cannot be easily categorised. Videos are much richer content than individual images, which are already hard to classify; they present a huge amount of raw data; and have very little prior structure (Hu et al., 2011). Existing systems rely heavily on the existence of metadata in a proprietary form, like YouTube tags. The volume of video content is so large, the rate of creation so fast and the channels of distribution too open and uncontrolled, so that this approach cannot guarantee effectiveness. As a measure, in June 2012, it was noted by officials at YouTube that roughly 72 hours of video are uploaded every minute[3].

In this thesis, sound is not considered, instead the work is restricted to the visual content of videos, following the distinction made by (Hu et al., 2011). This is also because, although microphones and recording equipment has been available and popular for at least as long as cine cameras and camcorders, there are not the same popular distribution channels for audio content, nor the same rate of user content creation.

At its most basic technical level, video is a series of images/frames, displayed at a regular rate. The way the human brain meshes these images together to give them cohesion and then proceeds to interpret them is what gives them meaning. What this effectively translates to is that any given video contains large amounts of information meant to be accessed in a specific linear way over a specific amount of time. This fixed nature is adequate for the main goal of video, namely conveying content, be it for entertainment (Movies, TV episodes, Sport broadcasts); information transmission (News broadcasts, tutorial videos); or other purposes (surveillance videos, etc.). After all, video is an (inexact) analogue for the human experience, - auditory and visual stimuli experienced in linear time.

Yet, it is this very nature that creates many problems when interacting with video content not as a passive viewer, but in other roles. That is, when the task at hand is not to ‘watch’ a video linearly, whether it was designed (directed) that way or not, but to explore or to ‘browse’ that video for a variety of reasons. There are countless different tasks that may need to be accomplished with video content: from finding a scene or segment; to detecting instances of physical objects (e.g. ‘Find the red car’); or to identifying abstract meanings and constructs (e.g. ‘Find the Deus Ex Machina scene’). Successfully accomplishing such interpretive tasks requires the ability to ‘scan’ and ‘process’ the entirety of the video quickly and efficiently. Linear and time-expressed by their very nature, videos are not tailored for such tasks. The issue is further compounded by the fact that the meaning and context of most of the information is discernible only to humans.

More specifically, the vast majority of useful information contained in a video only takes form due to human understanding. The context and nuances require human interpretation; therefore it is a very challenging endeavour to automate such tasks. There are many fruitful avenues of work on automatic video content analysis, however, for many issues, human judgement is still highly valued. For instance, a company [4] was launched whose main business was to ‘crowd source’ the monitoring of footage from surveillance cameras and report any suspicious or abnormal behaviour. The

[3] http://www.youtube.com/t/press_statistics?gl=GR
[4] http://interneteyes.co.uk

rationale for this company is evidence of the need to supplement automatic methods and make use of the abilities of humans for interpretations of events, objects, locations, and actions.

Of the systems and user interfaces that have evolved to service the needs of video retrieval, many are based predominantly on visual feedback intended for human rather than machine scrutiny (Del Fabro and Böszörmenyi, 2012). They rely on the perceptual ability of humans to recognise and understand features even in small or even distorted images (Oliva and Torralba, 2007; Torralba et al., 2008; Ahlstrom and Schoeffmann, 2012). These interfaces have one key element in common, the thumbnail image, as described in (Hürst and Darzentas, 2012) :

> “Thumbnail images have become ubiquitous in our daily digital life as representations of, for example, larger photos or video clips. The old adage; “a picture is worth a thousand words” is well and truly verified in this regard. Ever since the thumbnail first replaced the icon as a visual preview representation of a digital image (or video) file, it has become accepted as the de-facto standard practice (….) A simple small image can convey large amounts of information in an instant. As a fast and effective way of visually browsing (with no other assistance such as metadata) through image or video items, thumbnails are undisputedly one of the most effective methods hence their widespread adoption in all forms of digital galleries, both on desktop systems and on mobile handheld platforms.” (page 1)

Video retrieval systems often have relied on metadata: that is the existence of background (textual) information describing features of the videos. This means that the interface design follows text based information retrieval, offering results from video queries as a list of videos with metadata, and an accompanying single descriptive thumbnail (e.g. as in YouTube) (Christel, 2006).

However, accurate metadata is notoriously hard to achieve, even for text based information. Combinations of automatic annotating tools and human crafted metadata ‘tagging’ often co-exist to try to achieve the optimum. Thus a reliance on metadata is also one of the key weaknesses of such video retrieval systems. As mentioned previously, the vast amount of content available and the rapidity of creation, not to mention the complexity surrounding the actual organisation of the metadata due to fragmentation and system evolution, mean that it is often the case that the metadata information is absent or worse, false. Therefore there need to be other approaches developed.

Finally, besides the question of automatic and human involvement in retrieval systems, there is another major element that affects how this issue is approached and that is the platform chosen for the consumption of video content. The previously mentioned digital video ‘boom’ is due, in no small part, to the just as impressive increase in personal computing. But personal computing is also constantly changing, moving from the desktop environment to the mobile environment. It is a fact that more digital content is already being consumed on mobile platforms and specifically video content is increasing5.

This poses a number of issues because while these devices have very attractive characteristics such as portability, ease of use and intuitive interaction; these advantages are counterweighed by major disadvantages as well, such as very limited processing power and restricted screen sizes. Thus the

[5] http://venturebeat.com/2012/02/15/streaming-video-consumption-q4/

question arises as to whether these platforms are suitable candidates for thumbnail based video retrieval interfaces. Intuitively, the smallness of the screen and limited interaction tools (buttons, etc.) -what is known as the 'form factor'- would seem to argue that these devices are too small for comfortable human recognition of video and of thumbnails. This was the first issue to be addressed in previous works (Hürst et al., 2010, 2011; Hürst and Darzentas, 2012) and the findings of these research endeavours returned extremely positive results regarding use of video and thumbnails on small screens that were encouraging and provided a firm basis for this MSc project.

Thus the objective of this MSc project is to explore the use of mobile platforms, such as touch screen equipped smart phones and tablets, for video retrieval applications and produce and evaluate an innovative prototype alternative interface.

The MSc project builds upon the research undertaken for a preceding (experimentation) project. That work validated the viability of complex thumbnail based interfaces for mobile platforms. It provided the foundation for technical parameters such as thumbnail size, quantity, and layout. Also documented were interesting findings in the domain of human perceptual abilities and how they pertain to the field of video retrieval on mobile devices. The results of the experimentation project were presented in a conference paper (Hürst and Darzentas, 2012) (The paper is in Appendix 1). Finally, this work also led to a number of new research questions which form the basis of the MSc thesis.

More specifically the MSc thesis examines open questions posed by previous research such as the effects of different haptic based scrolling methods on grid interfaces and evaluates the effectiveness of alternate video browsing interfaces. These issues are addressed in two interrelated tasks. The first task is to perform a more extensive study and analysis on the effects of physical user interaction, in this case touch gestures, on the performance of video retrieval related tasks; this is an issue that is directly derived from the findings of the experimentation project. The second task builds upon the first: it utilises the results from the first task's findings in combination with the other previous work in order to design, implement and evaluate a thumbnail based video browser for mobile platforms. The aim of this browser was to take full advantage of the interaction characteristics of the platform and what we have learnt of the abilities of human perception. The browser, which is called "HiStory" (Hierarchical Storyboard), provides the users with a seamless (i.e. no scrolling) and dynamic overview of the content of a video while giving them the option to control the granularity of the layout. The working hypothesis was to understand whether this would enable effective retrieval task solving of the Known-Item-Search (KIS) type. Also considered were principles of human computer interaction regarding HiStory's usefulness, usability and acceptability.

The effectiveness of HiStory was measured by conducting a user study, which gathered quantitative data on accuracy and speed, and qualitative data in the form of live feedback and questionnaires. Also taken into account were the evaluation methods stipulated by the international academic community involved in the field of video retrieval. The final implementation was to be submitted to the Video Browser Showdown 2013 in order to gather expert feedback in addition to that gained by the user study[6].

[6] The call for papers and description of the showdown can be found in Appendix 1, along with the paper written for submission.

This thesis is organised as follows: in the next chapter (Chapter 2), the general background to video retrieval research is presented, in order to situate the work of this MSc thesis. Chapter 3 explains the directly relevant background to the thesis in terms of established hypothesis and open issues. Chapter 4 continues to move into the details of the work undertaken to develop the prototypes explaining the mostly technical approaches and challenges. In Chapter 5, the experimental design; the technical implementation; the set-up of the experiments and the carrying out of the experiments are described. Chapter 6 presents the results of the analysis of the data, and the interpretations of those results. Finally, in Chapter 7 ends with conclusions; a weighing up of the work carried out; and discussion regarding the directions future work might take.

2 Related and background work

The MSc thesis is situated in an area of work that is known as video retrieval. In this area, a number of streams of research can be distinguished and there are several very helpful surveys that map out the extent of the on-going work (Geetha and Narayanan, 2008; Hu et al., 2011). In addition, the work has been supported in a structured way by the TRECVID conference series (Smeaton et al., 2006). This is a community working on information retrieval in Video since 2001, under the aegis of the U.S. National Institute of Standards and Technology (NIST). The goal of the conference series is "to encourage research in information retrieval by providing a large test collection, uniform scoring procedures, and a forum for organisations interested in comparing their results"[7]. This helps researchers to align their research approaches and not to reinvent the wheel, and provides for a wider evaluation base (Christel, 2007).

2.1 Automated video content analysis

Briefly, in the area of video retrieval, there are a number of efforts which work on automated video content analysis (VCA). Of these, some rely on structural approaches, breaking down videos into units such as scenes, shots and frames as depicted in figure 2.1 below (Kanagavalli and Duraiswamy, 2012), while some seek the semantics (concept based approaches) (Snoek and Worring, 2009). There are of course combinations of these approaches and new approaches continue to be devised, for instance, distinguishing poses that people adopt in videos as markers of meaning (Jammalamadaka et al., 2012).

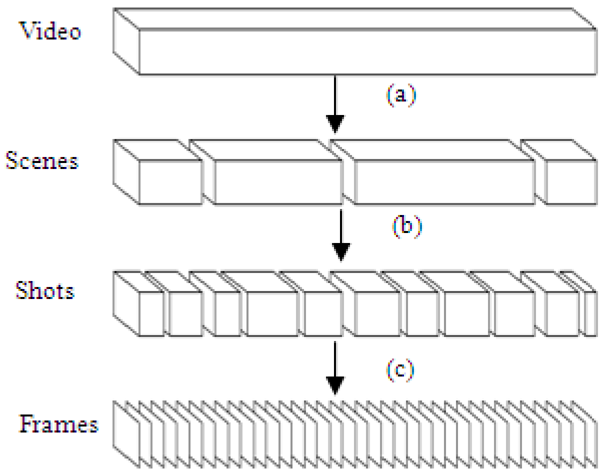


Figure 2.1 - (a) Video into scenes (b) scenes in to shots (c) shots in to frames from Kanagavalli & Duraiswamy, 2012

2.2 Video retrieval and human processing abilities – Video Browsers

Another strand of work relies on human processing abilities, (Schoeffmann et al., 2010; Del Fabro and Böszörmenyi, 2012). These focus on the design of the interaction to support the perceptual abilities of users, while giving them tools and interfaces that are intuitive and do not pose an extra cognitive load on a process (information seeking) which is already cognitively demanding for the users. They also experiment with visualisations of the data and of the interface components. These approaches, termed 'video browsers' sometimes eschew content analysis, that requires processing time, and where it is understood that the best algorithms are unreliable and computationally heavy

[7] http://trecvid.nist.gov/te

(Ahlstrom and Schoeffmann, 2012). This has been an important consideration for work carried out in the Known Item Search (KIS) tasks in the TRECVID work

In their review paper, (Schoeffmann et al., 2010), having reviewed more than 40 different interfaces classify them into 3 main categories.

The first category contains those that build upon traditional video-player like interfaces, that are well known and understood by users (seek bars, left to right progression, etc.) Several approaches have been used to make the scrolling more sensitive and to add more control for the user e.g. Elastic Scroll Bars, Variable Step, etc.

The second category, termed in the review 'video retrieval applications' are those where the users have specific queries. That is, they are searching, rather than browsing (or exploring) through video content. For this, users need more support for their searches. This means that often the video is, according to some criteria, displayed as a grid of thumbnails, or a storyboard. The storyboard is intuitive, and a number of applications use it in combination with other querying mechanisms (some based on content analysis) to help the users find what they are looking for. In user tests, (Christel et al., 2008), note that the storyboard can be very good for finding precise items, but is not so useful for more exploratory type interactions.

In the third category are those video browsers based on 'video surrogates and unconventional visualisation'. They use, for instance, fish eye layouts and video trees as depicted below in Figure 2.2.

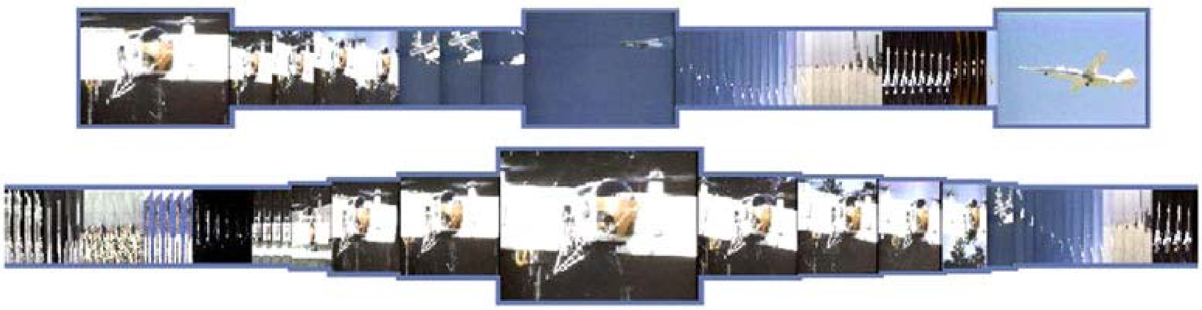


Figure 2.2 - The squeeze and fish-eye layouts for improved fast-forward and rewind. From (Divakaran et al., 2005)

The use of surrogates can be very efficient, for example, user studies showed that users perform better in search tasks with VideoTrees (Jansen et al., 2008). However, users declared they preferred the more classical storyboard approach, and the researchers have understood that that the surrogates may need to be differently designed each time to reflect the idiosyncrasies of different types of video content.

2.3 Combined approaches

There are of course many examples of combinations of automated content analysis and video retrieval with human processing abilities approaches, as well as of structural and semantic approaches, attempting to leverage these for better overall results. As an example of the combined approach (Pongnumkul et al., 2010) present an innovation on the traditional timeline slider control. Recognising the problem that the slider's effectiveness and precision degrade as a video's length grows, they propose a 'content-aware dynamic timeline control' that 'understands' the salient parts of the video via an algorithm that detects these, and then offers users a good rate of control when skimming over salient parts. In this way, they alleviate the users from needing to scan all the frames of video, but help them scan those parts with scenes of importance to users, and leverage the human abilities where they are best needed.

Against this background, the work of this MSc thesis is situated in the video browsing area. It is based upon exploiting human perceptual abilities and providing support for the interaction to be as intuitive as possible. It is also oriented towards mobile platforms, an area of work that will be next discussed.

2.4 Video content retrieval on mobile platforms

A newer element is that the work here is oriented towards mobile platforms. This is because of the increased viewing of video content on tablets and smart phones[8]. Already in 2003 researchers recognised the desire of people to consume content on their mobile devices (Fan et al., 2003), however these efforts concentrated on overcoming display constraints for single frames. Recently, the challenge has been reformulated by (Hürst et al., 2010, 2011; Hürst and Darzentas, 2012). This research exploits human perceptual abilities and expanded device capabilities, such as multi-touch screens and powerful hardware. To describe this work more fully it is useful to contrast it with the systems reviewed by (Schoeffmann et al., 2010).

These Video Browser systems reviewed by (Schoeffmann et al., 2010) were all designed as desktop interfaces and therefore had plenty of screen space with which to both display thumbnails and to have other windows with other functionalities, like search boxes, etc. Thus it is a concern that small screen sizes such as those on smart phones and tablets would severely limit the expansiveness, complexity and probably the usefulness of advanced video retrieval on these devices. Even the basic building block, the thumbnail is rendered smaller on a mobile device, and therefore may no longer be so effective.

Motivated by these questions (Hürst et al., 2010) carried out an evaluation where users were presented with different sizes of single solitary thumbnails in and asked to undertake common video retrieval tasks. The experiments also examined whether static or dynamic thumbnails were more effective. The conclusions showed that rather small images can still be reliably used for search tasks and if the thumbnails are dynamic even much smaller thumbnails still lead to a reasonable retrieval performance. In particular, thumbnail sizes as small as 80 pixels for static and 60 pixels for dynamic ones, respectively, have been determined as a viable conveyance of information on a mobile platform.

In a follow up experiment, (Hürst et al., 2011) identified that these thumbnail size results that were valid in the case of a solitary thumbnail representation, did not hold for designs in which single linear strips of five to nine thumbnails are shown. Users preferred larger thumbnails. The experiments also revealed that users were not taking advantage of the option to play individual thumbnails within the filmstrip as dynamic thumbnails – despite their superiority for search tasks. Reasons for this might be the higher cognitive processing loads resulting from the additional need to interact with moving as well static images; the increase in available information since more information about a video's

[8] http://newstex.com/2012/04/13/1-out-of-3-americans-will-watch-video-content-on-mobile-devices-by-2016/

content is shown with dynamic thumbnails; and the resulting ‘busy- ness’ and clutter of the interface.

Overall, the results of this work showed that it is worthwhile to design advanced interfaces for video retrieval on mobile platforms, but that more work is required to investigate open issues, in particular the effects that more complex layouts and arrangements would have on the users and their ability to complete certain video browsing tasks.

In the experimentation project reported on in (Hürst and Darzentas, 2012), the research question was phrased in the following way:

> “If optimum thumbnail sizes increase when we switch from classifying solitary thumbnails to film strip representations, will storyboard-like matrix representations result in a need for even larger thumbnails? And if so, how does the resulting need for interaction influence retrieval performance and subjective search experience?” (page 2)

Following (Christel, 2006), interaction was defined in this project as both retrieval performance and as how well the interface facilitates effective browsing. It was decided to restrict the user testing to static thumbnail representations, based on the assumptions of (Hürst and Darzentas, 2012) that simultaneous playback of various images as dynamic thumbnails would further distract users and most likely have an impact on performance and overall effectiveness when searching for information.

The major parameters explored were thumbnail size and the related ability of users to find content represented by them; and the role that interaction has on the effectiveness of the interface. As these are independent of the actual source of the data represented, the search tests were performed on single video files. The argument proposed was the expectation that the results could be generalised to search in multiple videos, where, for example, each row in the matrix represents the time-ordered sequence of thumbnails extracted from single videos.

The analysis from the experimentation project user tests showed that the static thumbnail grid layout achieved and maintained very high accuracy over all the different thumbnail sizes. This confirmed the results of previous work (Hürst et al., 2010, 2011), that the effectiveness of thumbnail based interfaces that was observed can be successfully transferred to more complex layouts. Another result was that an increase of optimum thumbnail size from 80 to 110px as reported by (Hürst et al., 2011) when moving from solitary thumbnails to film strip representations could not be observed when moving to the even more complex matrix-style storyboard layout – where thumbnail sizes of about 130px achieved an equally high performance rate of 90% as the much larger 200px size version of the interface. This was particularly encouraging, since the layouts evaluated in the experimentation project are far more applicable to real world scenarios.

These same levels of successful results were not similarly observed in the interaction tests where scrolling grid layouts were used. The analysis showed that uncontrolled, Continuous scrolling grid layouts are not a clear improvement over the static grid layout. It would appear that the added layer of interaction complexity for the user rendered the system less effective. This confirms the assumption about the critical influence of the interaction design on retrieval performance.

Thus the experimentation project recommended that further refinements must be made to maintain the high success rates of the static layout and at the same time, remove its inherent restrictions. Possible options might be interface designs where users can switch between various static grid arrangements (similarly to swiping through different screens on modern smartphone interface designs) in contrast to the Continuous scrolling evaluated in the experimentation project.

In the next section, the established hypotheses and open issues in mobile video retrieval are presented.

3 Established hypotheses and open questions

In the previous section a summary of the work and research that has been conducted in the area of video retrieval and especially mobile video retrieval was presented. Highlighted were the findings of the experimentation project (Hürst and Darzentas, 2012), which confirmed a number of hypotheses, such as the optimal size and quantity of thumbnails and the validity of complex arrangements. But it also led to a number of new questions that need to be addressed.

3.1 Grounded hypotheses

The first hypothesis that can be retained from previous research is the viability of mobile devices as effective platforms for video retrieval. Their value for video consumption is undisputed. The findings thus far show that while some users, particularly of older generations, do not feel immediately confident in using mobile devices with their limited screen sizes for complex video retrieval tasks such as known item search, their quantitatively measureable high levels of performance show they do perform well. Therefore the worth of pursuing further research into the area is scientifically sound, as well as contemporary.

The next important hypothesis is that thumbnails on mobile devices are still very effective at conveying large amounts of information and assisting in decision making. Despite instinctive reactions that the images would be too small, the results on isolated Images (Hürst et al., 2010) show that human perception is more than capable of interpreting the visual data into useful information. Indeed the effect is magnified when the thumbnails are of a dynamic nature, either in the form of slide shows or as short video clips. The reported sizes can be as small as 30 pixels, measuring just a few millimetres in width, however but this hypothesis has not yet been substantiated for complex layouts (Hürst and Darzentas, 2012).

Furthermore, the latest findings show that the ability to extract information from these small thumbnails hold even when the number of thumbnails is increased dramatically and they are arranged in complex and dense layouts. At this point the previously disparate images gain context and meaning as part of a video and the validity of thumbnails as vital elements of video retrieval interfaces is affirmed. Situations with thumbnails quantities as large as 80, per individual screen, combining dimensions as small as 8mm were tested and evaluated as viable and effective. The best results though were in more moderate setups combining 48 to 36 thumbnails with sizes approximately 10-12 millimetres. Quantitative performance measures for accuracy for these sizes averaged above 80%.

3.2 Open issues

These questions revolve around the user **interaction** with thumbnails interfaces, the type of the thumbnails, the tasks performed with these interfaces and finally the semantic context of the videos and the thumbnails.

The issue of interaction must first be more concretely defined as the term can apply to a large multitude of concepts in any given situation. For the purposes of this research, interaction refers to the physical actions undertaken by the user in order to manipulate the information that is presented on the screen of a mobile device. Therefore in such a case the physical interaction in question is predominantly of a haptic nature and includes such actions as tapping, scrolling and swiping. This clarification is necessary as the focus is specifically on these actions as they are directly pertinent to the user interface design that is utilised and the type of content that is consumed.

More specifically, in the experimentation project (Hürst and Darzentas, 2012) the users, apart from the simple interaction of tapping their selections, were given two types of interfaces to navigate, one where no scrolling was required and all available information was provided on a single screen and a second where there was more content available than could fit concurrently on the screen and therefore scrolling was necessary. The results showed that in the second case, the additional interaction negatively affected the performance of the users. However, these results alone could not condemn scrolling as a practice to be avoided in such scenarios. Further investigation was warranted in order to determine whether the action of scrolling is indeed as detrimental to performance as initially shown and whether it could be be further refined for increased effectiveness.

Issues were also raised about the **type of tasks** given to the users to complete. Initially the tasks were modelled after the Known-Item-Search tasks as defined by TRECVID (Smeaton et al., 2006) which serves as the standard for benchmarking video retrieval interfaces. These issues revolved around the ambiguity of the text that provides the 'known item' description that the users are searching for. Such issues have been raised before (Cao et al., 2010), and care needs to be taken in order to eliminate personal interpretations or ambiguity that may be implicated in low task performance.

A further avenue of research that could also be of great interest is the use of **dynamic thumbnails**. These are thumbnails that are not simply single static images but are either a set of images shown as a slideshow or are short segments from the video played back. These thumbnails are by their nature only applicable to video content. Previous works examining dynamic thumbnails, either isolated (Hürst et al., 2010) or within small groups of static thumbnails (Hürst et al., 2011), showed that they increased the amount of information available to the user exponentially and greatly aided performance and decision making. These though were all cases where there was only one dynamic thumbnail that was 'animating' at any given time. Their effectiveness in complex layouts has not been examined and neither has the case where more than one dynamic thumbnail is 'animating' concurrently.

Finally, a very important aspect of thumbnail based video retrieval interfaces has been left unexplored by this research and that is to analyse and utilise the **human-centric semantic context** of the thumbnails. Currently, any thumbnails selection is purely temporally based and the actual meaning that could be derived from the content is ignored. This is not a decision made lightly as the domain of analysing video based on its content is a field in its self, and one that is completely open to interpretation. Methodologies range from relatively 'simple' tactics such as scene segmentation - assuming one has a concrete definition of what a 'scene' is - to for instance, direct manipulation approaches such as that proposed by (Dragicevic et al., 2008). Following the instrumental interaction approach of (Beaudouin-Lafon, 2000), the direct manipulation approach proposed a method for browsing videos by directly dragging their content. For example, if users are interested in finding out about the action of a pool ball, they can examine this by dragging the ball to the point where it hits another ball. This work relies again on the human to provide the criteria to search over, and the computer to provide the processing for the points of interest.

From these open issues, the MSc thesis focuses on trying to answer questions regarding the role of interaction in the effectiveness of the interfaces. The goal was to measure the effects of physical user interaction, and in particular touch gestures, on the performance of video retrieval tasks on mobile platforms. In the next chapter the approaches and challenges for these tasks are presented.

4 Approaches and challenges

A general issue when considering video retrieval interfaces on mobile platforms, especially those purely based on visual cognition, such as thumbnail based interfaces, is the simple fact that the system has to provide the user with large amounts of information that cannot all be displayed onscreen at the same time. This issue is derived from the combination of the physical characteristics of the mobile platforms, i.e. small screens, and of the inherent characteristics of video content and the visual interface designs used to analyse them. Therefore the most pressing issue in this avenue of research is to design, develop and evaluate methods that approach and deal with this issue, whether by optimising time tested techniques such as scrolling or moving in different directions such as hierarchical patterns and advanced visualisations.

4.1 Examining haptic interaction – Scrolling Methods

As mentioned above, further research was warranted into the specifics of physical user interaction with video retrieval interfaces on mobile devices. The previous research experiments that highlighted the issue utilised a straightforward, 'default', implementation of a scrolling interface. The grid of thumbnails that was presented to the users contained more items than could fit on the screen at any one time and therefore in order to view the entire content the user could freely scroll either upwards or downwards, as shown in Figure 4.1 below.

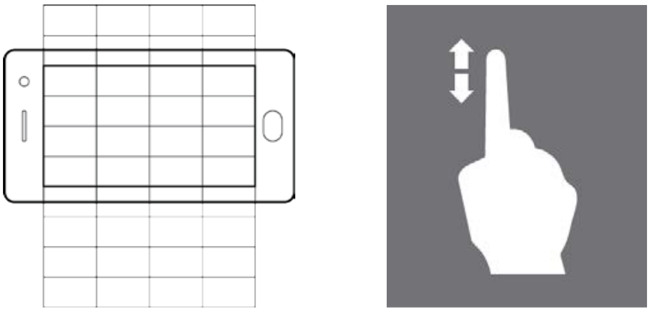


Figure 4.1 – Scrolling grid on a mobile device.

The physical interaction required for this was in the form of a haptic gesture that has become quite familiar in recent years due to the widespread existence of touch-screen devices. Users could simply hold and drag the onscreen content in the direction they desired. Technically this gesture is referred to as 'dragging' and is identical to physically moving an object from a fixed view point, intrusively reading down a sheet of paper by moving the paper instead of moving one's eyes or head.

This scrolling interaction is one of the most natural actions a user can make in an area which is by definition unnatural, but nevertheless the results showed that its inclusion had a significant impact on performance. Yet the basic problem of having to provide the user with more information than can fit on the limited screen of a mobile device remains and therefore some form of scrolling is necessitated

The first step in examining scrolling was to revisit the scrolling system that was used in the experimentation project and ensure the technical implementation would not affect performance. The next step was to investigate and test alternative methods of scrolling. Users of contemporary touch-screen equipped mobile devices will be familiar with the two main methods for scrolling through content on such devices. The first is the aforementioned method, termed 'Continuous scrolling’ where the user has complete control over his position in within the content. The other is ‘Paged’ or ‘Discrete’ scrolling. In this method the content is segmented into discrete pages and each scrolling action, in this case a swipe gesture, moves though the content exactly one page or step. In this type of scrolling, the user sacrifices absolute control over positioning for unambiguity and simpler controls.

In a cognitive approach, in a theory similar to that posed by research in keyhole navigation (Mehra et al., 2006; Hürst and Bilyalov, 2010), the Paged scrolling approach takes full advantage of the human brain’s capacity for spatial thought and spatial memory as the position of each thumbnail on each screen is retained and remains the same when the user scrolls back to a screen previously viewed. This consistency can subconsciously assist the brain in creating spatial associations, and improve recognition speed.

Both scrolling methods have overt and subtle differences and apparent advantages and disadvantages. To properly determine whether one of them is more suitable over the other for the purposes of video retrieval interfaces, an experiment is needed which directly contrasts the methods against one another using quantitative and qualitative measures.

4.2 Examining alternative methods - HiStory

Investigating the merits of conventional approaches such as scrolling was one objective, but the fluid and modular nature of digital platforms allows for exploring other, alternative, options, with different interface and interaction paradigms.

The experimentation project and the research it was based on provide a grounded starting point and guidelines for the conceptualisation, implementation and evaluation of a video browsing interface. This was achieved by utilising thumbnails as the basic building block and achieving appropriate functionality on mobile devices, taking advantage of their unique capabilities. The work was further guided by other research conducted on the human interpretation of thumbnails (Ahlstrom and Schoeffmann, 2012) and the standards and evaluation methods dictated by the TRECVID video retrieval research community.

So far the designs evaluated have been based on a breadth search paradigm. That is users are presented with a large number of thumbnails covering the length of a video with a fixed time interval between the sampled thumbnails. This expansive design essentially operated at a fixed ‘depth’ of the video. The thumbnails represented the whole video but at a specific granularity, conceptually, this is similar to having a single printed map of an area. Depending on the scale of the map there is a fixed amount of information and detail, which may be correct for a specific task but otherwise too little or too much for others. Using this analogy of a map, and wanting to avoid the restriction of a fixed scale, the proposed interface follows a variable granularity approach, emphasizing depth over breadth. What this means is that it can display less thumbnails at any given time, but allows control over the displayed content. The idea can be compared to contemporary map navigation methods. Dynamic Vector maps, such as those featured in GPS navigation systems and Google Maps, allow users to freely pan and zoom onto the area of their interest, optimally framing it at the required scale and therefore level of detail.

The interface concept, entitled Hierarchical Storyboard (HiStory), consists of two elements, a storyboard-style grid of thumbnails, taking up most of the available screen space and a narrow vertical bar on the right side of the screen. The grid serves as the primary interaction and visualisation point and the bar acts as a non-interactive visual aid that indicates which part of the video is visible at any time. The grid features a familiar and effective storyboard design with temporally ordered thumbnails that represent still images extracted from the video. Depending on the size and resolution of the mobile device used, the dimensions of the grid and the thumbnails vary.

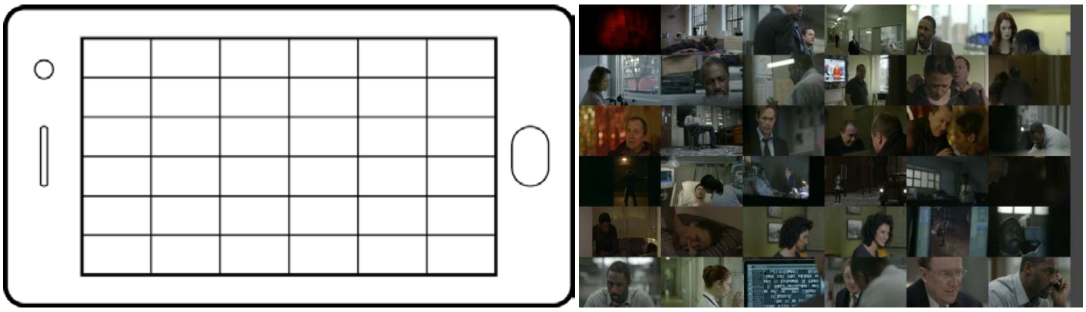


Figure 4.2 – Example of a segmented video

Figure 4.2 above, illustrates an example from a common mobile phone where the video has been split into 36 thumbnails arranged in a traditional storyboard layout. The indicator bar on the side functions as a reference, representing the currently viewable portion of the video. That is, when at full length, spanning the height of the screen, it indicates that the entire length of the video is currently arranged on the grid. When the viewed portion changes, the length and vertical position of the bar also changes, in order to provide a useful and fast positional reference. Figure 4.3 depicts the HiStory method on a conceptual level. A user selects a thumbnail and a new grid is created around that thumbnail with a different granularity. The indicator bar on the right side shrinks to properly indicate the hierarchical level and portion of the video that is being viewed.

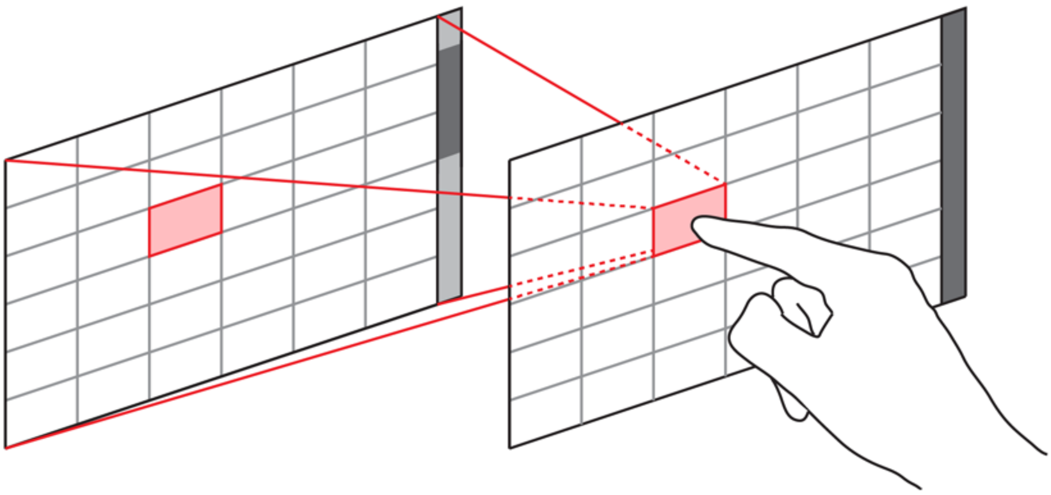


Figure 4.3 - HiStory illustrated

.

The core mechanic of HiStory is the ability of the user to dynamically change the granularity of the grid in a process similar to 'zooming'. The user specifies a thumbnail and the grid is rebuilt in a lower hierarchy with a smaller time interval around this 'anchor' thumbnail. This means that the viewed range shrinks (illustrated by the indicator bar) but the time interval between each thumbnail also becomes smaller, leading to more detailed information (finer granularity). Thus the previously disassociated thumbnails gain context as scenes as shown is Figure 4.5. Intuitively, the technique is similar to changing the scale of a map, as mentioned previously, affording more detail of a specific area while eschewing the general overview. The user can 'zoom in' multiple times, until the time interval between the thumbnails shrinks enough for a frame-by-frame representation. Inversely, they can also 'zoom out', backtracking through previous choices all the way to the top level or choosing a new thumbnail to use as an anchor point to go down a level. Parameters such as the number and size of the thumbnails are based on previous research in the area as is the question of cropping or distorting the aspect ratio of the images (Ahlstrom and Schoeffmann, 2012).

Figure 4.4 below illustrates the History approach in action. The image on the left displays the entirety of the video. The image on the right shows only a small segment of the video, evidenced by the indicator bar. In this case the granularity is much finer.

Figure 4.4 - HiStory - Highlighted thumbnail is the anchor, with the grid changing around it.

The next section details the mostly technical challenges that had to be overcome to implement the working prototype so that it is was robust enough for user testing.

4.3 Technical Challenges

In order to create the working prototypes of both the scrolling (Continuous and Paged) grid and the HiStory prototypes, a number of challenges to do with the processing power, the formats used in videos etc. had to be overcome.

4.3.1 Thumbnail Extraction

The first of these challenges was that of thumbnail extraction and in particular of dynamically extracted thumbnails. Thumbnail extraction was needed for the scrolling grids, and dynamically extracted thumbnails were needed for the HiStory prototype. Thumbnail extraction is a deceivingly complex task not only for technical reasons but also for interpretation reasons. How a representative thumbnail is chosen for a video is a field of research on its own and the issue is further compounded where there are multiple thumbnails to be extracted. In such a case one must also account for the semantic meaning linking each thumbnail. For the purposes of this research these considerations are set aside and the thumbnails sets extracted from each video are chosen purely based on a timed interval i.e. 'every n seconds'.

On a reasonably powerful desktop or laptop personal computer the process of extracting thumbnails from a video file is a relatively fast and straightforward task. Much effort has gone into optimising this process by taking advantage of Graphics Processors and Multi-Core CPUs in order to decode and manipulate video content as fast as possible. This is most evident in production environments and setups for video processing where effectiveness hinges on fast response. Having said that, these tasks are still extremely intensive and the fact that High-Definition content is becoming ever more prevalent and is itself becoming more complex (2K and 4K) means that there is a constant race between hardware and optimisation abilities and the 'weight' of the data to be processed.

4.3.2 Formats

As seen in the previous section, one of the key problems when considering any video retrieval system on a mobile platform, is the fact that there is substantially less processing power available for the necessary intensive tasks. The issue is further compounded when the desired system is based upon thumbnails. The complexity of dealing with digital video content is expanded to monstrous proportions due to the innumerable formats in use and circulation. There is a constant 'war' between proprietary and open digital video format standards all vying for dominance as the de facto industry standard. With the future and trends of digital video set to take and hold the centre stage of the digital world for many years to come, it is without exaggeration that one could describe the situation as similar to the 'format wars' of old such as VHS versus BetaMax and Bluray versus HD DVD but this time gone 'nuclear'.

What this means for developers, and finally for the end users, is that many efforts to 'break the mould' and be innovative in how video content is handled, are hampered by platform and standard fragmentation. It is very difficult to design and implement a consistent and reliable user experience across the multitude of platforms and content formats. In the course of this project, the aforementioned problem showed itself in the form of our chosen platform's inability to handle video in any meaningful way beyond simple playback, and then only in a couple specific formats (.mp4 and .3gp).

Additionally, the problem is further exacerbated by a number of factors, hinging on the platform of choice. Mobile hardware, despite its rapid advancement, is still far behind dedicated and desktop systems both in performance and in support. That is, mobile phones and tablet devices lack both the physical processing abilities and the necessary software support in order to manipulate video content.

Designers and manufacturers of such devices - or rather of the operating systems of such devices - consistently give video support lower priority over other aspects, such as 3D graphics support and rely on inefficient and proprietary standards. Put bluntly, extracting a given number of thumbnails from a video file at a set time interval is currently impractical on a mobile device. This statement is not made lightly as great efforts were made to overcome this obstacle. The next paragraphs detail the issues.

The first issue is that of codec support. The currently predominant operating systems for mobile platforms, Google's Android and Apple's iOS have only limited support for video formats. By default both operation systems only support a very small subset of formats including MPEG-4, H.264 and H.263. As is the case with web technologies, emphasis has been given to the H.264 standard as it has proven to be both powerful and flexible. Web browsers such as Chrome and Firefox, and Web based content providers such as YouTube utilise H.264 as their format of choice. A choice which is not unjustified as the standard since it is, as mentioned above, very powerful but also it is becoming increasingly common. It is also the codec used for contemporary physical media such as Blu-Ray.

Having said this though, a vast amount of information is encoded with different codecs, such as MPEG-2 or WMV, for which the mobile platforms in question have no inherent support whatsoever. For end users, playback support for these formats is easily achieved with third-party software (or apps as is the term nowadays) but actually incorporating the ability to decode other video codecs for other uses requires substantial commitment.

4.3.2.1 ADDENDUM

Shortly before the submission of this document, a press release [9][10]by the Moving Picture Experts Group (MPEG) presented the draft for a new international video compression standard, informally titled h.265. This standard can provide high efficiency compression with accompanying high quality and the preliminary plan is for it to be fully supported by 2013, especially on mobile platforms. Such an event would greatly assist in future research endeavours in the area of mobile video retrieval and should dramatically increase the technical feasibility, viability and effectiveness of complex interfaces.

4.3.3 Processing power

For the purposes of this project, the issue with the media support of mobile platforms does not directly lie with their ability to perform playback at an acceptable rate but more with the ability to extract certain frames in a timely manner, a task which is, un-intuitively, much harder. Extracting the the currently displayed frame from a playing video is a trivial task, a simple matter of recording and saving the image buffer, but extracting large amounts of frames from different points in a video is a more complex task.

Technically, any video manipulating application has a seeker, i.e. a reference point within the length of the video. Moving though the video, whether it is actually displayed or not, is called 'seeking'. One cannot arbitrarily request a specific frame of a video to be extracted, e.g. 'extract the nth frame', but instead the seeker must seek to that position, decode the video stream and then extract the frame. This process must then be repeated for each individual frame requested, the more frames requested, the more time is needed to complete the task. The effectiveness and speed by which this process is completed is heavily dependent on the software implementation of the used codec. If the implementation does not take advantage of all the possible optimisations then there are significant penalties in performance. Advances in multicore processing and graphics processor utilisation on personal computer systems have led to impressive optimisations in software dedicated to handling and manipulating video content, such as Adobe Premier[11], Avid[12], etc. In such software packages, whose effectiveness directly lies with their ability to rapidly display and edit digital video content at

[9] http://www.engadget.com/2012/08/15/mpeg-drafts-twice-as-efficient-h-265-video-standard/
[10] http://www.ericsson.com/news/120814_mpeg_244159018_c
[11] http://www.adobe.com/products/premiere.html
[12] http://www.avid.com

the frame level, real time extraction and display of thumbnails is achieved, albeit with a pre-processing cost and a significant reliance on hardware.

However, this is not an option for a mobile platform. The available hardware, while improving at an impressive rate, is nowhere near as powerful as that of a dedicated desktop solution. More importantly, as mentioned above, software support is still in primitive stages, and not optimised for the tasks needed. More specifically, the approaches outlined earlier require real time extraction of thumbnails with an acceptable quality. This process is not an issue for small quantities of thumbnails, especially on personal computer systems which are optimised for the task, but for the numbers needed for the scrolling grid and HiStory interfaces, the task becomes critical. In order to proceed with the project, efficient solutions needed to be found.

4.3.4 Solutions to Codec issue

A number of approaches were evaluated and tested involving both hardware and software. The first approach was to improve the inherent support of the platform for video codecs. The most logical and efficient course of action was to examine the most prevalent desktop solutions and investigate their application to a mobile operating environment. One of the most popular and mature frameworks for handling multimedia is the open source LGPL licenced libavcodec which is part of the FFmpeg [13]project. The chosen Android development environment is Java based but includes support for a native C++ software development kit which allows for the integration and use of libraries developed in C++, albeit with significant difficulties. A series of attempts were made to compile and utilise FFmpeg, this framework utilises libavcodec and is one of the most widely used multimedia framework distributions and incorporated into widely used applications such as VLC[14], MENcoder [15]and Handbrake[16].

At the inception of this project there was no provision for such an endeavour, existing examples of porting multimedia libraries to Android were tenuous at best, sacrificing both reliability and features for proofs of concept. At the time of writing the situation has been significantly improved. With the growing demand for mobile video consumption, serious attempts are being made by third party application developers to provide video player solutions with the widest possible format support. It is worth noting that despite the apparent success and popularity of these applications, their performance is still sketchy with frequent problems and incompatibilities reported. The majority of these issues are not related to the design and capabilities of the mobile devices but are more due to the fragmentation of the platform situation due to the extreme variations in hardware and software. Additionally all these solutions focus on video playback and not video manipulation, which is what this project requires. Indeed even the released video editing applications for both Android and Apple platforms are applicable only to a very small subset of formats, notably those produced by the device itself.

Nevertheless, the attempts to incorporate proven multimedia frameworks into a mobile application were successful, both in the case of ground-up implementation and in the case of utilising third-

[13] http://ffmpeg.org/
[14] http://www.videolan.org/vlc/index.html
[15] http://www.mplayerhq.hu/design7/info.html
[16] http://handbrake.fr/

party compiled libraries like FFmpeg4Android[17]. At this point the next and crucial issue came into prominence, namely the comparatively underpowered mobile hardware. An FFmpeg command executed on a relatively high-end mobile device required at worst, seven times and at best three times, more time to complete than on an average desktop system.

Therefore once the problem of attaining the ability to decode a video stream was achieved, the main issue became solving the processing issue. The combination of weak hardware and inefficient software could not be easily overcome within the time and budget constraints of the project. It is true to say that advances in both aspects are being made daily with multiple core devices readily available and optimised software taking advantage of this are becoming ever more common, therefore in the long-term these barriers will be overcome. But for the purposes of this project an immediate solution had to be found.

4.3.5 Solutions to low processing power

The next logical step was to consider approaches where the processing intensive task would be offloaded onto dedicated hardware, delegating the mobile hardware purely to a display and interaction role. Essentially, a number of client-server arrangements were tested and evaluated, with the best contender being chosen as the technical basis for the experimental implementation.

Two main directions where investigated, the first was to setup a server running an optimised and fully featured compilation of FFmpeg and implement a server-side application, written in PHP that would handle the parsing of the necessary FFmpeg commands as requested by the client application on the mobile device. This approach mimics the established methods used by digital content providers in production situations.

The second direction was to leverage the power and flexibility of HTML 5, a contemporary version of the established web technology, which features extensive support for multimedia content and its manipulation. The side benefits of utilising such a technology were numerous, including portability and extensibility as the HTML 5 standard is already widely supported among many devices and is gaining momentum in the informatics world at a constantly increasing rate. Indeed, the first feasibility tests showed an impressive adaptability and an extremely simple implementation.

Having both possible setups a short feasibility experiment was staged that tested the reliability and flexibility of both on a variety of devices; different Android versions; and with a wide selection of content. The HTML 5 based solution while very easy to port to different devices was unfortunately not able to provide a consistent experience across different media types and was a sufficiently large departure from established mobile application design to prove problematic. The traditional server-based service solution, while inflexible and requiring considerable refactoring for flexibility, was able to provide the required results consistently, efficiently and with significantly increased speed once optimised.

With these results in mind, the server-client model was adopted as the method of choice for supporting the experiment. From this point on the implementation of the experiment became a feasible and realistic prospect.

[17] http://ffmpeg4android.netcompss.com/

Following this concept and having overcome the technical challenges described in the above section, a working prototype was implemented in order to test and evaluate the feasibility and viability of the design. In the next chapter the rationale; implementation; the functionality of the prototype and the execution of the experiments are presented and analysed.

5 The experiments

In order to answer the scrolling interaction questions posed in Chapter 3 (Continuous vs. Paged scrolling) and to scientifically validate and evaluate the HiStory concept, two experiments were designed, implemented and carried out. These experiments were administered together, sequentially, to a pool of 26 test subjects with a range backgrounds, ages, technological skills and attitudes to technology. The resulting data and findings are documented and analysed in the next chapter. In this chapter we describe the implementation of the prototype and its functionality as well as the user tests.

5.1 Source content and segmentation

The videos used for the experiment were sourced from popular television series, this was a decision that had proved popular in the experimentation project and had served to maintain the test subjects' interest throughout the experiment. It also serves to trial the designs for real-world situations as such content is one of the most prevalent video content types in circulation and under scrutiny.

The videos were all 30 minutes in length, in colour and with similar aspect ratios. For the purposes of the first experiment, thumbnails were extracted every 10 seconds, resulting in 180 thumbnails per video, in the following sections the utilisation of these thumbnails will be explained in detail.

5.2 Hardware setup

The hardware setup for the experiment consists of two elements, the client mobile phone device upon which the display and interaction takes place on the user end, and a personal computer, in this case a laptop running a custom server package in order to perform the processing-intensive tasks.

The mobile device in question is a Samsung Galaxy S i9000 model, this device is a good representation of the current abilities of mainstream devices. It does not feature cutting edge advantages such as High definition resolutions and multi-core processors which are fast becoming standard features of devices on the high-end spectrum at the time of writing. This is an advantage as its abilities and form factor are much more in line with the majority of hardware used by the population. It features a 4.0 inch AMOLED screen with and aspect ratio of 15:9 and a resolution of 800 by 480 pixels. This is currently one of the most popular resolutions on mobile devices and affords the display a respectable 233 pixels per inch density. Figure 5.1 shows the device in question being used during the experiment.

Figure 5.1 – Moble device used in experiments.

The CPU is a single core ARM Cortex A8 which implements the ARM v7 instruction architecture and the GPU is a PowerVR SGX540. In comparison with one of the top end devices at the time of writing, the Samsung Galaxy S3 i9300 has a Quad-core 1.4 GHz Cortex-A9 CPU and a Mali-400MP GPU. Although the CPUs maintain the same ARM v7 architecture, they still cannot be directly compared as the addition of multiple cores means that meaningful improvements are a matter of software optimisations. This is also the case with the graphics processors and it is of importance to note that up until version 4.0 of the Android Operating system (codenamed 'Ice Cream Sandwich'), Android devices had no provision or support for hardware accelerated graphical user interface at all. This means that the entirety of the GUI experience was processed by the CPU on a single thread (aptly named the 'UiThread') leading to mediocre user experiences. The two aforementioned devices lie on each side of this evolution.

The device utilised for the experiment featured a customised distribution of the operating system at version 2.3.6, the last Android version before the jump to version 4.0. Currently, 60% of the android devices in circulation utilise a 2.3.x version of the operating system[18], eschewing any hardware acceleration. This demographic is shrinking by the day with more contemporary devices running version 4.0 and above being activated every day. In the context of this MSc project, the increased computational power and the optimised software that take advantage of the GPUs are of paramount importance for future work. The limitations outlined earlier necessitated the reliance on a server for heavy duty processing.

The server in question is in this case a moderately dated laptop computer running Windows 7 Ultimate edition. The hardware specifications of this laptop were an Intel Core 2 Duo CPU with a clock running at 2.0Ghz, 2GB of RAM and a NVidia 8600M discrete graphics card. More importantly on the software side the server used a distribution of XAMPP which utilises the popular and robust

[18] http://developer.android.com/about/dashboards/index.html

Apache web server at version 2.4.2 and version 5.5 of PHP. This setup allowed the server to execute PHP scripts that interface with the also installed distribution of FFMPEG. The server hosted the video files and performed the necessary thumbnail extraction on request of the user through the mobile application client.

5.3 Mobile Application

The mobile application that was used for the experiments was implemented for the Android environment by using the Android software development kit. The environment is Java based and allows for the rapid development of applications that can run, with varying effect, on a multitude of Android powered devices.

Inspired from the application developed for the Experimentation project, the application was designed from the ground up to have a modular nature. That meant it could act as an experiment framework which would allow for the experiments to act and questionnaires, to be combined in any order. In this case the application was setup to interface with the server and to present the two experiments sequentially with relevant data gathering questionnaires before; between; and after the experiments. In the background the application is designed to silently gather all relevant data required for evaluation such as task success rates; the time required to complete each task; the type of answer given; and the questionnaire data. Upon completion of the experiments, the application generates a CSV sheet containing the results in a format useful for the final analysis, which it then uploads to the server. Additionally, as a redundancy, it dispatches the results via email and also creates a local copy on the mobile device.

5.4 User Testing Process

The 26 participants were tested singly in a variety of settings (home and office). For each test, care was taken to ensure conditions were quiet, without distractions; that ambient lighting was consistent and that the subjects were not pressured by time constraints. The experiment test times ranged from 30 to 50 minutes depending upon how quickly each participant completed the experimental tasks and accompanying questionnaires.

Below, figure 5.2 shows the individual phases of the experiments in the order that they were undertaken.

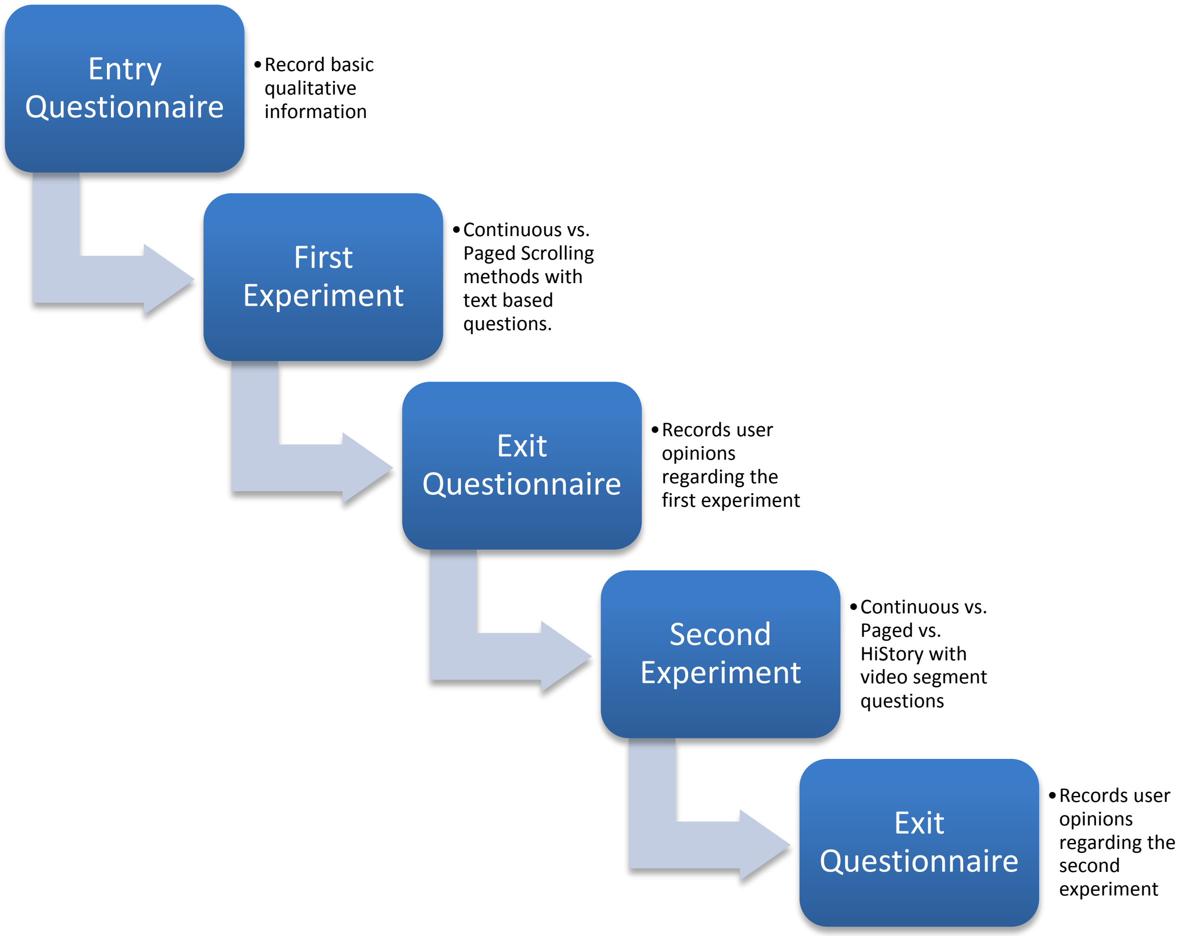


Figure 5.2 - Test phases

5.5 The questionnaires

As the two experiments were administered together, in a single sitting, standard data gathering techniques, such as questionnaires, were administered in the following way and as shown by Figure 5.2. That is, an entry questionnaire before the first experiment began, an exit questionnaire in-between the experiments. This was designed to help to indicate to the test subjects the change between the first experiment, that used text based questions, and the second one, that used video segment based questions and also included the HiStory interface. Finally, the second experiment ended with its own exit questionnaire. All questionnaires were completed using on screen forms as depicted in figure 5.3 below which shows the Entry Questionnaire.

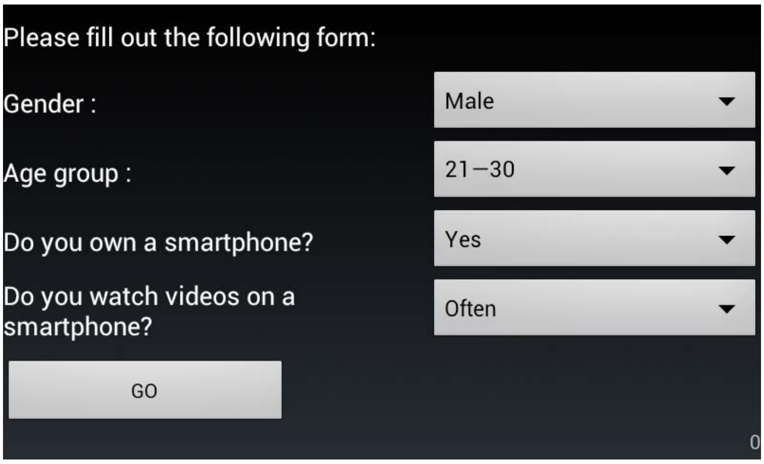


Figure 5.3 – Screen capture of the Entry Questionnaire.

The Entry Questionnaire aimed to gather basic qualitative and quantitative demographic information about the participants with the questions shown above in figure 5.3. The age group choices were broken into groups of -20, 21-30, 31-40, 41-50, 51-60 and 60+. For the question regarding smartphone ownership, a smartphone was defined as a contemporary touchscreen equipped device. And finally, the question regarding whether users watch videos on their smartphone, also queried the viewing frequency.

The second and third questionnaire gathered data relevant to the experiments that subjects had just completed and asked users about their opinions of the experience. These will be described below following each experiment (section 5.6.3 and section 5.7.1 experiment exit questionnaire descriptions) in order to preserve in this narrative the phases of the user testing.

5.6 First Experiment - Scrolling interaction

The first experiment revolves around the issues regarding scrolling across large grids of multiple thumbnails as described in the first section of the previous chapter. The main aim of this experiment was to disambiguate whether the type of scrolling, 'Continuous' or 'Paged' has a significant impact on the performance of simple Known-Item-Search tasks. The photo in Figure 5.4 shows an 'action shot' of a female subject scrolling through grids on a mobile device, her finger is out of focus for that reason.

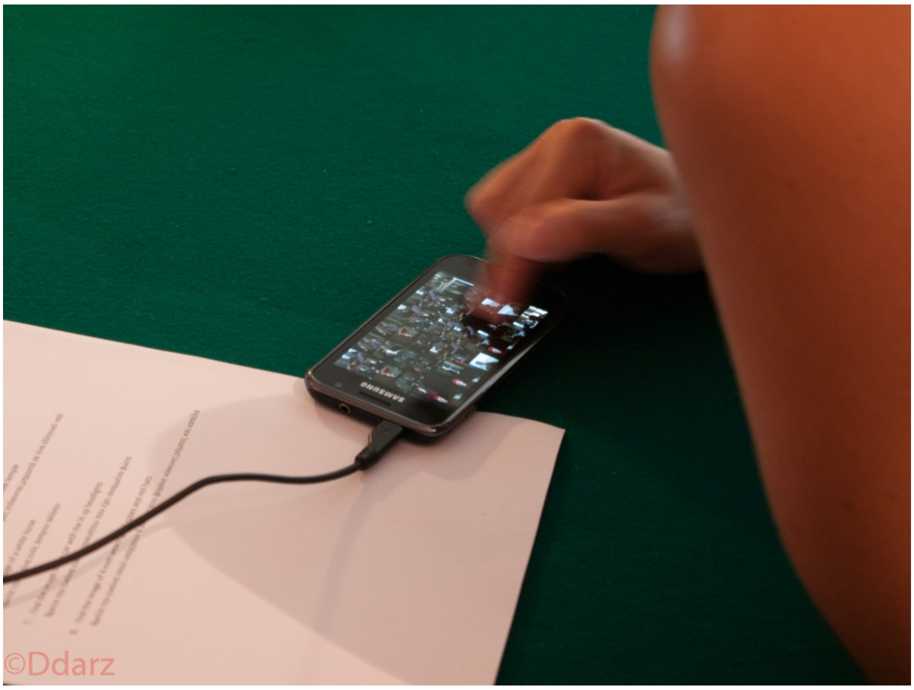


Figure 5.4 – A user scrolls through the grid.

The design and flow of the interface and experiment were a continuation of the experiments performed during the experimentation project (Hürst and Darzentas, 2012) which are described in the extract below:

> "Scrolling Interaction Experiment Interface Rationale & Design
>
> To begin investigating the effects of presenting multiple thumbnails simultaneously on video browsing, a number of different layouts were considered, including grids, linear scrollable strips, three dimensional perspective representations and other less conventional arrangements. As a first step we settled on a grid layout reminiscent of a traditional storyboard, a layout that is intuitive and effectively communicates its purpose. It caters to user expectations as it follows the conventional left to right and line by line reading styles prevalent in all (western) media, such as printed works, comics, etc.
>
> More importantly, the grid layout also allows for the maximum quantity of thumbnails possible in a two dimensional interface without any overlap. Specifically, for the static grids, the number of possible thumbnails that would fit on the screen of the device was determined by the dimensions of the thumbnails. For the smallest size thumbnails (80 pixels) the number of thumbnails was 10x8 = 80 images. This amount decreased progressively as the thumbnail sizes increased. For the largest (200 pixels) images, only 12 images could fit on-screen […]
>
> In the case of the scrolling grid […] while there was no upper limit on the total amount of thumbnails, the quantity that could be displayed on screen at any given time was still held by the same restrictions. For consistency, across all thumbnail sizes we limited the total amount of thumbnails so that the total "scrolling distance" would be about two and a half single screen grids." (page 3)

5.6.1 Known – Item – Search Tasks

The basic concept of the experiment is to present the test subjects with a number of tasks, specifically Known-Item-Search tasks. Practically, this means that the users were given a question in text form e.g. "Find the man wearing a red hat in the video" and they must then utilise the interface to search through the thumbnails in order to find the item described in the question. (The 16 questions asked for the Known-Item-Search tasks can be found in the appendix (9.3) along with a screen shot, of the grid with the correct thumbnail). Once the item was found, users could submit their answer by pressing on the thumbnail and move on to the next question that came up automatically on the screen. They could at any time submit a neutral answer, stating that they cannot find the requested item.

For the purposes of the experiment, the requested item always existed in the video and this fact was made known to the subjects. This was intentional as the experiment is designed to determine the efficiency of the method therefore the test subjects could either find the object they knew existed, or by declaring they could not find it, implicitly say that the interface did not assist them in the task.

In the course of this experiment, a total of 16 questions were asked, 8 for the Continuous scrolling method and 8 for the Paged scrolling method. The users were given 2 paper based question sheets, each with the 8 questions for the particular method. This was a deliberate decision, so that the users could refer to the questions easily. In the experimentation project, where the text based questions were included in the test environment, it had been observed that test subjects stopped searching in order to re-read the questions. Figure 5.5.below shows a test subject with the mobile phone and the question sheet.

Figure 5.5 – A user reads the KIS questions in text form.

Not all of the test subjects were native English speakers -although all had certified proficiency in the language- the questions were in both the Greek (the mother tongue of most subjects) and the English language. One test subject noted that the dual language format was useful as it helped to rule out any ambiguities.

Since ambiguity is a recognised problem with text based video retrieval tasks, special care was taken with the design of the questions. Of course, as each test was administered by the researcher, there was someone on hand to assist the test subjects, but for greater validity, the questions were piloted with three different volunteers, before being given to the 26 member pool of test subjects.

More specifically, the known item search questions were focused on objects/items, as opposed to event search, which is another category in TRECVid. That is the questions asked are of the type “find the birthday cake” as opposed to “find the birthday party”.

Care was taken that questions should not be:

- prone to ambiguity: Ambiguity is caused by many factors. Many, but not all, are language based. Besides having the questions in two languages, care was also taken not to use descriptions that were open to misinterpretations in the context of viewing the thumbnail; e.g. “find the man wearing glasses” where ‘glasses’ means ‘sunglasses’.
- based on colour: This is because the human eye can quickly scan for colour. Where colour is mentioned in the question, e.g. ‘find the woman in the green dress”, the thumbnails contained many other instances of the colour green. At the same time, since colour blindness is not uncommon, especially in the male population, further descriptive details were included in the question, so that even if the colour of the dress could not be distinguished, there were other ways to find the item. In this case the full question read “Find the image of a woman in the long green dress standing in front of the staircase”
- based on faces (as far as possible): Again, this is because face recognition is a task that humans excel at, in particular at recognising (and interpreting) expressions.
- too obvious: One way to ensure this was to check that the item searched for was not the main focus of the thumbnail. Thus, several questions asked subjects to find items that were part of the background of the thumbnail images.
- in same sections of the thumbnail sets: Care was taken to make sure that the items were distributed throughout the sets, so that they would not fall into a pattern of always being ‘towards the middle’. Indeed, one question asked for an item that was located in the very first thumbnail of the selection.

Finally, care was taken that the items asked for should be unique: there should not be more than one instance of the item in the thumbnail set.

5.6.2 Scrolling Interaction

In the Continuous method, the subjects had complete control over their vertical scroll position, being able to navigate up and down the grid as they saw fit. In the Paged method the users were restricted to specific groupings of the thumbnails and could swipe up and down at will, in order to move through the groups of thumbnails.

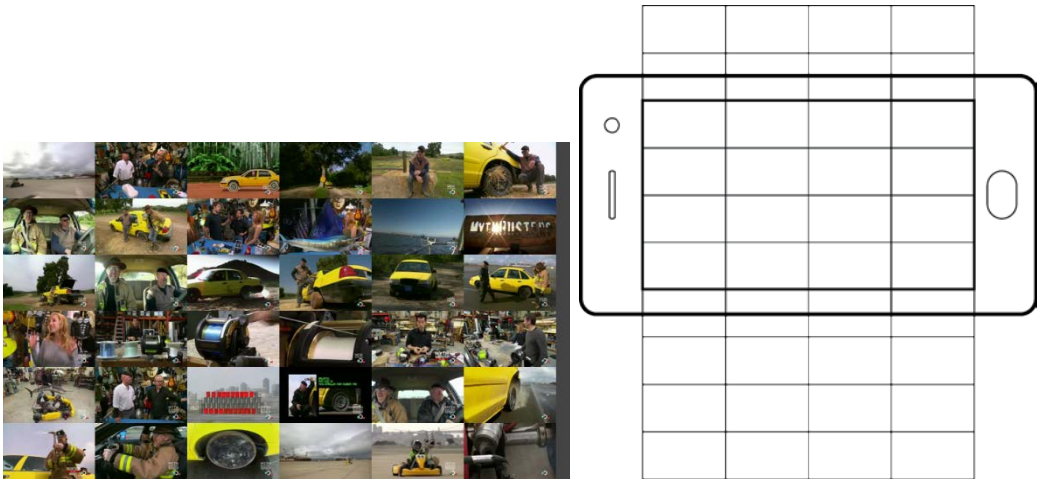


Figure 5.6 – Scrolling Grid.

Based on the results of the experimentation project, a number of refinements where made in order to isolate the desired variables. To begin with, the size of the thumbnails was fixed at 130 pixels, a size which had been determined during the analysis of the experimentation project results to be the size that was deemed the most comfortable to view by the majority of the users and also had the highest success rate in the given tasks. With this specified size, the number of thumbnails that could fit on screen at any given time was 36.

As stated above, the total number of thumbnails extracted per video was 180, this means that a total of 5 screens of thumbnails were viewable. It also means that, in the case of the Paged scrolling model, there were 5 distinct sets or 'pages' of thumbnails to view. In the unbounded Continuous scrolling model, the users are free to scan through the thumbnails at their own preferred scrolling pace.

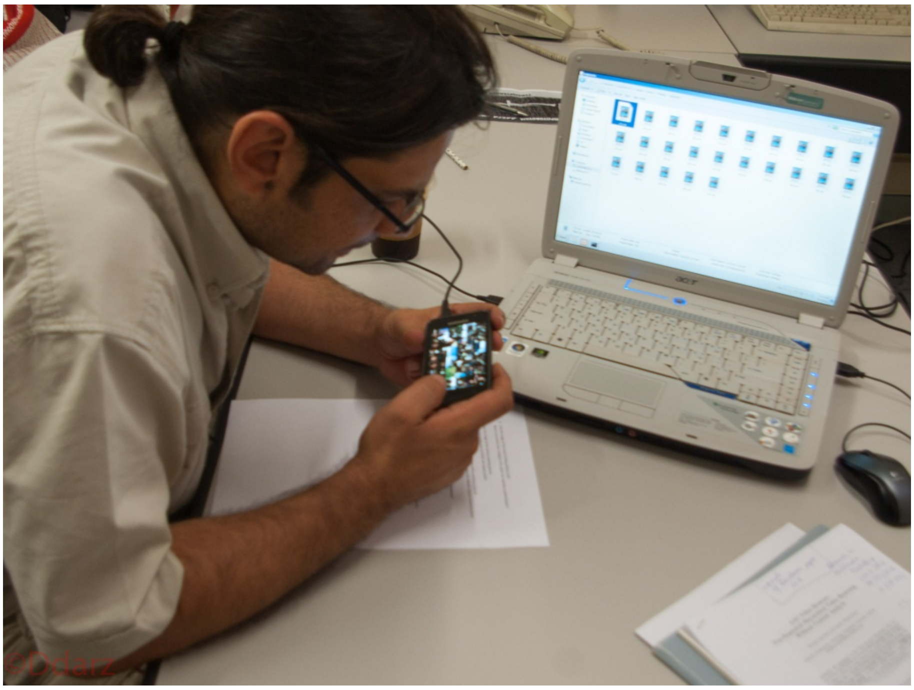


Figure 5.7 – Most of the users found the layout and size of the thumbnails comfortable.

In the background, unknown to the test subjects, the application logged and stored the time, measured in seconds, that it took for each question to be answered. Additionally, the type of answer given was also recorded: 'positive' if the users declared they had found the requested item and 'negative' if they declared that they could not find the item (despite knowing it exists). Finally the application recorded whether the answer given was correct or not (accuracy measures). Figures 5.7 above and 5.8 below provide photographic evidence that subjects were comfortable and not pressured by the set-up.

Figure 5.8 – users were not overtly pressured by time constraints.

5.6.3 First Experiment Exit Questionnaire

Once users had answered all 16 questions, whether by submitting an answer or giving a negative response, then a questionnaire was presented on screen posing a set of questions regarding the user experience, as shown in Figure 5.9.

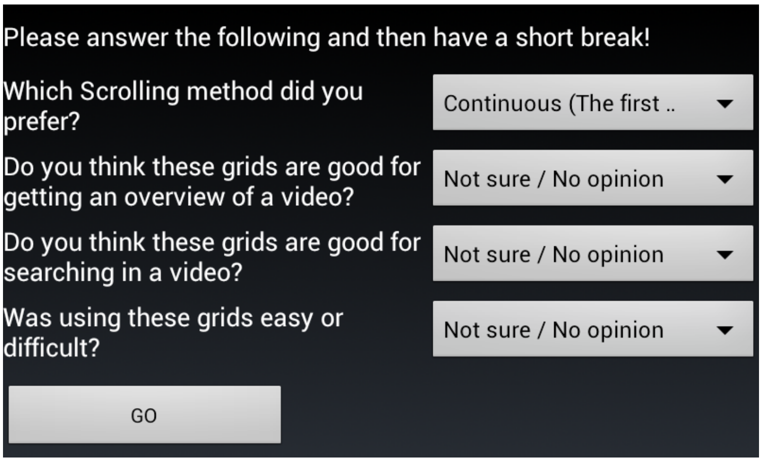


Figure 5.9 - First experiment exit questionnaire

The exit questionnaires asked for various opinions from the users. Specifically users were directly asked to make a decision about the scrolling method. Also questions were asked about search aids, such as video overviews, that were described as 'getting an idea of what the video is about or what it contains', and whether the interfaces helped searching within a video, described as the tasks the user had just performed. Finally the users were questioned about the ease of use of the interface.

The questionnaire is meant to gather qualitative data regarding the test subjects' opinion on the interface, its uses and possible merits or issues. In addition, it served often to engage the test subjects who offered some further comments to the researcher, and these were noted and have been included in the results reported in Chapter 6.

5.7 Second Experiment – Scrolling methods versus HiStory interface

Following the questionnaire pertaining to the scrolling experiment of the previous section, the users were then seamlessly presented with the next experiment.

The experiment consists of 32 questions in 3 sets of 8. The first 16 follow the same interaction pattern as seen in the previous experiment. That is, the first 8 using the same Continuous scrolling interface and the next 8 the Paged scrolling, both with 5 screens of 180 thumbnails. The final 8 questions used the HiStory interface described in chapter 4. The overall goal is to directly compare the three methods by evaluating qualitative and quantitative measures while at the same time eliminating variables such as question ambiguity. This was achieved by giving users a video segment to view instead of a text based question. This follows the format for the 2013 Video Browser Showdown[19].

Figure 5.10 – The 20 second segment provided the users with cues to search for.

As shown above in Figure 5.10, instead of text questions, the user is presented with a 20 second segment extracted from the target video. The users were tasked with locating the viewed segment

[19] http://mmm2013.org/Video_browser_showdown.htm

in the video using the same interface they have been using so far. Practically this means that, the users must search through the video for thumbnails that contain frames from the viewed segment. Once located, the users can submit those thumbnails as answers. Any frame that was contained in the segment is considered a valid answer. For the experiment, the decision was made to allow the test subjects to view the segment as often as they wished, see Figure 5.11 below. The rationale for this was that with text based questions the users were allowed to consult the questions as often as they wished.

Figure 5.11 – Users could view the required segment as often as they wished.

The decision to replace the text description of the known item with actual video segments is based on a number of factors. Previous experiments showed that there was a measure of ambiguity regarding the meaning and interpretation of the text descriptions, especially when language barriers were involved. Additionally, the academic video retrieval community has, for the same reasons, moved away from the use of text based test descriptions, as evidenced by the use of video segments in the Video Browser Showdown and in research publications such as (Cao et al., 2010) which highlight the issues presented by purely textual descriptions in Known-Item-Search tasks.

Once the users had completed the first 16 questions, with the only apparent difference so far being the different question form, the users were presented with the HiStory interface. There were 9 tasks to be completed with this interface. The first one was a tutorial task which was completed with the assistance of the supervising researcher. During the tutorial, the researcher explained to each user the basic functionality of the interface elements such as the non-scrolling grid, the reference bar and the ability to change the granularity of the video by ‘zooming’ in and out. (These features were explained in detail in the Chapter 4 Section 4.2.). Figure 5.12 shows an example of a zoomed in grid, where a particular user has gone one step down.

Figure 5.12 – HiStory, one step down. The selected thumbnail acting as the anchor is circled.

5.7.1 Second Experiment Exit Questionnaire

Following the completion of all the tasks the test subjects were presented with the final Exit Questionnaire which, as with the previous questionnaires, aimed to gather qualitative data regarding the users' experience and opinion of the interfaces, with emphasis on the HiStory interface in this case.

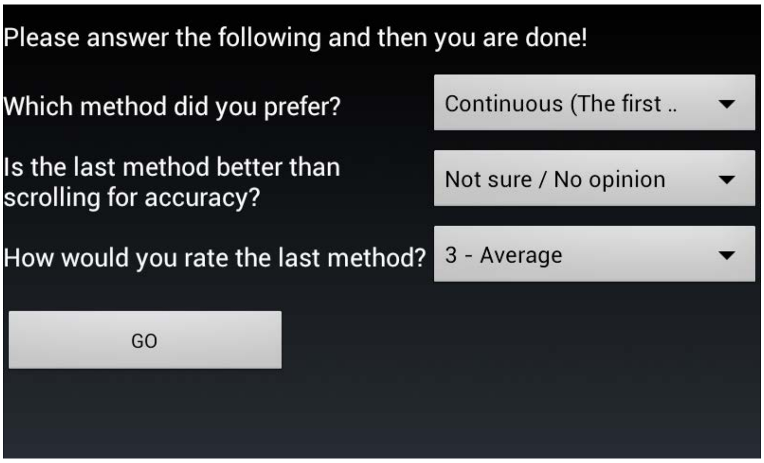


Figure 5.13- Second experiment onscreen Exit Questionnaire.

The details of this questionnaire are as follows:

For the question, "Which scrolling method was preferred?" The choices offered were "Continuous", "Paged" or "HiStory". For the next questions that used HiStory as point of reference, the questions focused on accuracy, not time or any other factor, while the test subjects were asked to rate the HiStory concept, but not its execution. As with the other exit questionnaire, users engaged in comments that were recorded by the researcher.

The logged results of the task execution of both experiments along with the 3 questionnaires (1 entry, 2 exit) are analysed and interpreted in the next chapter, Chapter 6. Added to the logged results are some qualitative data arising both from test subjects' comments to the researcher, as well as their observed reactions to the experiments and test environment.

6 The results

Following the execution of the experiments a number of useful and interesting findings were gleaned from the resulting data. In short, these findings indicate the following:

- For the case of the first experiment which evaluates scrolling
  - The majority of users prefer Continuous scrolling.
  - Paged scrolling is shown to be slightly slower, but significantly more accurate.
- For the case of the second experiment evaluating HiStory and comparing with scrolling
  - HiStory is more accurate with comparable speed
  - Users unanimously believe it is more accurate
  - The majority of users believe is a very good concept
  - Users were split between preferring Continuous scrolling or HiStory

In the following sections, the results and corresponding findings will be analysed in more depth.

6.1 Users

The experiments were administered to a total of 26 tests subjects with a variety of backgrounds. Gender wise, there were a larger number of male participants with just over a quarter of the subjects being female, as illustrated in figure 6.1. This is not a major issue as the use of mobile devices, and video consumption on such devices has, so far, spread equally across both genders (Economides and Grousopoulou, 2008).

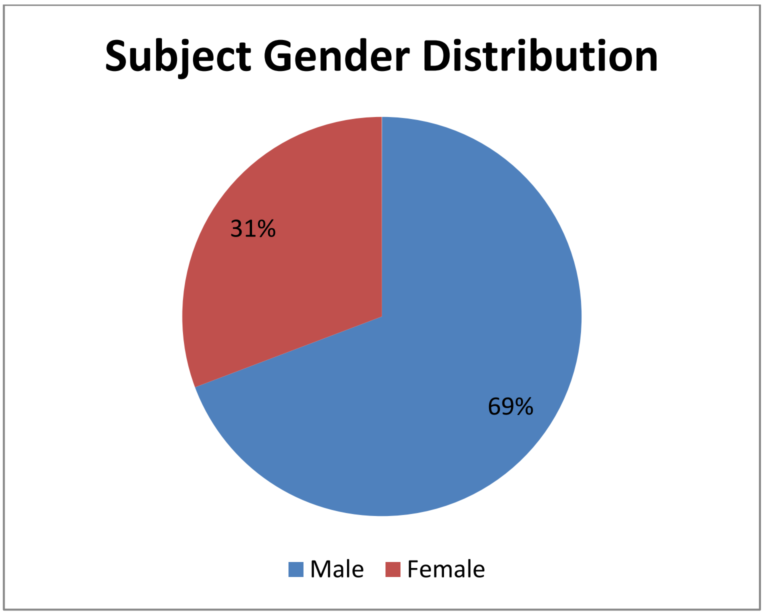


Figure 6.1- Subject Gender Distribution

The subjects were drawn from a variety of age groups covering a large spectrum. This allowed for the creation of a clearer picture regarding the opinion of mobile devices users as such technologies are not just the province of younger users. Indeed the fastest growing user demographic has been the Baby Boomer generation[20]. Figure 6.2 uses a pie chart to illustrate the distribution of the test

[20] http://www.newmediatrendwatch.com/markets-by-country/17-usa/855-mobile-devices

subjects into predefined age groups. The distribution covers a large portion of the age spectrum, from very young users (12 years old) to older users (60+ years old).

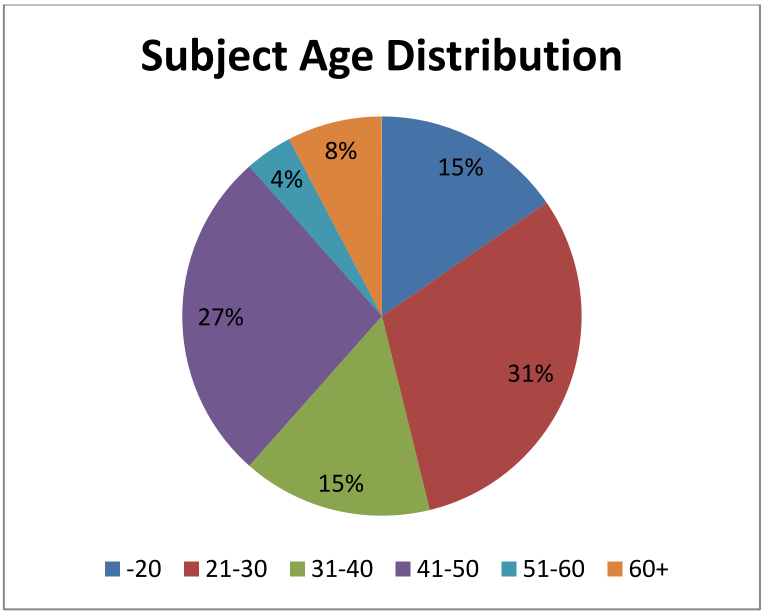


Figure 6.2 - Subject Age Distribution

The test subjects were additionally queried on whether they owned smartphone devices and whether they watched videos on mobile devices (regardless of ownership) and if so, with what frequency. Figure 6.3 illustrates the percentages of users who owned smartphones and used them on a daily basis and the percentage of users who did not own a smartphone or who owned one and used it too seldom to count as proficient users. For instance, one test subject explained: "I purchased a smartphone but returned to using an older model because I prefer physical buttons."

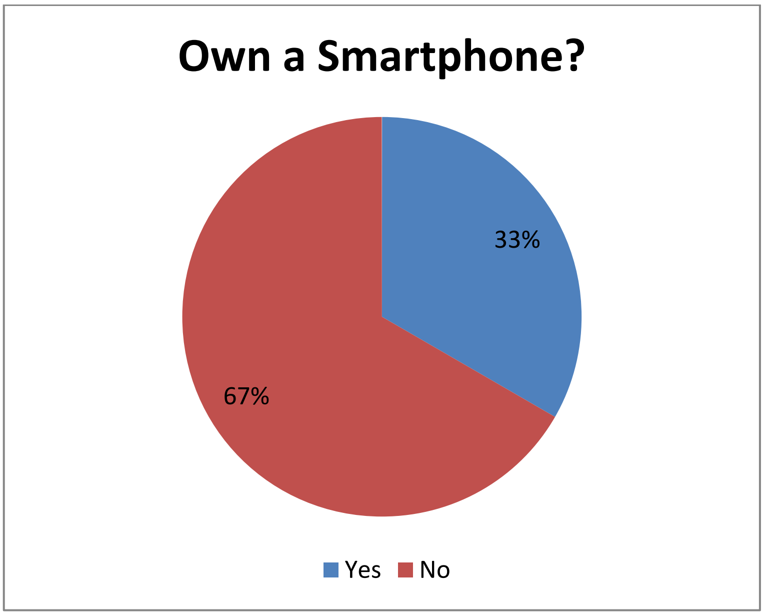


Figure 6.3 - Percentage of users who owned a smartphone and used it daily.

As illustrated, comparatively fewer users 'owned' smartphones. This is not a direct reflection on their technological skills however. Users who owned and used a smartphone regularly were not always comfortable with the latest technology and purchased the device due to lack of interest in researching alternative options or even simply as a status symbol. Conversely, users who did not 'own' a smartphone could be quite proficient with its use and be comfortable with technology. They either could not afford a device of their own, or had other –non technical- reasons. This information came from users' remarks and comments to the researcher when answering the Entry Questionnaire. Figure 6.4, below illustrates how often the test subjects utilised mobile devices to watch videos. Videos, in this case, are defined as short video clips to full length feature films.

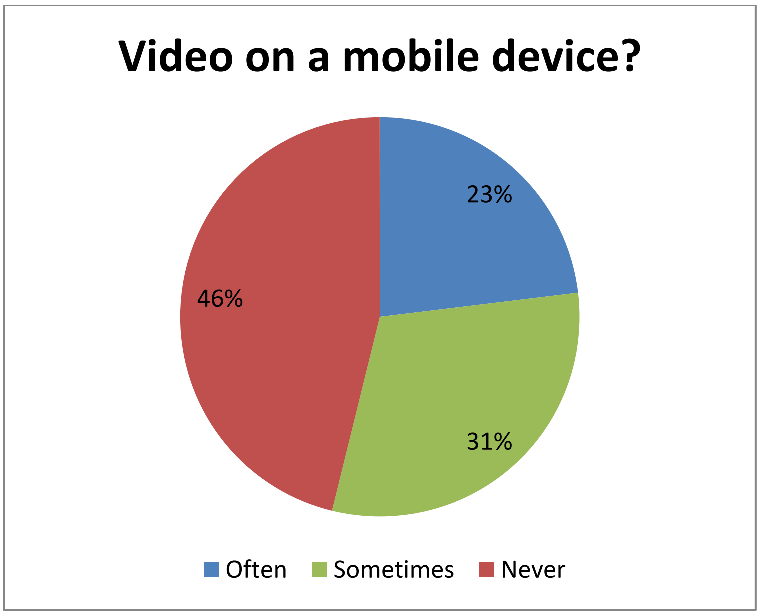


Figure 6.4 - How many users actively watched videos on mobile devices and with what frequency.

The above figure illustrates the point that, regardless of ownership or proficiency, almost half the users queried stated that they "Never" watch video on a mobile device. Most users cited their lifestyle as the reason for this, stating that when they had the opportunity, or were in the mood to watch a movie, they would simply watch it on a large screen television. Users who voiced such opinions were often in the more advanced age groups. Members of these age groups were often very entrenched in their ideas of media consumption, having experienced many changes in media trends, some even going back to complete lack of televisions. Therefore selective viewing and personal media consumption were eclipsed by concepts such as fixed television programming ("I will watch whatever is on"). On the surface these findings appear to conflict with reports such as (Taneja et al., 2012) that people watch videos on mobiles but on further investigation, the users were not opposed to the idea and stated that, had they the chance, time and financial means, they would gladly adopt the practice. Already, on mobile devices that have larger sizes, such as tablets, a recent press release[21] by comScore indicated that the majority of users watch video on their device and indeed pay for the privilege.

In the following sections the particular findings of each experiment will be presented in detail accompanied by further user related qualitative data such as responses and opinions.

---

21 http://www.comscore.com/Press_Events/Press_Releases/2012/6/Majority_of_Tablet_Users_Watch_Video_on_their_Device

6.2 First experiment

The first experiment pitted Continuous scrolling directly against Paged scrolling in a series of 16 tasks split into two groups of 8 text Known-item-search questions for each scrolling method. The application gathered data with regard to the validity of the answers given by the users in each of the tasks (Correct or Incorrect) and measured the time it took each user to complete each task. The following figures illustrate the results.

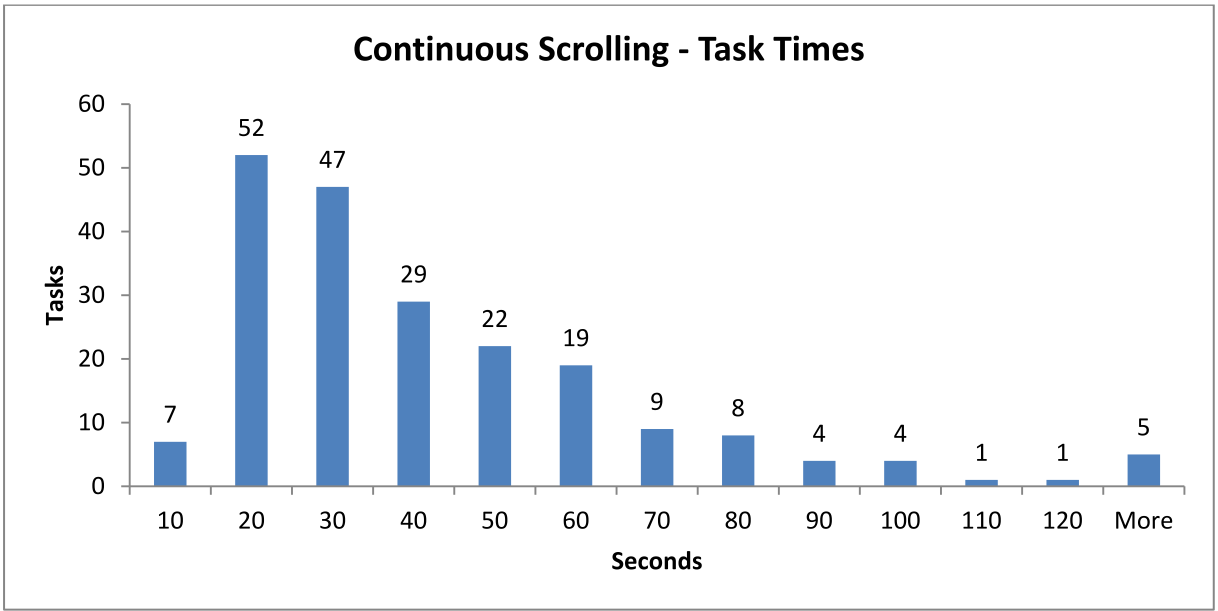


Figure 6.5 - Time required to complete the Continous scrolling tasks.

Figure 6.5 above, illustrates that the majority of all the (208) tasks given to the 26 test subjects (26 x 8 = 208) were completed in less than 60 seconds with most of them completed in less than 40 seconds. The average time across all tasks was 38.7 seconds. In relatively few cases users required more than 60 seconds and in a few outlying cases, more than 120 seconds. This falls well within acceptable limits, both practical and academic, such as those stipulated in the Video Browser Showdown[22].

In the following figure 6.6, the accuracy of the Continuous scrolling method is illustrated by displaying the number of incorrect and correct answers to the total of 208 questions. Neutral answers, meaning when the user declared that they could not find the requested item, were counted as incorrect given that the item in question always existed and the users had knowledge of this fact.

[22] http://mmm2013.org/Video_browser_showdown.htm

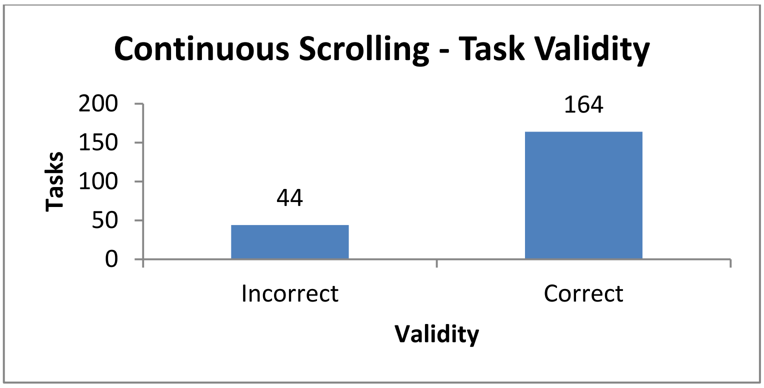


Figure 6.6 - Success rate for the Continuous scrolling tasks

Specifically, just 21.2% of the answers given were incorrect and the remaining 78.8% were correct giving the Continuous scrolling method a very high percentage of success. When compared to the results of the Experimentation project (Hürst and Darzentas, 2012), the results of this experiment show a different picture, where the tasks featuring scrolling interaction scored much worse than those without. Yet considering the varying degree of proficiency and temperament towards technology shown by the test subjects in this experiment and the diverse age groups involved, a success rate of more than 75% coupled with the fast completion times indicates fairly clearly the effectiveness of the method.

Immediately following the 8 Continuous scrolling tasks, the 26 users were administered 8 Paged scrolling tasks.

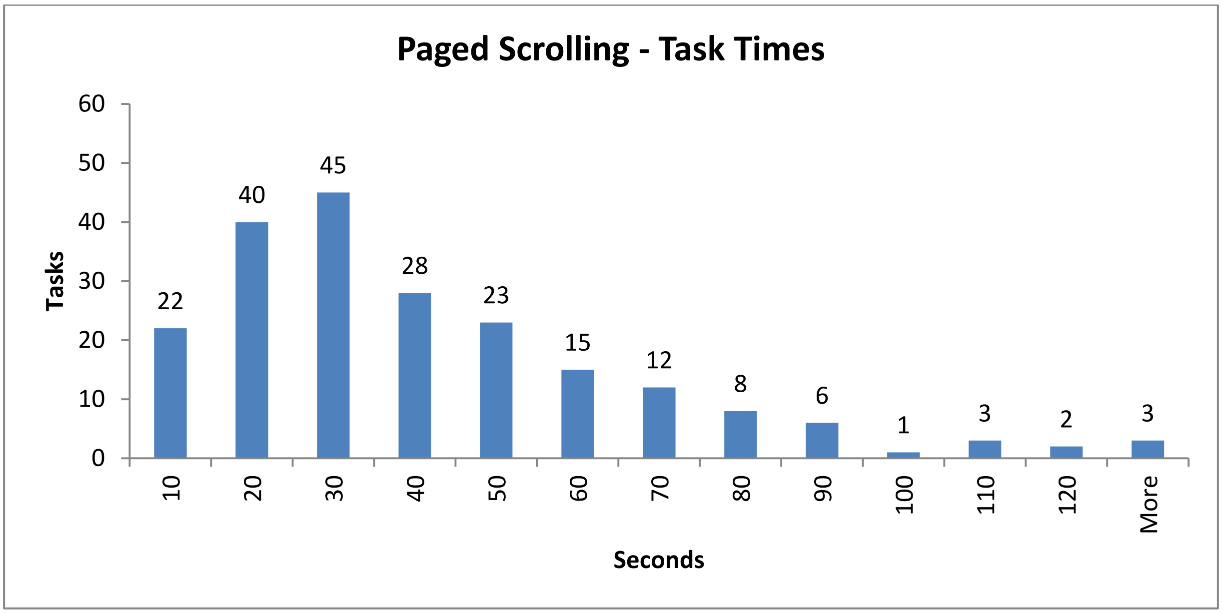


Figure 6.7 – Time required to complete the Paged scrolling tasks.

The above chart shows the recorded time for completion score, in seconds, for the 208 Paged scrolling tasks. As with the case of the Continuous scrolling method, the 208 tasks are comprised of the 8 tasks given to each of the 26 test subjects, totalling 208 readings. The average time to complete a Paged scrolling tasks was 37.2 seconds which is comparable with Continuous scrolling. Again the majority of the tasks were competed in less than 60 seconds and most of those in about

30 or less seconds. The following figure 6.8 illustrates the success rate for the Paged scrolling method.

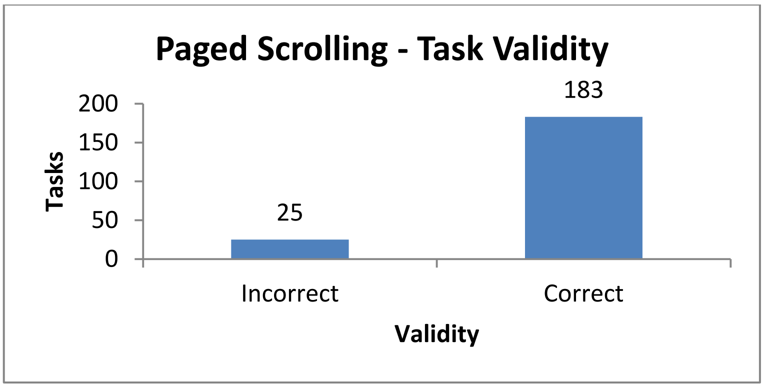


Figure 6.8 – Success rate for the Paged scrolling tasks

While Continuous scrolling achieved a high success rate of about 78%, Paged scrolling achieved a much higher success rate of 87.9%, almost 10% better results, as illustrated in Figure 6.11. Combined with the slightly faster completion times Paged scrolling appears to be a more effective method than Continuous scrolling, but the qualitative data discussed later creates a different picture.

In order to concretely deliver a quantitative comparison of the two methods, a statistical test (t-test, see appendix 9.4.1) was conducted. In the following figure, (Figure 6.9), a direct comparison between the mean times required for completion, for each method, is shown.

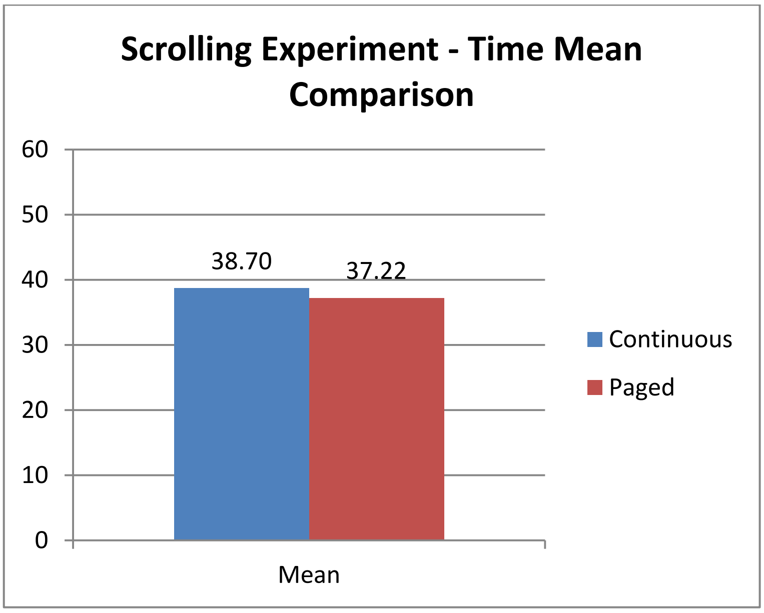


Figure 6.9 – Comparison between the average times for Continuous and Paged scrolling.

In figure 6.10 below, the task times for both methods are displayed in a direct comparison. Here, it can be observed that the Continuous and Paged scrolling remained competitive with neither being overtly faster than the other.

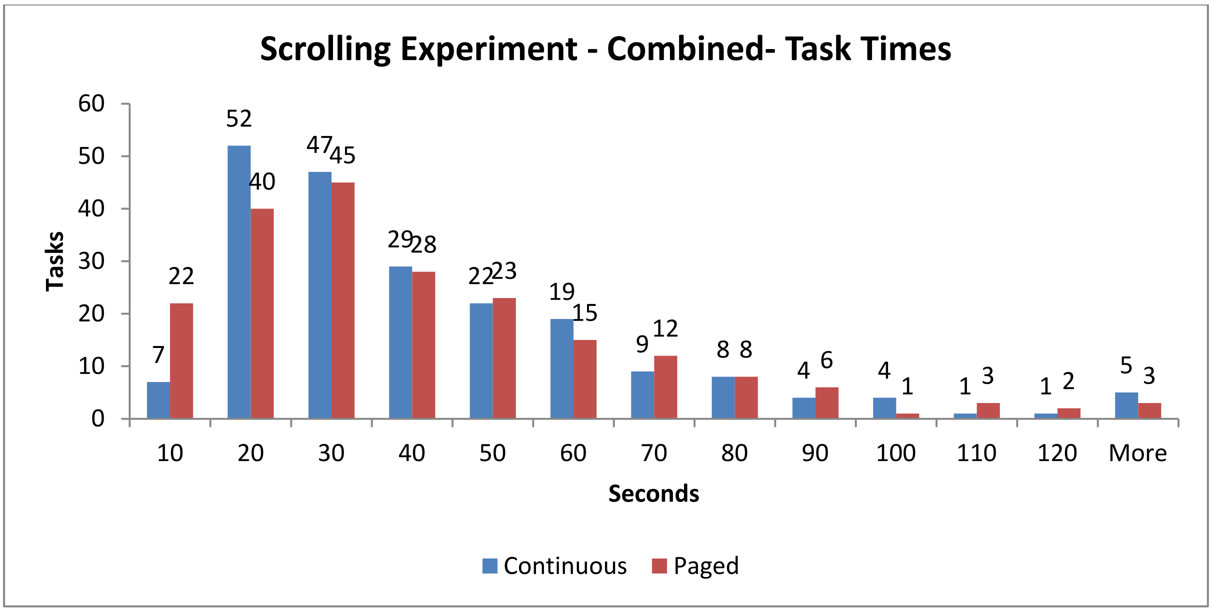


Figure 6.10 – Combined task times illustration for the scrolling experiment

To confirm this, the statistical testing showed no evidence that there is a significant difference in the times required to complete the tasks of each method. Therefore it can be surmised, that one method is not significantly faster than the other. However, this is not the case for the success rate (accuracy), as illustrated below in figure 6.11.

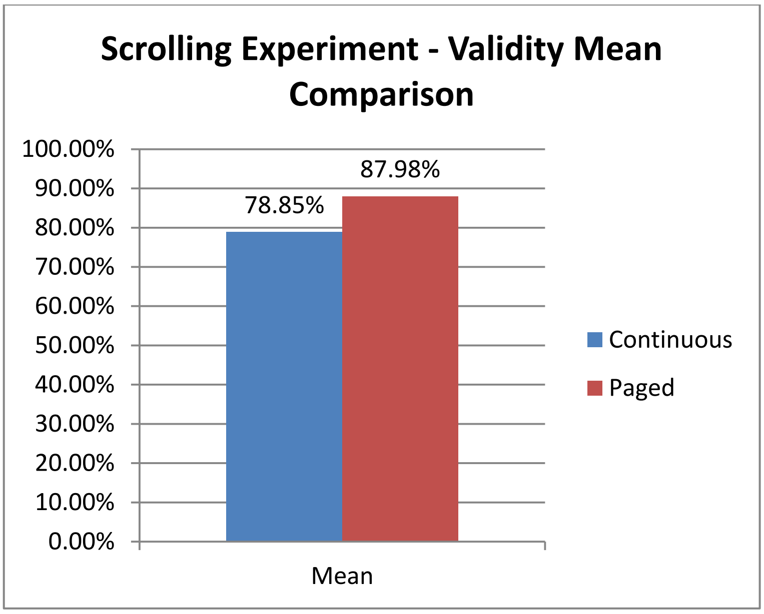


Figure 6.11 - Comparison between the average success rates for Continuous and Paged scrolling.

In the case of accuracy, as shown in the figure above, there is a considerable difference in the average success rates for the two methods. A finding that validated the worth of the method, despite its relative 'unpopularity' with users who preferred Continuous scrolling as described next.

Following the conclusion of the 16 tasks administered to each user, a questionnaire was filled out with questions pertaining to the two methods (Continuous scrolling/Paged scrolling). The first question asked the users to subjectively state their preferred method. As shown in the following pie chart in figure 6.12.

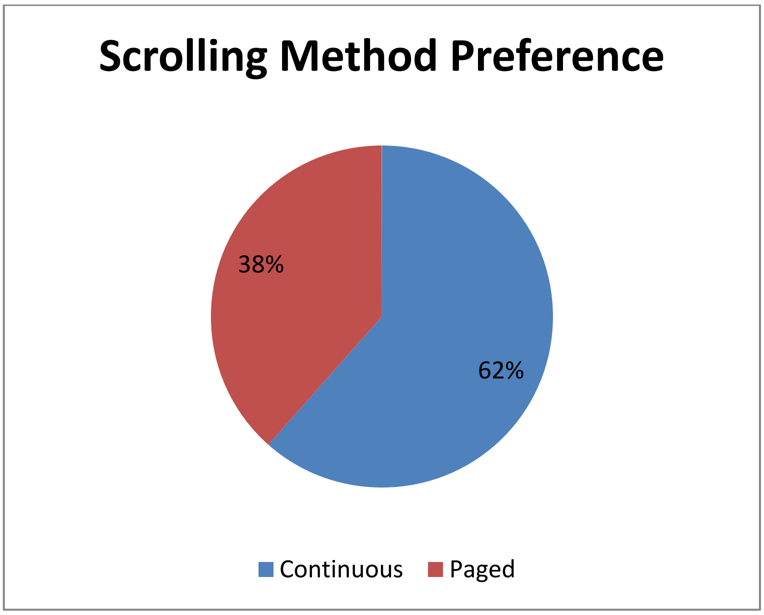


Figure 6.12 – User preference between Continuous and Paged methods.

The majority of the users preferred the Continuous scrolling method and many voiced their approval or disapproval with conviction. Users preferred the freedom of Continuous scrolling, stating that it allowed them to scroll at their own pace and segment the thumbnails as they required. Users who did not prefer the Paged scrolling method stated that it felt slower and that they worried that they might have missed a line of thumbnails. This is an opinion directly opposite to that of the users who did prefer Paged scrolling, who expressed that the most powerful advantage was the guarantee that they had not missed any thumbnails. Additionally some users, it would appear, recognised that the Paged scrolling method afforded them the advantage of spatial memory as, when browsing through the grid, the thumbnails always appeared in the same position. This fact, in all probability, contributed to the high success rate (Accuracy) of the method.

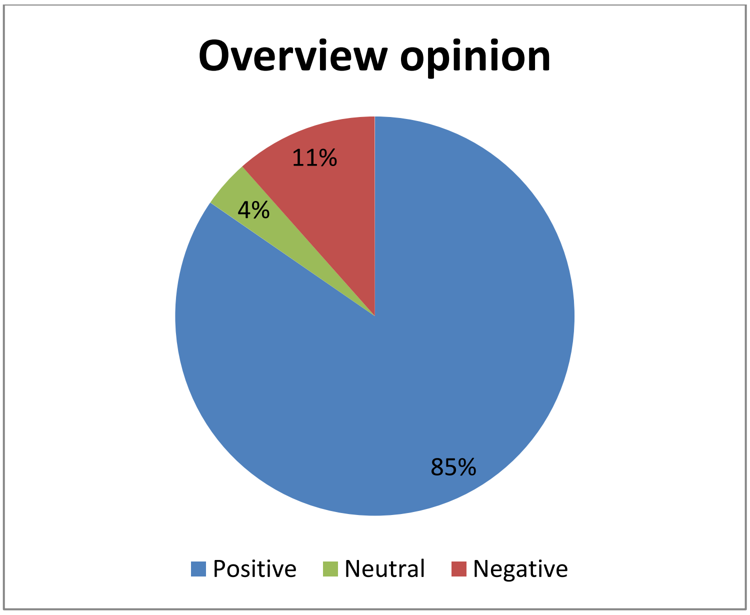


Figure 6.13 – User opinion on whether the interface provides an Overview.

Figure 6.13 above, illustrates the users opinions on whether the Grid interface afforded an effective overview of the content of a video, regardless of interaction. The users overwhelmingly held the belief that the thumbnails grid offered a very effective overview of a video with some even expressing concerns over 'spoilers', a case of course that does not apply to all content e.g. surveillance Footage, but is a valid concern in other applications such as films, sports, etc.

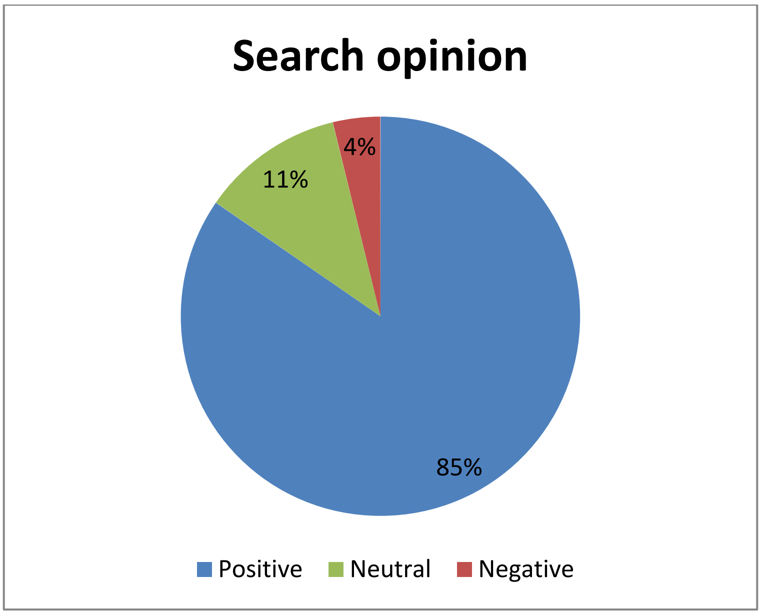


Figure 6.14 – User opinion on whether the interface serves for searching.

Next the users were queried on their opinion of whether the grid interface facilitates search tasks within a video, such as those they had just completed. Again the majority of the users stated that they found the methods effective. A few, especially for the more advanced age groups and those needing reading glasses found the system effective, but uncomfortable, due to the size restrictions, not the layout. The question of ease of use is discussed below.

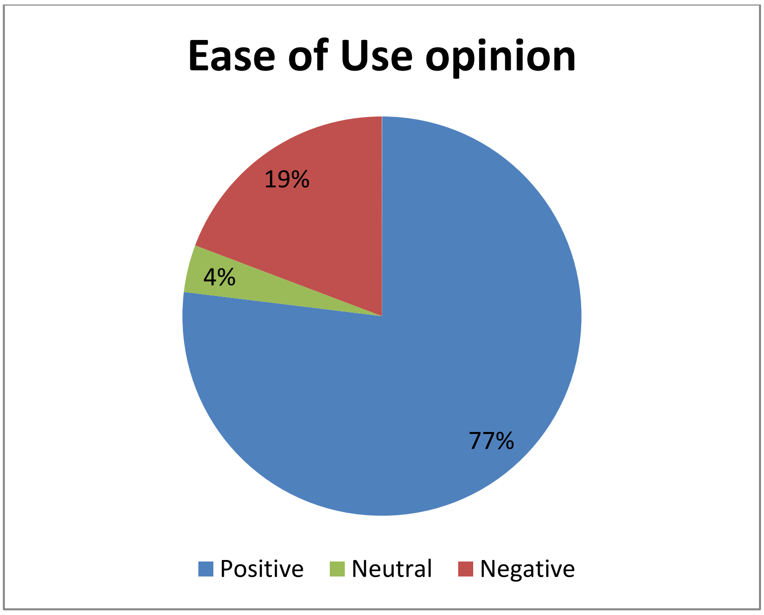


Figure 6.15 - User opinion on whether the interface was easy to use.

Above, Figure 6.15 shows the subjective opinion of the users regarding how easy the overall grid method was. Again most of the users found the concept easy to grasp and the functionality straightforward. These findings are in line with the conclusions of the experimentation project (Hürst and Darzentas, 2012) which showed that users were fast to adopt the mechanisms of the grid and fast to develop effective strategies to complete the given tasks. Complaints that were voiced tended to be from older users who stated deteriorating eye-sight combined with the small thumbnails made them unsure about their choices and they felt that they might be making mistakes. Nevertheless their results were quite accurate.

To summarise, the thumbnail grid method has been validated both in the experimentation project and in this experiment as an intuitive interface for the completion of Known-Item-Search tasks within a single video on a mobile device. On the matter of interaction style, (i.e. Continuous versus Paged scrolling), the principle variable of this experiment, the findings are of a contradictory nature. Firstly, the success rates for the tasks, both with Continuous and Paged scrolling are higher and more consistent that those displayed in the similar portion of the Experimentation Project. This is a finding that indicates that additional physical interaction is not a detrimental factor for the thumbnail grid based interfaces as initially feared.

Secondly, regarding the question of which scrolling method is better suited for the task at hand, the picture, quantitatively speaking, is quite clear, but less so in the case of the qualitative data. The Paged scrolling tasks were completed with a higher accuracy. However, most users firmly held the belief that Continuous scrolling was easier; more pleasant due to increased control; and much faster. It is worth noting that the relatively fewer users who preferred Paged scrolling were the ones who quickly adopted specific and methodical strategies in order to complete the tasks.

Finally, another outcome of this experiment was that in spite of the best efforts to make the task questions as unambiguous and as clear as possible, several test subjects voiced concerns over the textual descriptions of the items they needed to find (the known item search questions) as there were issues of interpretation and nuance This is an issue addressed in the next experiment where the 'known item' cue is given by actually watching the segment that is needed to be found.

Based on these findings, a clear recommendation cannot be made for one scrolling method over the other. A clear quantitative advantage does not overrule the user preferences shown by the qualitative data. Indeed in a production implementation, it would be best to include both methods as options. And perhaps to give the users the option to switch between them as they think fit.

6.3 Second Experiment

The second experiment is larger in scope and examines two major aspects, the first being the departure from text descriptions for the Known-Item-Search tasks and the second is the inclusion of the HiStory interface. Of the 24 tasks given to each of the 26 test subjects for this experiment; the first 8 tasks used the Continuous scrolling method; the next 8 tasks used the Paged scrolling method and the final 8; the HiStory method. Therefore the results of the first 16 tasks can directly be compared to the findings of the first experiment in order to determine the effects of switching from text 'known item' descriptions to actual user knowledge of the item by watching the required video segment. In order to avoid any potential issues with reliance on memory, the users were allowed to view the video segment as many times as they wished, whenever they wished. Finally, of the 26 users, the recorded data for two subjects had to be discarded as one was obliged to finish the experiment hurriedly due to a personal emergency that arose and the other subject's reading glasses broke unexpectedly. Therefore the quantitative results presented below are calculated without these two outliers, but the qualitative data from the questionnaires remains, as the two users submitted valid data through the questionnaires.

The first presented results are the recorded times required to complete each task. Figure 6.16 below illustrates the results for the first 8 tasks which featured Continuous scrolling.

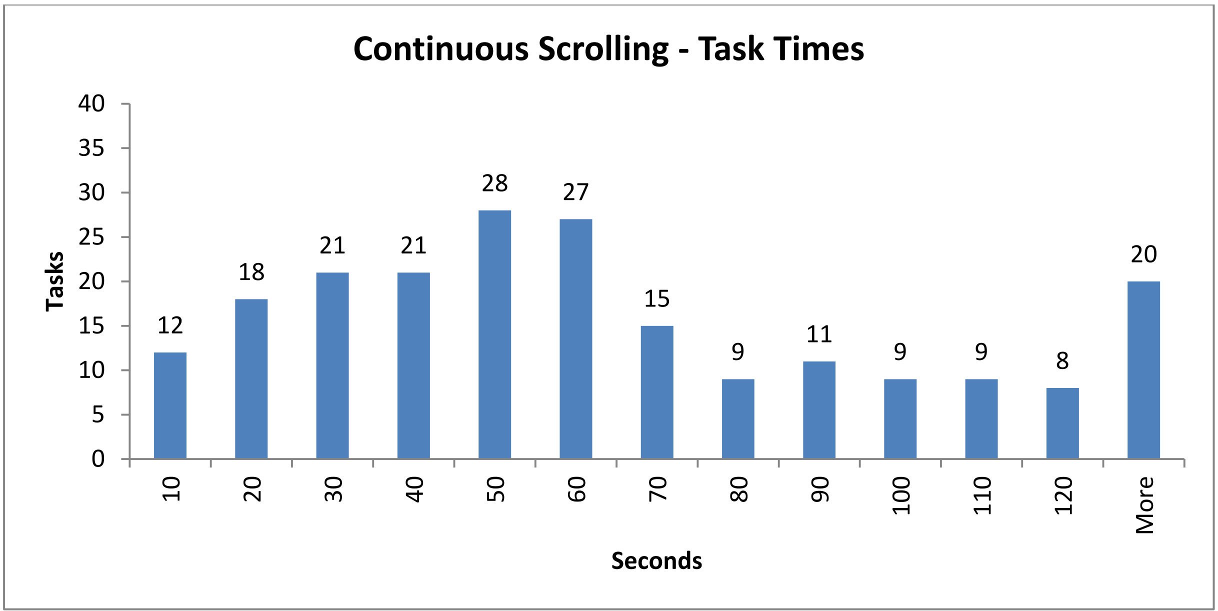


Figure 6.16 - Time required to complete the Continous scrolling tasks.

Despite an average time of 61.5 seconds, most users were still able complete each task in less than 60 seconds with the higher average being the result of a number of users who, by repeatedly watching the requested segment, required more time to answer.

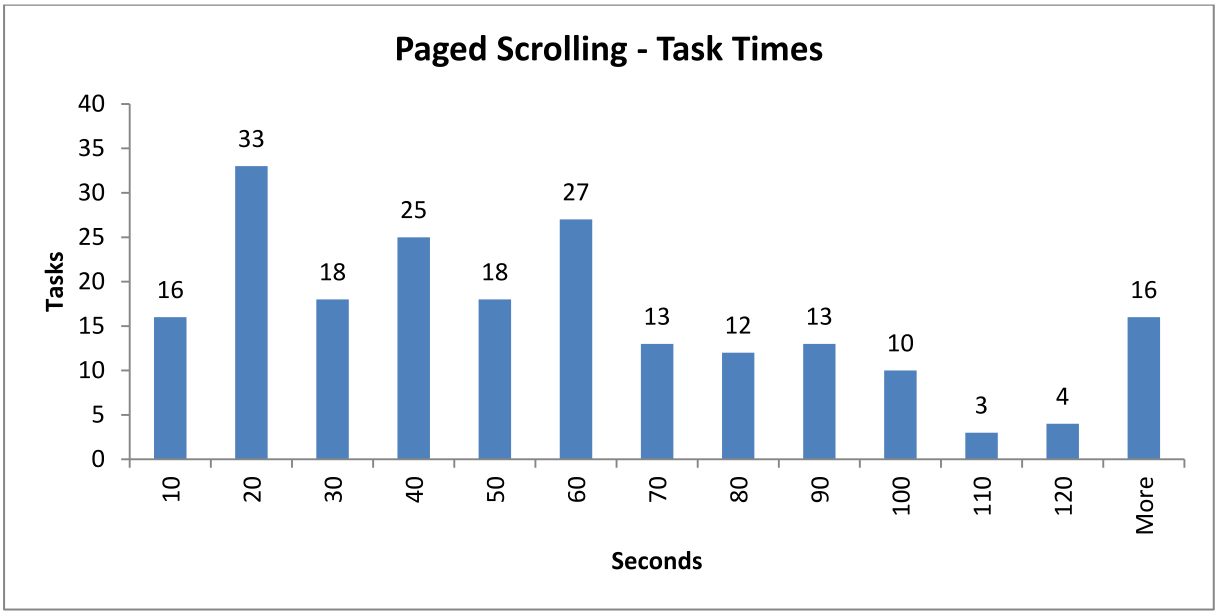


Figure 6.17 - Time required to complete the Paged scrolling tasks.

Unlike in the previous experiment, the Paged scrolling method, unintuitively according to some users, achieved faster scores with regards to the time required to complete each task. With an average of 55.6 seconds, a statistical test (t-test, see appendix 9.4.2) showed that there is no evidence of significant difference between Continuous and Paged scrolling, as far as time required is concerned, further validating the finding than that there is no appreciable speed difference between Continuous and Paged scrolling.

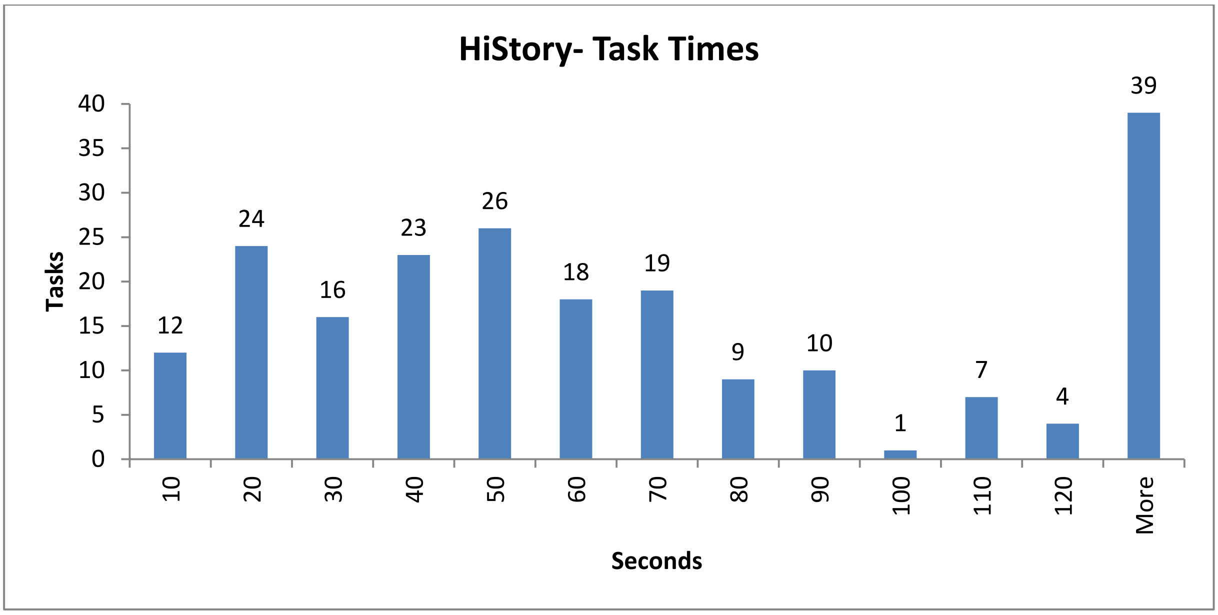


Figure 6.18 - Time required to complete the HiStory scrolling tasks.

A different picture is shown above in figure 6.18 where the time scores for the HiStory method are illustrated. At first glance, a large number of the tasks appear to have required more than 2 minutes to complete. With an average of 71.6 seconds, there is a considerable increase from the scores of the previous methods. Following statistical tests (t-test, see appendix 9.4.3 and 9.4.4), there is evidence that there is a significant difference between the HiStory method, and the other two methods in the time required to complete a given task. This phenomenon has its roots in a number of reasons, with the principal issue being the inability of the hardware to recreate the thumbnail grid fast enough, therefore creating considerable time delay. Thus the HiStory method cannot be immediately penalised for the overlong completion times.

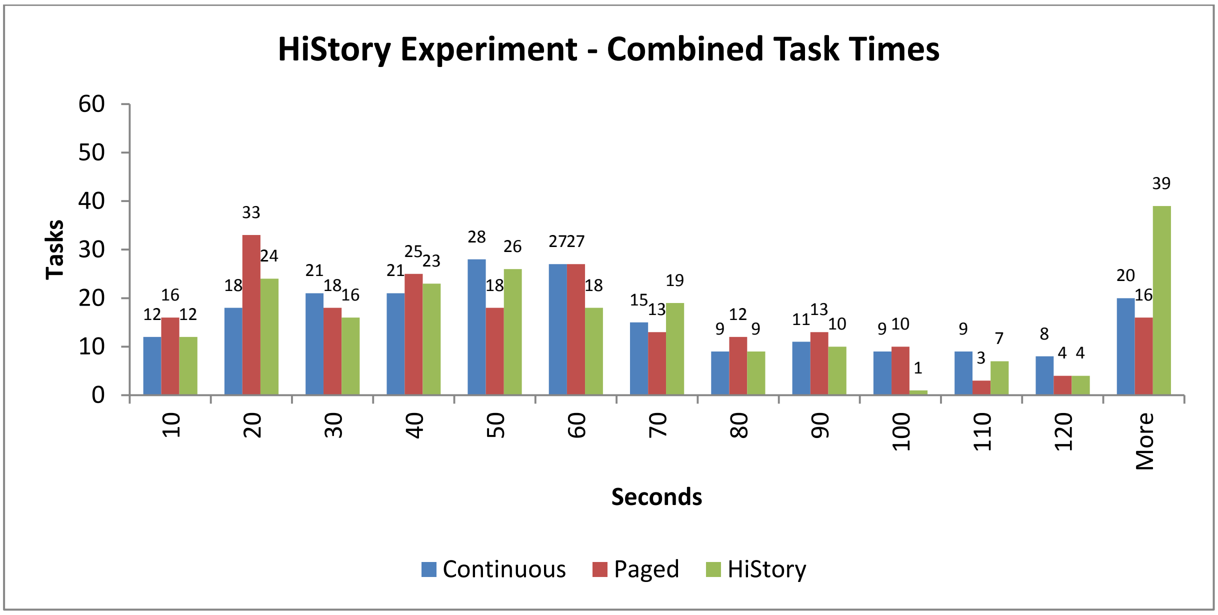


Figure 6.19 – Combined times for the HiStory Experiment

The above figure 6.19 combines the times of each method into a single chart for easier comparison Indeed, despite the technical handicap, HiStory held its own in a considerable number of tasks with the exception of the tasks that required more than 120 seconds, purely for technical reasons. Continuous and Paged scrolling remained close competitors as with the previous experiment. Next analysed are the results for method accuracy.

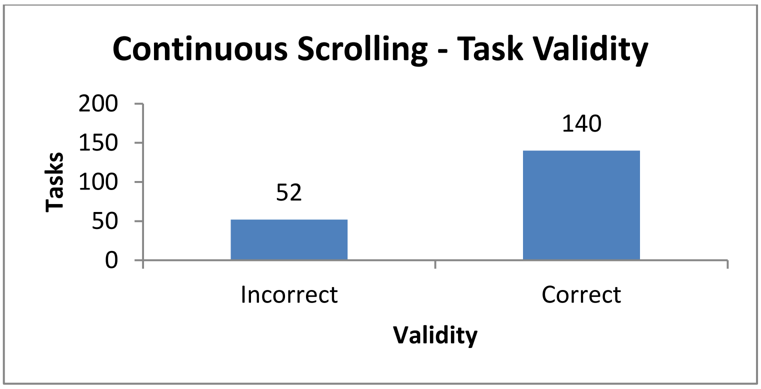


Figure 6.20 - Success rate for the Continuous scrolling tasks.

Above in figure 6.20, the recorded results for the Continuous scrolling method accuracy are displayed. Continuous scrolling retained a high success rate with an average of 72.9% while below, in figure 6.21, Paged scrolling had a similar accuracy rating as in the previous experiment, with an average success rate of 69.7%.

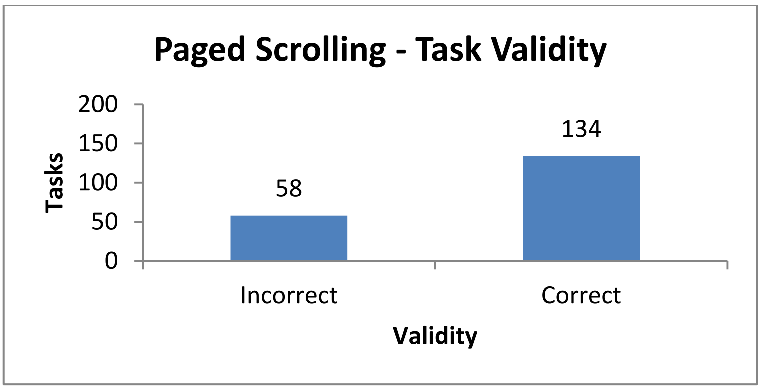


Figure 6.21 - Success rate for the Paged scrolling tasks.

Below in figure 6.22 is an illustration of the success rates of the HiStory method. In this case the success rate was considerably higher than the other methods with an average rate of 82.2%. This can be directly attributed to the nature of HiStory which eliminated guesswork as the users can ‘dive’ or ‘zoom’ into the video repeatedly, gaining a finer viewing granularity down to a frame by frame level and find the exact item they are searching for.

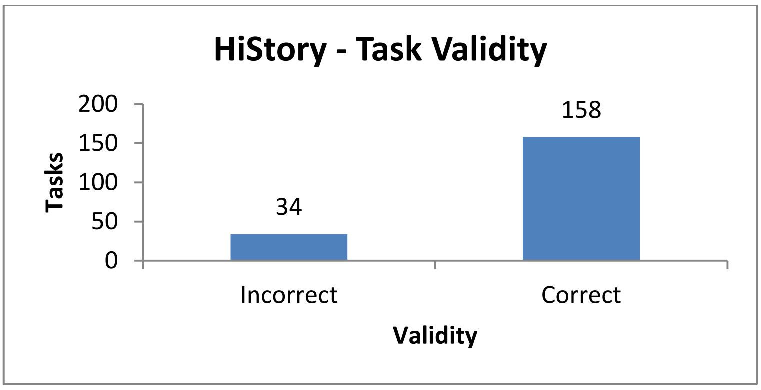


Figure 6.22 - Success rate for the HiStory scrolling tasks.

The following figures illustrate the comparison between the mean values of the three methods, both for the time required to complete (Figure 6.23) the tasks and the accuracy (Figure 6.24) through the validity of the answers given.

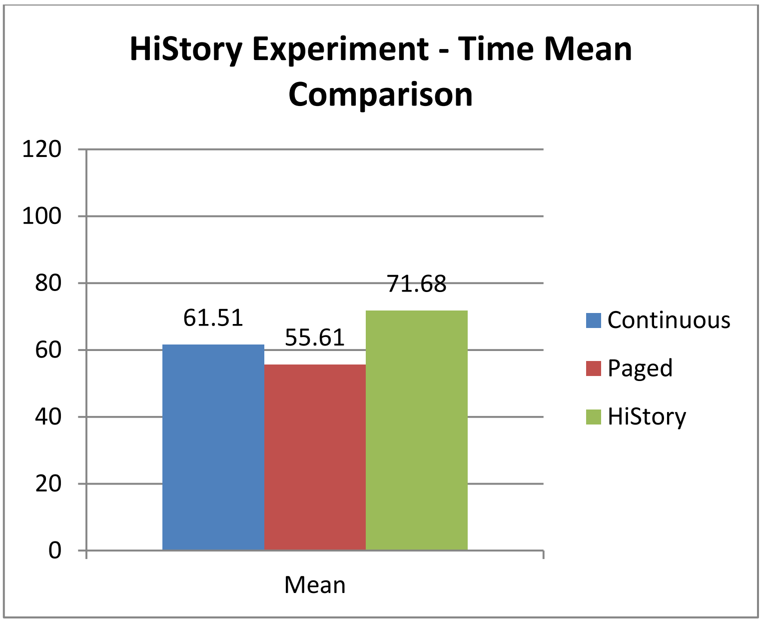


Figure 6.23 - Comparison between the average times for Continuous and Paged scrolling.

As mentioned earlier, the Paged scrolling method again had the fastest average time-to-complete, especially when compared to the technically handicapped HiStory method. However it was not statistically significantly faster than the Continuous scrolling method (See appendix 9.4). Figure 6.24 below illustrates the direct comparison between the average success rates for the three methods. Here the much higher average success rate achieved by the HiStory method, compared to the average success rates of Continuous and Paged scrolling method should be noticed.

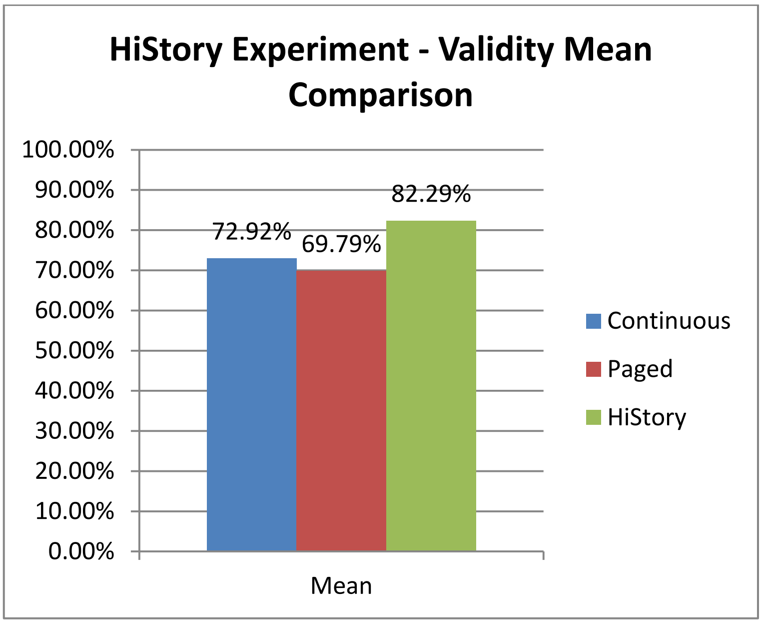


Figure 6.24 - Comparison between the average success rates for Continuous and Paged scrolling.

Following the completion of all the tasks of the second experiment, the users were again presented with a questionnaire, in this case focused on the HiStory method. The users were queried regarding their method preference, their opinion of HiStory method's accuracy and finally their opinion of method's Concept.

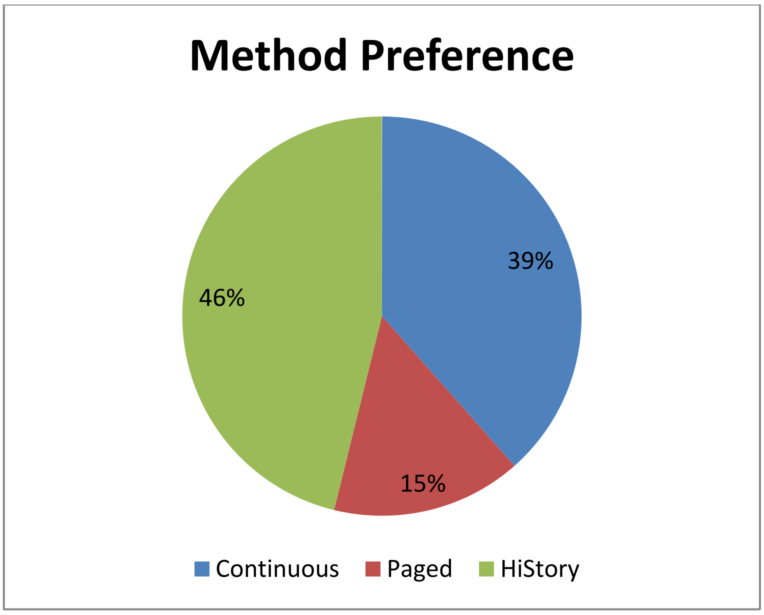


Figure 6.25 – User preference between Continuous, Paged and HiStory methods.

Above, in figure 6.25, is illustrated the users' opinion of which method, of the three they were presented with, they preferred. Almost half the users stated that they preferred the HiStory method as it allowed them much greater flexibility and accuracy. In fact, the results of the second question, which queried users on whether they believed that the HiStory method, with its ability to 'dive' to a frame-by-frame granularity level, had higher accuracy, are not illustrated in this report. This is become all the users (100%) unanimously stated that they believed that HiStory method allowed for the most accuracy in retrieving specific items from a video. As with the previous experiment, Continuous scrolling remained a more popular choice than Paged scrolling, with the same objections about control and freedom voiced.

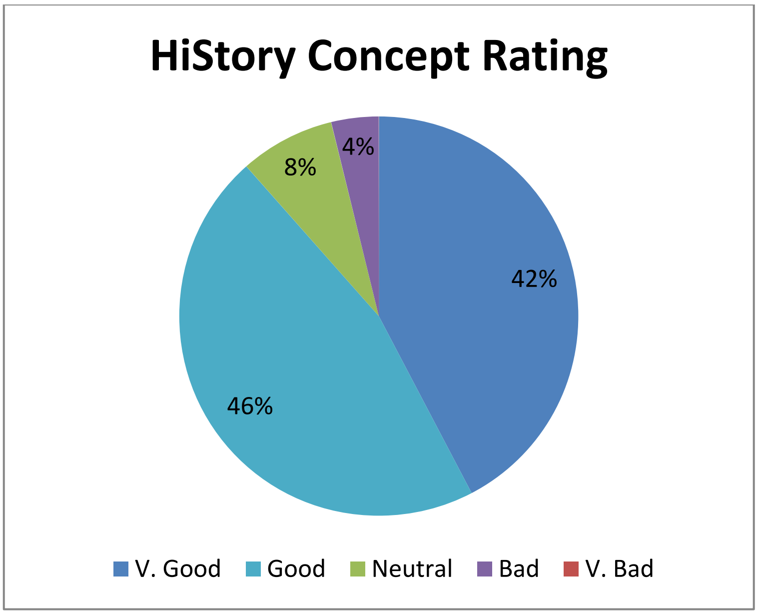


Figure 6.26 – User rating for the HiStory concept.

The last question, whose results are illustrated above in Figure 6.24, requested that the users rate the concept of the HiStory method on a five point scale, ranging from 'Very Bad' to 'Very Good'. The large majority of users had a positive opinion of the concept with more than 85% of the test subjects splitting their rating between 'Very Good' and 'Good'. The few detractors stated that they preferred the straightforwardness of the scrolling methods to the dynamic nature of the HiStory method.

To summarise the findings of the second experiment, the introduction of displayed video segments instead of text descriptions for the Known Item Search tasks applied a penalty to all methods in terms of the time required to complete each task. However, this overhead was deemed acceptable especially when taking into account the complete lack of ambiguity regarding the item that was to be found for each task. Users praised the change stating that while with the text questions they were essentially guessing as to what they were looking for and felt insecure as a result, the video segments gave them greater assurance. The HiStory method displayed longer times for completion, compared to the scrolling methods, but this can be directly attributed to the technical time delays imposed by the hardware limited prototype used for the experiment. Between each interaction of the user with the interface there occurred some delay between each granularity reconfiguration of the grid. Some users voiced concerns over this but they correctly assumed that such an issue would not exist in a production version.

On the matter of accuracy, the Paged method suffered a slightly lower rate than the Continuous method, contradictory to the results of the first experiment but the difference was insignificant. The HiStory method was indisputably more accurate with very few mistakes made by the users. Indeed given the lifting of any time constraints self-imposed by the users, the success rate could raise even more as the users had the option to reach a very fine granularity.

Most telling of all are the user opinions regarding their preference. Despite none of the users having dealt with a similar concept of a video browsing interface before, they were quick to adopt its mechanics and functionality and formulate effective browsing strategies in order to rapidly complete tasks. Indeed it was observed that the users were getting more proficient with the completion of every test, showing that while the method had a learning curve; it was very short and led to very effective results.

6.4 HiStory

Since HiStory, in contrast to the scrolling methods, was a complete prototype, with its own alternative search paradigm, it is further evaluated in terms of its perceived success in terms of real world demands and human computer interaction (HCI) principles.

When measured up against the guidelines for events such as the Video Browser showdown, that stipulates that users must find a 20 second segment, shown to them on a separate screen, within a time allotment of approximately 2 minutes, the HiStory method shows great promise. Especially when considering that there was no prior processing or analysis of the videos, as is the case with Content Analysis based methods and the fact that the users were utilising the interface on a mobile device.

Since this work is based firmly in video retrieval using human processing abilities, it is interesting to be evaluated in terms of measuring up to various human computer interaction principles, guidelines and standards (Norman, 1990; Nielsen, 1993; Dix et al., 2003; Shneiderman and Plaisant, 2004). The

HiStory interface and interaction can be examined along the high level principles of whether it was useful, usable and acceptable to users (Dix et al., 2003). The main HiStory interface elements were the storyboard grid and the vertical scroll bar, while the main interaction relied upon the paradigm of going from larger to finer granularity by zooming in and out of the storyboard with the user action of selecting a thumbnail by tapping on it. The users then needed to understand where in the video they were (by referencing the vertical scrollbar) and how the thumbnails reconfigured (the new storyboard grid is reconfigured around the selected 'anchor' 'thumbnail).

Usefulness here can be taken as meaning whether the interface elements and the interaction were useful for the task at hand, in this case finding known items. Also along with the principle of usefulness, it is important to see the principles of effectiveness (did users achieve what they set out to do) and efficiency (did users achieve what they wanted without undue effort and resources)[23].

Our results show that users did find HiStory useful for the task, particularly the hierarchical grid, which gave an overview of the whole video. The logged results for accuracy show that it was effective, but in terms of efficiency, due to the aforementioned problems with the processing time required for actually carrying out the reconfiguration, there was some overhead on the time-to-complete (users had to wait for the thumbnails to reconfigure).

In terms of usability, the principles ask for such things as whether the interface and the interaction was easy to learn (learnability); easy to remember (memorability); whether there is visible feedback on user interactions (visibility); and whether there is tolerance for error (feedback and tolerance); meaning whether users can go back and undo erroneous actions. HiStory was introduced to users with a short tutorial. The users were quick to understand the operating concept, which was familiar to them from map applications. Thus the interface and interaction made use of the principle of familiarity and was also consistent with user expectations. The use of the vertical scrollbar provided visible affordance that they found easy to learn and hence to remember. As mentioned above in the qualitative results, the users got more proficient in the use of the interface and made comments to that effect. They voluntarily stated that they felt in control and confident that they would be able to find exactly what they were looking for.

Finally, in terms of acceptability, this equates to whether the users found it pleasant to use; whether it removed anxiety regarding task completion; and perhaps even increased pleasurable anticipation, as in wanting to solve the task as though it were a puzzle or a game. As stated above, HiStory was accepted by users; they enjoyed the increased accuracy the prototype gave; and with it the sense of security that they would find the answers they sought. On the negative side, the users did experience some irritation caused by the limitations in the processing power, which meant they had to pause while the thumbnails reconfigured, but as mentioned before, this was understood to be a temporary problem. One older user also expressed that the ability to be able to enlarge a thumbnail, by double tapping it, (or by using "two finger scissors motion" or some other haptic gesture), might be useful for users who want to momentarily inspect a particular thumbnail at a larger size, before deciding whether to change the granularity. This user felt that this might help those users who felt their eyesight problems were handicapping them. However, such a feature would have to be weighed up against the overhead that increasing the functionality of the interface brings. Furthermore our results that showed that users, despite feeling uncomfortable sometimes do

[23] http://www.iso.org/iso/catalogue_detail.htm?csnumber=16883

actually make accurate choices, a phenomenon prevalent in the experimentation project as well. Finally, in regard to user satisfaction, for some users, and in particular, the youngest test subject (aged 12 years) the HiStory interface felt a bit like a game to them, and they wanted to know how well they scored! This would indicate that the merits of Gamification (Deterding et al., 2011) could be leveraged for video retrieval tasks.

In the next and final chapter, the overall results are summarised, and further experiments are suggested along with future avenues of research.

7 Conclusions and Discussion

In this chapter a summary of the findings of the two experiments will be presented, followed by a report of possible further experimentation that could be conducted in the same area. Finally future avenues of research are discussed regarding the next possible steps in the field.

7.1 Summary of conclusions

In short, following the analysis of the results of the two experiments, a number of grounded conclusions can be made about the suitability and effectiveness of the tested interfaces for video retrieval tasks. Generally, a number of positive findings were made, further validating the conclusions made in previous research endeavours. Specifically:

- The nuances of scrolling interaction were examined with a number of results confirmed.
  - Users prefer the flexibility and control of Continuous scrolling over Paged scrolling.
  - Paged scrolling, despite appearing slower, is not significantly slower than Continuous scrolling.
  - Paged scrolling allows for significantly higher accuracy.
  - Scrolling Grid interfaces do not negatively impact search tasks as initially feared.
- Alternative visualisations and interface concepts, were shown to be effective:
  - The HiStory Interface was highly rated by the test subjects as an effective concept.
  - Users also considered HiStory as the most desirable interface when given a choice.
  - Users unanimously believed that HiStory was the best option when accuracy in completing tasks is of paramount importance.
  - Quantitatively HiStory achieved much higher accuracy rates than the other methods, with very few errors made by users.
- A few general findings regarding video retrieval research were also made:
  - Textual descriptions for Known-Item-Search create numerous issues regarding ambiguity and interpretation.
  - Video segments as descriptions of the requested item of a search task were welcomed by users due to their unambiguity and immediacy.

The above conclusions can act as a foundation for further experimentation and research. Nevertheless there are some aspects of the project which would merit from further investigation, as discussed in the next section.

7.2 Further Experimentation

On the matter of further examining the intricacies of scrolling with relation to video retrieval tasks, a number of investigations could be made. Scrolling remains very important in this context as in the technical domain of mobile devices, accompanying limited screen size means that more often than not, the content the user wishes to view cannot be effectively presented on a single screen. Refinements to the physical haptic interaction can be investigated, as well as combinations of scrolling methods. Additionally, the effects of dynamic thumbnails combined with complex interfaces have not been explored. There is a strong possibility that a mixture of dynamic thumbnails and scrolling interfaces or alternative interfaces like HiStory, which is considered below, might unlock yet more flexible and powerful capabilities for Video Retrieval applications on mobile devices.

For the purposes of the HiStory interface, further experimentation could be made with modifications to the presentation of the content. An experiment could be designed on the interaction methods as well as the visual feedback and layout of the interface. For example, an interesting experiment would be to investigate the optimal reconfiguration strategy of the grid; is it more beneficial for the selected thumbnail to stay in place, as is the case with the current iteration, or is it better for the selected thumbnails to be moved to the centre of the screen with the grid reconfiguring itself around it with an equal weight before and after?

These are parallel questions that would provide very useful insight and information for future endeavours.

7.3 Future Work

The domain of video retrieval is enormous in its scope and variety and focusing specifically on mobile solutions does not in any way limit that scope, quite on the contrary it expands it to a great degree when one considers the possibilities. The above findings and conclusions can be further refined with more experiments but they can also serve as the base for a number of parallel and tangential research avenues. It is possible to continue along the lines of traditional two-dimensional interfaces, incorporating all the refinements garnered from research investigating the best way to optimise and humanise the consumption of content in mobile devices but it is also possible to combine all the above with the boundless and exciting research being conducted on the intricacies of three-dimensional content on mobile devices. As these devices become more technically powerful and diverse, incorporating features such as advances haptic controls, 3D visualisations and augmented reality capabilities, the options and possibilities for exploration are countless.

8 References

Ahlstrom, D., Schoeffmann, K., 2012. A Visual Search User Study on the Influences of Aspect Ratio Distortion of Preview Thumbnails, in: Proceedings of the 2012 IEEE International Conference on Multimedia and Expo Workshops. IEEE Computing Society.

Beaudouin-Lafon, M., 2000. Instrumental interaction: an interaction model for designing post-WIMP user interfaces, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '00. ACM, New York, NY, USA, pp. 446–453.

Cao, J., Zhang, Y.D., Pang, L., Feng, B.L., Li, J.T., 2010. Known-Item Search by MCG-ICT-CAS.

Christel, M., 2007. Examining User Interactions with Video Retrieval Systems. Computer Science Department.

Christel, M.G., 2006. Evaluation and User Studies with Respect to Video Summarization and Browsing, in: IN MULTIMEDIA CONTENT ANALYSIS, MANAGEMENT AND RETRIEVAL 2006, PART OF THE IS&T/SPIE SYMPOSIUM ON ELECTRONIC IMAGING.

Christel, M.G., Lin, W.-H., Maher, B., 2008. Evaluating audio skimming and frame rate acceleration for summarizing BBC rushes. ACM Press, p. 407.

Del Fabro, M., Böszörmenyi, L., 2012. AAU video browser: non-sequential hierarchical video browsing without content analysis, in: Proceedings of the 18th International Conference on Advances in Multimedia Modeling, MMM'12. Springer-Verlag, Berlin, Heidelberg, pp. 639–641.

Deterding, S., Dixon, D., Khaled, R., Nacke, L., 2011. From game design elements to gamefulness: defining "gamification", in: Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments, MindTrek '11. ACM, New York, NY, USA, pp. 9–15.

Divakaran, A., Forlines, C., Lanning, T., Shipman, S., Wittenburg, K., 2005. Augmenting fast-forward and rewind for personal digital video recorders, in: Consumer Electronics, 2005. ICCE. 2005 Digest of Technical Papers. International Conference On. pp. 43–44.

Dix, A., Finlay, J.E., Abowd, G.D., Beale, R., 2003. Human-Computer Interaction, 3rd ed. Prentice Hall.

Dragicevic, P., Ramos, G., Bibliowitcz, J., Nowrouzezahrai, D., Balakrishnan, R., Singh, K., 2008. Video browsing by direct manipulation, in: Proceedings of the Twenty-sixth Annual SIGCHI Conference on Human Factors in Computing Systems. pp. 237–246.

Economides, A.A., Grousopoulou, A., 2008. Use of mobile phones by male and female Greek students. Int. J. Mob. Commun. 6, 729–749.

Fan, X., Xie, X., Zhou, H.Q., Ma, W.Y., 2003. Looking into video frames on small displays, in: Proceedings of the Eleventh ACM International Conference on Multimedia. pp. 247–250.

Geetha, P., Narayanan, V., 2008. A Survey of Content-Based Video Retrieval. Journal of Computer Science 4, 474–486.

Hu, W., Xie, N., Li, L., Zeng, X., Maybank, S., 2011. A Survey on Visual Content-Based Video Indexing and Retrieval. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 41, 797–819.

Hürst, W., Bilyalov, T., 2010. Dynamic versus static peephole navigation of VR panoramas on handheld devices, in: Proceedings of the 9th International Conference on Mobile and Ubiquitous Multimedia, MUM '10. ACM, New York, NY, USA, pp. 25:1–25:8.

Hürst, W., Darzentas, D., 2012. Quantity versus quality: the role of layout and interaction complexity in thumbnail-based video retrieval interfaces, in: Proceedings of the 2nd ACM International Conference on Multimedia Retrieval, ICMR '12. ACM, New York, NY, USA, pp. 45:1–45:8.

Hürst, W., Snoek, C.G.M., Spoel, W.-J., Tomin, M., 2010. Keep moving!: revisiting thumbnails for mobile video retrieval, in: Proceedings of the International Conference on Multimedia, MM '10. ACM, New York, NY, USA, pp. 963–966.

Hürst, W., Snoek, C.G.M., Spoel, W.-J., Tomin, M., 2011. Size matters! how thumbnail number, size, and motion influence mobile video retrieval, in: Proceedings of the 17th International

Conference on Advances in Multimedia Modeling - Volume Part II, MMM'11. Springer-Verlag, Berlin, Heidelberg, pp. 230–240.
Jammalamadaka, N., Zisserman, A., Eichner, M., Ferrari, V., Jawahar, C.V., 2012. Video retrieval by mimicking poses, in: Proceedings of the 2nd ACM International Conference on Multimedia Retrieval, ICMR '12. ACM, New York, NY, USA, pp. 34:1–34:8.
Jansen, M., Heeren, W., van Dijk, B., 2008. Videotrees: Improving video surrogate presentation using hierarchy, in: Content-Based Multimedia Indexing, 2008. CBMI 2008. International Workshop On. pp. 560–567.
Kanagavalli, R., Duraiswamy, K., 2012. A Study on Techniques used in Digital Video for Shot Segmentation and Content Based Video Retrieval. European Journal of Scientific Research 69, 370–380.
Mehra, S., Werkhoven, P., Worring, M., 2006. Navigating on handheld displays: Dynamic versus static peephole navigation. ACM Trans. Comput.-Hum. Interact. 13, 448–457.
Nielsen, J., 1993. Usability Engineering, 1st ed. Morgan Kaufmann.
Norman, D., 1990. The Design of Everyday Things. Doubleday Business.
O'Hara, K., Mitchell, A.S., Vorbau, A., 2007. Consuming video on mobile devices, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 857–866.
Odlyzko, A., 2001. Content is not king. First Monday 6.
Oliva, A., Torralba, A., 2007. The role of context in object recognition. Trends Cogn. Sci. (Regul. Ed.) 11, 520–527.
Pongnumkul, S., Wang, J., Ramos, G., Cohen, M., 2010. Content-aware dynamic timeline for video browsing, in: Proceedings of the 23nd Annual ACM Symposium on User Interface Software and Technology. pp. 139–142.
Schoeffmann, K., Hopfgartner, F., Marques, O., Boeszoermenyi, L., Jose, J.M., 2010. Video browsing interfaces and applications: a review. SPIE Reviews 1, 018004–018004–35.
Shneiderman, B., Plaisant, C., 2004. Designing the User Interface: Strategies for Effective Human-Computer Interaction, 4th ed. Addison Wesley.
Smeaton, A.F., Over, P., Kraaij, W., 2006. Evaluation campaigns and TRECVid. ACM Press, p. 321.
Snoek, C.G.M., Worring, M., 2009. Concept-Based Video Retrieval. Found. Trends Inf. Retr. 2, 215–322.
Taneja, H., Webster, J.G., Malthouse, E.C., Ksiazek, T.B., 2012. Media consumption across platforms: Identifying user-defined repertoires. New Media Society 14, 951–968.
Torralba, A., Fergus, R., Freeman, W.T., 2008. 80 Million Tiny Images: A Large Data Set for Nonparametric Object and Scene Recognition. IEEE Trans. Pattern Anal. Mach. Intell. 30, 1958–1970.

9 Appendix

9.1 Quantity vs. Quality

Quantity versus Quality – The Role of Layout and Interaction Complexity in Thumbnail-based Video Retrieval Interfaces

Wolfgang Hürst
Department of Information and Computing Sciences
Utrecht University, Princetonplein 5,
3584 CC Utrecht, The Netherlands
huerst@uu.nl

Dimitri Darzentas
Department of Information and Computing Sciences
Utrecht University, Princetonplein 5
3584 CC Utrecht, The Netherlands
d.p.darzentas@students.uu.nl

ABSTRACT
In this paper we present our findings from an experiment designed to test the effectiveness of complex, thumbnail based layouts and the interaction methods required, in video retrieval scenarios for handheld devices such as smartphones. Our evaluation explores the relationship between the number of thumbnails (Quantity) visible on screen and their discernible detail (Quality) with regards to the related necessary amount of interaction. The results indicate that such layouts achieve a very high rate of accuracy and speed but that the nature of the interaction is of critical importance for the success of the system.

Categories and Subject Descriptors
H.5.2 [**Information Interfaces and Presentation**]: User Interfaces – *Evaluation/methodology, graphical user interfaces (GUI), screen design, style guides, user centered design*

General Terms
Design, Experimentation, Human Factors, Theory

Keywords
Mobile Video, Video Retrieval, Thumbnails, Mobile Cognition

1. INTRODUCTION
Thumbnail images have become ubiquitous in our daily digital life as representations of, for example, larger photos or video clips. The old adage: "a picture is worth a thousand words" is well and truly verified in this regard. Ever since the thumbnail first replaced the icon as a visual preview representation of a digital image (or video) file, it has become accepted as the de-facto standard practice. The ability of humans to perceive and recognize features in these small pictures has made them an important tool in non-automated video browsing [10]. A simple small image can convey large amounts of information in an instant. As a fast and effective way of visually browsing (with no other assistance such as metadata) through image or video items, thumbnails are undisputedly one of the most effective methods [16] hence their widespread adoption in all forms of digital galleries, both on desktop systems and on mobile handheld platforms.

With the latter becoming the most ubiquitous connection with the digital world, especially in form of smartphones, the question arises if and how we can create efficient and effective video search and retrieval interfaces given the unique form factor and small screen sizes of such devices. On desktop systems, interfaces for video retrieval have been heavily reliant on the presence of metadata accompanying the video data, and the design of these interfaces reflects this [5]. Results from video queries are most often returned as a list of videos, presenting each with its relevant metadata and accompanied by a single descriptive thumbnail. Other approaches, such as the popular CrossBrowser implementation [15] focus on the presentation of multiple thumbnails extracted from individual videos and representing their respective content. Individual videos are often represented by a temporarily sorted sequence of extracted thumbnails – commonly referred to as storyboards [1, 6].

While usability and usefulness of such interface designs have been verified by many scientific studies in desktop system environments, their applicability for popular smartphones seems questionable due to their much smaller screens – typically around 3.5 to 4 inch diagonal versus 17 inch or larger diagonal on desktop systems – and different interaction modes – typically touch screen interaction versus mouse and keyboard on desktop PCs. Motivated by the relevance and importance of thumbnails for video retrieval interfaces, [7, 8] present a series of experiments addressing the first issue, i.e. what are optimal sizes and type (static or dynamic) of a thumbnail on a mobile platform? Their encouraging findings showed the use of thumbnails is actually a very effective means of data retrieval on mobile platforms, despite their limited screen size. However, related tests were limited to individual thumbnails shown in isolation [7] and horizontally presented strips [8]. It is yet unclear if and how they generalize to more complex layouts as they are required in realistic video retrieval tasks. In addition, the experiments presented in [8] suggest that the need for interaction might hinder users from utilizing thumbnails in the optimum way that was identified in [7]. Hence, using these findings as a starting point, we apply their conclusions to more elaborate and real-world scenarios to test whether the results can be transferred from single solitary thumbnails and small strips, to more complicated layouts, in order to determine the next best step for the effective design of mobile video retrieval interfaces. In particular we were interested in the effectiveness of instance recognition tasks [13] when attempted in combination with these layouts.

Given the high popularity and relevance of storyboards for video retrieval, we decided to focus on a grid (or matrix) layout of thumbnails in this paper (as illustrated in Figure 1). This layout is applied to a group of thumbnails extracted from a single video clip in a time ordered sequence arranged on a grid. This conventional layout is also familiar from reading text left-to-right, bottom-to-top (at least in the case of western texts). Its use has also been endorsed by TRECVID [6, 13]. Focusing on search within a single video file was motivated by the related retrieval task, which will be discussed later.

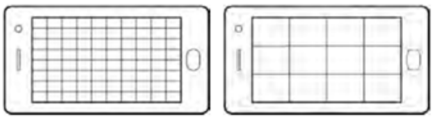


Figure 1. Storyboards-like grid-based thumbnail layouts

The parameters of a grid of thumbnails are essentially the size of the thumbnails themselves and immediately tied to that variable is their quantity. The smaller the size of the thumbnails, the more of them can fit on the grid. This leads to a ratio of size to quantity. What this means is that we can either have a lot of small images, and therefore greater granularity of information, or fewer larger images, which while more detailed individually, may not give enough information overview about the content of the video clip.

One way to break this stalemate is to lift the limit of the number of thumbnails by making the grid scrollable; this has the immediate benefit of allowing as many thumbnails as deemed necessary and therefore maintaining a finer granularity without necessitating smaller thumbnail sizes. The downside is that not all the thumbnails are displayed at the same time and an additional layer of physical and cognitive interaction is necessary. Figure 2 shows one possibility of such an arrangement and the one we chose to pursue. In our experiment we utilized "continuous scrolling" identical to the methods used by desktop applications instead of the "discrete scrolling" methods that enable a page-wise navigation and are often employed by mobile devices to, for example, switch between different "screens" of icons and digital book readers following the appropriate page metaphor.

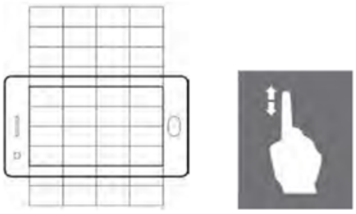


Figure 2. Scrollable grid layout for browsing larger quantities of thumbnails

This paper presents the findings of an experiment we set up to study the effects of the above variables (size, quantity, grid type) on the effectiveness of instance recognition tasks. We also evaluated the related interaction to determine whether such a setup would be appropriate and useful to real-world users and applications, such as video browsing for entertainment. By testing a 'bare-bones' static grid, and then systematically adding features and functionality, in this case the ability to scroll through more content, we plan to roadmap the creation of an effective thumbnail based layout for video retrieval on mobile handheld devices.

It should be noted that we were not immediately aiming to determine the optimal technical parameters for the thumbnails of a grid based on retrieval performance but to understand how well the layout facilitates effective video browsing and video retrieval tasks. In the following sections, we shall discuss the related work (section 2), elaborate further on the details and methods of the experiment (section 3), provide an analysis of the results (section 4), the conclusions, and related work (section 5).

2. RELATED WORK

Generally, we can assume that providing as much ancillary information about the content of a video as possible with a query result will lead to better search performances. This intuitive statement has resulted in many effective but rather complex interfaces which take full advantage of the large screen sizes (see [11] for a related overview). Considering video retrieval interfaces for contemporary mobile platforms, such as smartphones, which are fast becoming commonplace and are a good indication of current interaction technology, the situation is rather more complicated. Small screen sizes severely limit the expansiveness, complexity and therefore usefulness of advanced video retrieval devices. Indeed one may argue that these limitations could severely impact the effectiveness of the thumbnail representation as one of the most effective tools in the video retrieval arsenal.

Thumbnails have been accepted as the standard base on which nearly all video retrieval interfaces are built. Even the most barebones, text-query, interfaces strive to provide a thumbnail of the retrieved video file. Other thumbnail-based interfaces often use metaphors like filmstrips [3], Manga-like collages [2, 4], and the aforementioned storyboards [1] to represent the content of a video to the searcher. Thumbnails in all their forms are unquestionably effective on desktop systems for video retrieval. Motivated by the instinctive concerns about their size and practicality on small mobile screens [7] presented an evaluation where users were presented with different sizes of single solitary thumbnails in relation to common video retrieval tasks. The experiments also included the type of the thumbnails as a factor, examining whether static or dynamic thumbnails were more effective. The conclusions showed that rather small images can still be reliably used for search and target classification and if the thumbnails are dynamic even much smaller still lead to a reasonable retrieval performance. In particular, thumbnail sizes as small as 80 and 60 pixels for static and dynamic ones, respectively, have been determined as a viable conveyance of information on a mobile platform.

However, in a direct follow up experiment, [8] identified that these results are only valid in the case of a solitary thumbnail representation, but do not apply for designs in which single linear stripes of five to nine thumbnails are shown. In particular, the reported experiments suggest an optimum thumbnail size of 110 pixels, which still seems much smaller than originally anticipated. Maybe even more important, the studies revealed that users are not taking advantage of the option to play individual thumbs within the filmstrip as dynamic thumbnails – despite their superiority for search tasks that was demonstrated by [7]. Reasons for this might be a combination of a higher cognitive load resulting from the additional need to interact as well as the fact that more information about a video's content was shown at a time (thumbnails in the test have been extracted from a single video file).

Although providing encouraging information for the design of more advanced interfaces for video retrieval via mobile platforms, the results presented in [7, 8] leave several important questions unanswered. In particular: If optimum thumbnail sizes increase when we switch from classifying solitary thumbnails to film strip representations, will storyboard-like matrix representations such as depicted in Figure 1 result in a need for even larger thumbnails? And if so, how does the resulting need for interaction (cf. Fig. 2) influence retrieval performance and subjective search experience – especially considering the negative experience with interaction reported in [8]?

In order to further investigate the effects that more complex layouts and arrangements would have on the users and their ability to complete certain video browsing tasks, we present an experiment evaluating the parameters thumbnail size in a grid-based layouts as shown in Figure 1 under consideration of potential issues resulting from scrolling interactions as illustrated in Figure 2. We define interaction here as both retrieval performance and as how well the interface facilitates effective browsing [5]. For our tests, we decided to restrict them to static thumbnail representations, based on the same arguments as [8], i.e. that simultaneous playback of various images as dynamic thumbnails would further distract users and most likely have an impact on performance and overall effectiveness when searching for information. The major parameters to explore in our test are thumbnail size and the related ability for users to find content represented by them, and the role interaction has on the effectiveness of the interface. Because these are independent of the actual source of the data represented, we only tested search in single video files, but expect our results to generalize to search in multiple videos, where, for example, each row in the matrix represents the time-ordered sequence of thumbnails extracted from single videos.

3. EXPERIMENT

3.1 Interface Rationale & Design

To begin investigating the effects of presenting multiple thumbnails simultaneously on video browsing, a number of different layouts were considered, including grids, linear scrollable strips, three dimensional perspective representations and other less conventional arrangements. As a first step we settled on a grid layout reminiscent of a traditional storyboard, a layout that is intuitive and effectively communicates it purpose. It caters to user expectations as it follows the conventional left to right and line by line reading styles prevalent in all (western) media, such as printed works, comics, etc.

More importantly, the grid layout also allows for the maximum quantity of thumbnails possible in a two dimensional interface without any overlap. Specifically, for the static grids, the number of possible thumbnails that would fit on the screen of the device was determined by the dimensions of the thumbnails. For the smallest size thumbnails (80 pixels) the number of thumbnails was 10x8 = 80 images. This amount decreased progressively as the thumbnail sizes increased. For the largest (200 pixels) images, only 12 images could fit on-screen. The illustrations shown in Figure 1 illustrate the exactly the largest (left) and lowest (right) number of thumbnails used in the experiment.

In the case of the scrolling grid (cf. Figure 2), while there was no upper limit on the total amount of thumbnails, the quantity that could be displayed on screen at any given time was still held by the same restrictions. For consistency, across all thumbnail sizes we limited the total amount of thumbnails so that the total "scrolling distance" would be about two and a half single screen grids.

3.2 Experiment Design

In order to determine the effectiveness of the grid based layout, we adapted the experimentation methodology of [7] and set up an experiment to progressively test the ability of the user to complete instance detection tasks. Grid layouts were evaluated by the subjects in decreasing number of thumbnails (i.e. increasing thumbnail sizes) in order to benefit from learning effects and better identify subjective user perception of prefered thumbnail sizes (cf. [7]). A series of questions/tasks would be asked of the user with the parameters of the layout changing every few questions.

The layout parameters under examination are the size of the thumbnails, their number and whether the grid is static (limited to the current screen) or scrollable (bigger than the current screen and therefore scrolling is necessary in order to view all the thumbnails). An emergent parameter that is under scrutiny is the ratio of the thumbnail size to the thumbnail quantity. Determining the ratio that corresponds to the optimum balance between information depth and information breadth is one of the key issues in a related interface design (quantity – quality).

The profiles of the users are of course also an important factor. Therefore the experiment was designed to gather and catalogue the information from the users both prior to and after the experiment was conducted. Prior to the test, the users were queried on typical relevant factors such as gender and age and also on their attitude and experience towards technology, mobile devices and the use of technology for entertainment. After the experiment, the users were queried again, this time on their opinion of the interaction. Importantly they were specifically not asked for their opinion on the thumbnail's parameters, e.g. "which size did you find more convenient?" because it seems obvious that users would most likely agree that larger thumbnail sizes are better suited to identify the contained content. Instead, they were asked questions regarding the interaction method itself, e.g. "How intuitive/natural did you find the interaction?" because the need for more interaction activity is a natural trade-off between thumbnail size and the resulting number displayed at a time.

3.3 Implementation

To perform the experiment we settled on the creation of a single mobile application on the Android platform. We implemented a portable mobile application that would serve as a distributable test to gather data. The application encompassed the entirety of the experiment, including questionnaires, feedback mechanisms and the performable tasks themselves. It was designed in such a way so that it would serve as a modular, expandable and flexible test system for the possibility of future adapted tests or different experiments. It was also designed to be autonomous, gathering and dispatching the data to the researchers automatically to provide the additional option large scale distribution for gathering with large numbers of users should the need arise. Figures 3, 4 and 5 show the end result of the implementation. Thumbnails in the figures have been replaced for this publication with non-copyrighted material, but reflect the type and style of data that was used in the tests.

The source video data that was used to extract the thumbnails was a collection of various episodes from contemporary television series. The wide variety and today's prevalence of television media provided a consistent and convenient base of data that directly relates to real world use cases. The thumbnails for each task were extracted from different video clips. The choice of which thumbnails to extract from each video was based on the requirement that they should cover the whole length of the clip and be equidistant in frames.

Figure 3. Screenshot of the layout used in the experiment (with largest number of thumbnails)

Figure 4. Photo of the layout used in the experiment (with the smallest number of thumbnails)

Figure 5. Photo of the experiment procedure (with a question being asked via a text label popup)

The tests that we report on in this paper were all administered and supervised. This was done so that in addition to the quantitative data gathered by the application, we could also obtain some qualitative data. The quantitative data was in the form of transaction form analysis and task time, while the qualitative data arose from the observed user reactions and voluntary think-aloud opinions from the participants.

The overall testing process begins with a short questionnaire establishing details such as age, gender and experience with mobile technology. This is followed by a short onscreen tutorial that familiarizes the users with the imminent testing process by running them through some non-interactive mock questions with popups explaining the functionality and options. The users are then presented with a sequence of 24 questions. Each question is an instance recognition task (cf. below). The users peruse the thumbnails and can highlight any number of them that they consider as applicable. The users may also provide a negative answer when they are sure that the requested instance is not contained in the video. In the case where the users feel that they cannot commit to an answer, i.e. they are unsure, there is a third option available to them – i.e. a button indicating “unsure”.

Tasks were motivated and designed in accordance to the 2011 TRECVID “Instance Search” category. Examples for typical search motivations in relation to the used TV shows include situations such as: “I want to re-watch that funny scene from the show that I saw some time ago, where the one character was wearing this funny hat,” or “Is this the episode in which this character was wearing that funny hat?” However, since the participants were unfamiliar with the actual content of the data used in the tests, we phrased the questions neutrally. For practical reasons, they were also formulated as yes/no questions that could be answered by “I found it/It doesn’t exist/ I cant tell”, resulting in tasks such as “Is there a man wearing a read baseball hat in the video?” or “Is there a fire truck in the video?”

Of the 24 questions (or tasks), the first 12 are on static thumbnail grids and the rest are on scrollable thumbnail grids. The size of the thumbnails begins at 80 pixels, equivalent to 8.7mm in width, and increases in steps every 3 questions, finally reaching a size of 200 pixels, equivalent to 21.5mm. In the case of static grids this means that a single screen on the first 3 tests contains 80 thumbnails, this number finally shrinks to 12 in the last three tests. In the case of the scrollable grids, regardless of thumbnails size, the user is able to scroll downwards approximately two and a half screens.

The sequence of tasks was the same for all the participants so as to clearly indicate any trends and inconsistencies. We chose to have the progression of the tasks (small to large thumbnails) the same for all the test subjects, so that any learning curve can be identified and taken into account. Upon answering the final question, the user is presented with a closing questionnaire that queries them solely on the interaction experience.

4. DATA ANALYSIS AND RESULTS

The data collected from the supervised tests, when collated and analysed, presented very encouraging and interesting results. As the main distinction between the tasks was the type of grid (Static or Scrollable) it is beneficial for ease of comprehension to separately assess the results for each case.

Before moving on to this separate analysis, a few observations must be made on the users. We chose to collect a sample that was as balanced as possible, gender wise, with 14 male and 10 female participants. Tables 1 and 2 show their answers to questions regarding their experience with technology and more specifically mobile platforms.

Table 1. Relationship with Technology

Positive - Everyday use	8
Neutral - Circumstantial use	11
Negative - Avoided if possible	5

Table 2. Relationship with mobile platforms (smartphones)

Positive - Part of everyday life	11
Neutral - Dictated by circumstances	4
Negative - Avoided if possible	9

The tables illustrate the fact that many of the participants did not consider themselves experienced users, what would be colloquially known as "power users" and that some of them even have negative relationships with technology. What is of great interest though is that all 24 users, when asked the question about whether they used mobile technology for entertainment, answered without any observed hesitation, yes. This observation also holds across all age groups as the sample that was tested was distributed over a large range of ages, as can be seen in Table 3.

Table 3. Numbers of participants per age group (top row)

21-30	31-40	41-50	51-60
8	9	4	3

As previously mentioned the use of contemporary technology is not limited to a small "technological elite" but is a phenomenon widespread not only across multiple knowledge levels but also across age groups and generations of users. Indeed the fastest growing demographic for the use mobile technology is the baby boomer generation (55-64 year olds) [14]. Therefore we purposefully sought out users that would belong to the more advanced age groups, who have had to deal with the advances of technology over the course of their lives in favour of younger users who have become acclimatized to today's technological metaphors.

In the following sections, we shall have a detailed look at the individual results for the static and the scrolling grids and their comparison. Each question asked of the users is referred to as a task, with 24 tasks in all, 12 for the static layout and 12 for the scrolling layout. Each set of 12 is divided into 4 groups which share the same size parameters. Finally we analyse the user feedback on the interaction experience.

4.1 Static grid

For the case of the static grid it was apparent early on in the analysis that the levels of accuracy were high and very consistent. Indeed the average rate of success across all 12 tasks is 85% with half the users scoring above 87.5%. Within the static tasks the main variable parameter is the size and quantity of thumbnails. With the first 3 of the 12 static tasks featuring very small (80 pixels wide) thumbnails, the success rate began at a moderate 66% for the first task and 70% for the second, rising immediately for the third one as illustrated in Figure 6, which also shows the results for the three tasks of each other thumbnail size (100px, 133px, 200px, respectively).

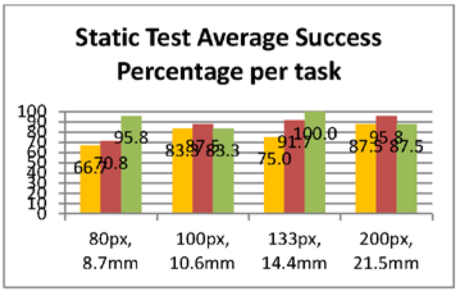


Figure 6. Success rates (static test, detailed results)

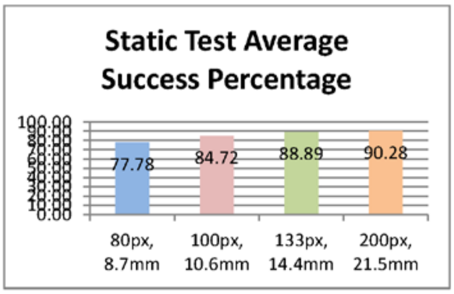


Figure 7. Success rates (static test, averages)

Figure 7 illustrates how, despite the lower performance of the first task start, the success rates rise fast, and continue rising throughout. Each bar represents the average success rate for each thumbnail size task group, 3 tasks for each group for a total of 12 static grid tasks shown in Figure 6. A plausible explanation, the low score on the first task is a necessary acclimatisation period by the users. As mentioned above in section 3.3 we anticipated a learning curve.

Obviously, the improvement in performance was to be expected as larger thumbnail sizes should result in a better visibility and thus better classification performance. It is surprising though that in contrast to the relatively large increase of about 7% from 80px to 100px thumbnail sizes, further improvements are less dramatic, i.e. only about 4% and 2% from 100px to 133px to 200px. This is particularly noteworthy with respect to the results presented in [7, 8] which show a strong increase in optimum thumbnail size when moving from solitary thumbnail representations to film strips. Apparently, this does not seem to be the case when moving from strips to full, matrix-like storyboards.

However, besides the high and consistent level of accuracy, also of some note was the drastic reduction in time required by the users to make a choice as the size of the thumbnails increased. Figure 8 illustrates this result with each bar again representing a thumbnail size task group. The overall average time to complete a task across all the static questions was 25 seconds.

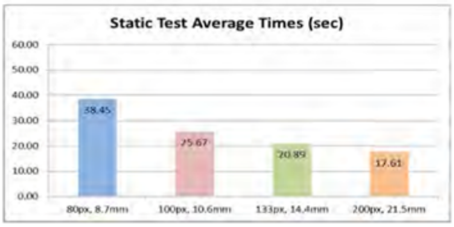


Figure 8. Time to solve the tasks (static test, averages)

The decrease in the time to answer might have been expected as the increasing thumbnail sizes naturally lead to smaller quantities (less individual thumbnails to look at) of thumbnails. However as the thumbnail sizes increase, so does the amount of visual detail that must be processed (bigger, more detailed images). It would appear that it is faster and easier to analyse fewer, more detailed images, than it is to scan through several less detailed ones.

As the experiment was designed so that some of the questions would have a number of correct answers, there are occurrences where some users did not detect all the correct answers, therefore the success rating for each individual task is measured as a percentage with a 100% score representing the case where a user has found all the applicable instances. Figure 9 illustrates the comparison of those who got at least one right answer and those who got all the right answers for each given task.

As is apparent from the graph, for the most part, users consistently found all the correct answers for each task, with a single outstanding outlier, which has been attributed to the relative ambiguity of the question.

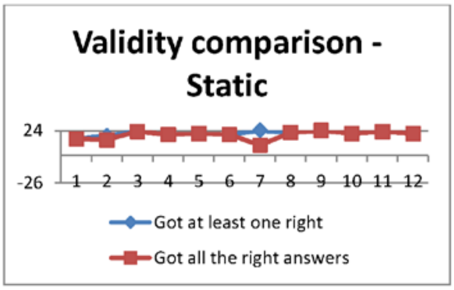


Figure 9. Correctness of the results (static test)

Therefore in the case of the static grid, we can state that the layout appears to be an effective, fast and accurate method of expediting instance recognition tasks. For real world applications though, these findings are not directly applicable, as the limited numbers of thumbnails that can be concurrently displayed on screen at any given time are too few for extensive video search tasks. Therefore, we move on to the next part of the experiment, lifting the limit on the quantity of thumbnails available (but not displayed) by using a scrollable grid.

4.2 Scrolling grid

The results of the scrolling grid layout tasks paint a different picture than for the static cases. The average success rate, across all the thumbnail sizes remained good, but lower than the static grid results, with a final average of 61% and half the users scoring below 56%. Figure 10 shows the success rates on the individual tasks for the scrolling grid.

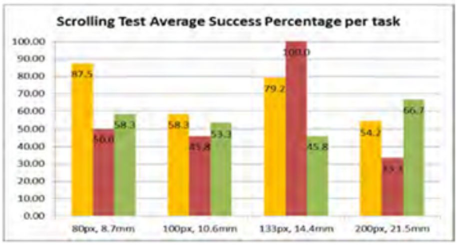


Figure 30. Success rates (scrolling test, detailed results)

Figure 11 below illustrates the average success rates for the 4 groups of thumbnail sizes. In the case of the third group (133 pixels) an outstanding outlier occurred where all users scored 100% on one of the given tasks, thus explaining the comparatively high success rate for thumbnail sizes of 133px. Although the otherwise observable decrease in performance for larger thumbnail sizes seems low, a comparison with the related increase of performance in a static scenario (cf. Fig. 6 and 7), confirms our expectation that the need for interaction can have a negative effect on retrieval performance.

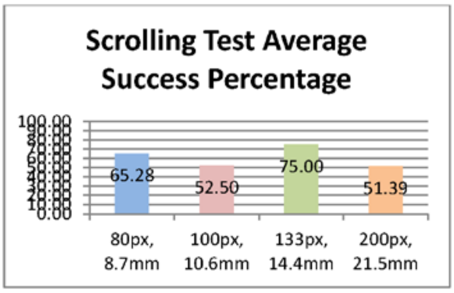


Figure 41. Success rates (scrolling test, averages)

The time performance was also lower in comparison with the static grid tasks, with the average time required to complete each task rising to 34 seconds. Figure 12 shows the average time required per task according to thumbnails size. The higher times can be attributed in some degree to the fact that there were physically more thumbnails for the user to peruse, approximately 2.5 times more. If the physical amount of time spent interacting with the interface, i.e. 'swiping' is taken into account and not added to the total time taken to answer then time difference between the static and scrolling layout tasks is diminished.

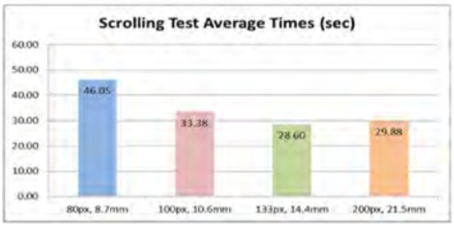


Figure 52. Time to solve the tasks (scrolling test, averages)

As far as the validity comparison results for the scrolling grid tasks, Figure 13 illustrates that, unlike in the static grid tasks, the users who did actually succeed in correctly completing the tasks found all the right answers for those tasks and not just some of them (hence there is no blue line on the graph). However the task success rates remain lower than for the static cases. Therefore, we can speculate that uncontrolled, continuous, non-discrete, scrolling is not an immediate 'upgrade' to the static layout and that further refinements must be made to maintain high success rates and lift the restrictions of the static layout.

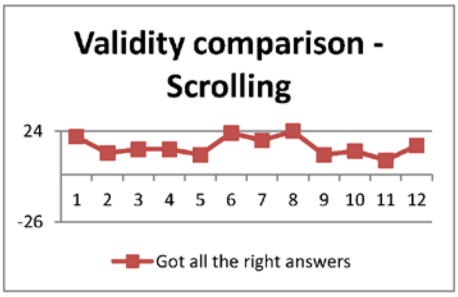


Figure 63. Correctness of the results (scrolling test)

4.3 Post-test user feedback

Immediately after finishing the 24 tasks set to them, the users were presented with a short questionnaire querying them on the experience and opinion of the interaction. There were no direct questions regarding the parameters of the tasks, such as eliciting preferences for thumbnail size and quantity, as the aim of this questionnaire was to determine how whether the setup was 'usable' from a human centric point of view. The following Table 4 illustrate their answers.

Table 4 Qualitative user feedback.

	Negative	Neutral	Positive
Was the interaction intuitive?	4	9	11
Was the interaction enjoyable?	3	14	7
Would you use it?	7	9	8

On the whole, the results show that the users' opinions did not indicate any significantly strong trends. There were not many downright negative opinions regarding the intuitiveness of the interactivity, but a noticeable amount of users rated interaction lower with respect to fun and enjoyment. Interestingly however, most of them also expressed their dislike about the images being too small in the settings where less interaction was required. Several users also expressed the opinion, during and after the test, that they found the static grid more comfortable and intuitive. Indeed it was apparent to the observer that nearly all the users preferred to tackle the problem as an isolated onscreen challenge.

Despite the more neutral rating regarding the characterisation of the interaction, a majority of users declared that they would use such an interface. Overall, user opinions indicate that a grid layout could indeed be a welcome part of video retrieval interfaces; this is also well substantiated by the quantitative data; despite the expressed misgivings of some of the users, their success rates were consistently high, as seen above in the data analysis. Even in the scrolling grids the success rate rarely dropped below 50%.

Although many users, especially those belonging to the advanced age groups, expressed that they felt daunted by the tasks that utilized the smallest size thumbnails and worried about the accuracy of their scores. However, the data showed that they fared no worse than users from younger age groups. We expected some variance of the success rates due to age related factors (eyesight, experience, etc.) but these were not immediately apparent in our test sample and will be revisited in the future with larger samples.

5. CONCLUSIONS & FURTHER WORK

Our analysis shows that the static thumbnail grid layout achieved and maintained very high accuracy over all the different thumbnail sizes. Compared to previous works in this area [7, 8], it is clear then that the effectiveness of thumbnail based interfaces that was observed in the related tests is successfully, if not quite fully, transferred to more complex layouts. Most importantly, a similar increase of optimum thumbnail size from 80 to 110px as reported by [8] when moving from solitary thumbnails to film strip representations could not be observed when moving to the even more complex matrix-style storyboard layout – where thumbnail sizes of about 130px achieved an equally high performance rate of 90% as the much larger 200px size version of the interface. This is particularly encouraging, since the layouts evaluated in our experiment can be far more applicable and useful to real world scenarios.

Unfortunately, these same levels of successful results were not similarly observed in the case of scrolling grid layouts, where the added layer of interaction complexity took its toll and rendered the system less effective. This confirms our assumption about the critical influence of the interaction design on retrieval performance. It is in line with the observation reported by [8] that users did not interact as much with the system as earlier results reported in [7] would suggest. The difficulties users experience with the scrolling interface can be likened to the issues that are apparent in all static peephole interfaces [9], where the additional spatial distortion of the content is added to the temporal distortion, further increasing the cognitive load of the user.

Therefore, it can be concluded that uncontrolled, continuous scrolling grid layouts are not a clear improvement over the static grid layout. Further refinements must be made to maintain the high success rates of the static layout and at the same time, remove its inherent restrictions. Possible options for further investigations include, for example, interface designs where users can switch between various static grid arrangements (similarly to swiping through different screens on modern smartphone interface designs) in contrast to the continuous scrolling evaluated in our experiment.

These conclusions of our study provide further insight into better use of complex thumbnail based layouts for mobile video retrieval tasks, but obviously, they are just one more step towards the ultimate goal of creating the perfect video retrieval interface for mobile devices. As already said above, interaction design is a crucial issue and the better, more intuitive, yet efficient and effective interaction modes need to found. Considering the actual layout, our results suggest that despite the small screen sizes of mobiles, even more complex interface designs such as scrollable linear strips, multi-tiered grids, and even 3D interfaces [12] could be possible. Likewise, the usefulness of dynamic thumbnails identified in [7] needs to be evaluated in more complex designs.

Further investigation is warranted into the exact specifications of the optimal balance between larger thumbnails (featuring more detail) and larger numbers of smaller (less detailed) thumbnails that a user has to analyze to complete a task. Is it faster and easier to examine fewer, more detailed images, than it is to scan through several less detailed ones or the opposite? The answer to this question is certainly dependent on the task at hand indicating the need for further research about other retrieval than the ones studied here. For example, how effective are the proposed layouts for "event recognition" tasks that will gauge the ability of users to accurately detect events and series of actions in a video instead of just instances of entities. And how should different domains and contexts be considered in the interface design?

6. ACKNOWLEDGMENTS

This work was partially supported by a Yahoo! Faculty Research Grant. We would like to thank our test subjects for their time and willingness to participate in the experiment.

7. REFERENCES

[1] Adcock, J., Cooper, M., Girgensohn, A., & Wilcox, L. (2005). Interactive Video Search Using Multilevel Indexing. *In Proc. CIVR*, (pp. 205-214).

[2] Boreczky, J., Girgensohn, A., Golovchinsky, G., & Uchihashi, S. (2000). An Interactive Comic Book Presentation for Exploring Video. *Proc. ACM CHI*, (pp. 185-192).

[3] Christel, M.G., Warmack, A., Hauptmann, A.G., & Crosby, S. (1999). Adjustable Filmstrips and Skims as Abstractions for a Digital Video Library. *Proc. IEEE Advances in Digital Libraries Conference*, (pp. 98-104).

[4] Christel, M.G., Hauptmann, A.G., Wactlar, H.D., & Ng, T.D. (2002). Collages as Dynamic Summaries for News Video. *Proc. ACM Multimedia*, (pp. 561-569).

[5] Christel, G. C. (2006). Evaluation and User Studies with Respect to Video Summarization and Browsing. *Multimedia Content Analysis, Management and Retrieval* 2006, Proc. of IS&T/SPIE Symposium on Electronic Imaging.

[6] Christel, M. G., & Yan, R. (2007). Merging storyboard strategies and automatic retrieval for improving interactive video search. *CIVR '07 Proc. of the 6th ACM International Conference on Image and video retrieval* (pp. 486-493).

[7] Hürst, W., Snoek, C. G., Spoel, W.-J., & Tomin, M. (2010). Keep moving!: revisiting thumbnails for mobile video retrieval. *MM '10 Proc. of the International Conference on Multimedia* (pp. 963-966).

[8] Hürst, W., Snoek, C. G., Spoel, W.-J., & Tomin, M. (2011). Size matters! How thumbnail number, size, and motion influence mobile video retrieval. *MMM'11 Proc. of the 17th International Conference on Advances in Multimedia Modeling* (pp. 230-240).

[9] Mehra, S., Werkhoven, P., & Worring, M. (2006). Navigating on handheld displays: Dynamic versus static peephole navigation. *ACM Trans. Comput.-Hum. Interact.* 13(4), (pp. 448-457).

[10] Oliva, A., & Torralba, A. (2007). The role of context in object recognition. *Trends in Cognitive Sciences*, (pp. 520-527).

[11] Schoeffmann, K., Hopfgartner, F., Marques, O., Boeszoermenyi, L., & Jose, J.M. (2010). Video browsing interfaces and applications: a review. *SPIE Reviews*, 1(1), (pp. 1-35).

[12] Schoeffmann, K., Ahlström, D., & Böszörmenyi, L. (2012). Video Browsing with a 3D Thumbnail Ring Arranged by Color Similarity. *Advances in Multimedia Modeling*, Springer, Berlin, Heidelberg, New York. (pp. 639-641).

[13] Smeaton, A. F., Over, P., & Kraaij, W. (2006). Evaluation campaigns and TRECVid. *Proc. of the 8th ACM International Workshop on Multimedia Information Retrieval* (pp. 321-330).

[14] Smith, N. (2011/08/23). Baby Boomers Up-to-Date on Smartphone Technology, Study Finds. *BusinessNewsDaily*. http://www.businessnewsdaily.com/1380-baby-boomers-smartphone-savvy.html

[15] Snoek, C.G.M., Worring, M., Koelma, D.C., & Smeulders, A.W.M. (2007). A Learned Lexicon-Driven Paradigm for Interactive Video Retrieval. *IEEE Transactions on Multimedia* 9(2), (pp. 280-292).

[16] Torralba, A., Fergus, R., & Freeman, W. T. (2008). 80 Million Tiny Images: A Large Data Set for Nonparametric Object and Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (pp. 1958-1970).

9.2 Showdown paper

HiStory – A Hierarchical Storyboard Interface Design for Video Browsing on Mobile Devices

Wolfgang Hürst and Dimitri Darzentas

Department of Information and Computing Sciences, Utrecht University
Princetonplein 5, 3584 CC Utrecht, The Netherlands
huerst@uu.nl, d.p.darzentas@students.uu.nl

Abstract. This paper presents an interactive thumbnail-based video browser for mobile devices such as smartphones featuring a touch screen. Developed as part of on-going research and supported by user studies, it utilises a hierarchical storyboard design which provides an interface metaphor that is familiar and intuitive yet supports fast and effective completion of Known-Item-Search tasks by rapidly providing an overview of a video's content with varying degrees of granularity.

Keywords: Video Retrieval, Video Browsing, Mobile Interfaces, Touch Interfaces

1 Introduction

We introduce an interactive thumbnail-based video browser for mobile devices such as smartphones and tablets featuring touch screens. Similarly to [1], our approach relies on an intuitive visualisation technique that is optimised for human visual perception and cognition in Known-Item-Search (KIS) tasks. The proposed interaction design is based on our previous detailed studies investigating optimum thumbnail sizes [2], film strip visualisations [3], and storyboard layouts [4] on mobile platforms. The method was implemented as part of on-going research in the field and was optimised for KIS tasks of known segments in single video files in accordance with the video browser showdown [5].

2 The Hierarchical Storyboard Design

Like most common interfaces for video browsing, our Hierarchical Storyboard (HiStory) approach uses thumbnails as building blocks. Our design consists of two elements, a storyboard-style grid of thumbnails, taking up most of the available screen space and a narrow vertical bar on the right side of the screen. The grid serves as the primary interaction and visualisation point and the bar acts as a non-interactive visual aid that illustrates which part of the video is visible at any time. Both elements are detailed below.

The grid features a traditional storyboard design with temporally ordered thumbnails that represent still images extracted from the video. Depending on the size and resolution of the mobile device used, the dimensions of the grid and the thumbnails vary. Figure 1 illustrates an example from a common mobile phone where the video has been split into 36 thumbnails arranged in a traditional storyboard layout.

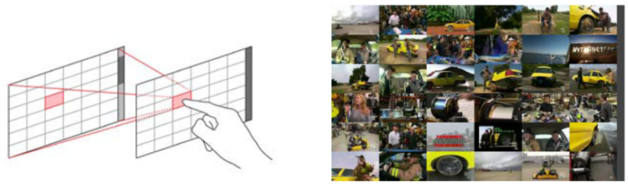


Fig. 1. Illustrated on the left is a diagram of the Hierarchical Storyboard concept. On the right is an example of how a video appears on a device's screen.

The indicator bar on the side functions as a reference, representing the currently viewable portion of the video. That is, when at full length, spanning the height of the screen, it indicates that the entire length of the video is currently arranged on the grid. When the viewed portion changes, the length and vertical position of the bar also changes, in order to provide a useful and fast positional reference. The core mechanic of HiStory is the ability of the user to dynamically change the granularity of the grid in a process similar to 'zooming'. The user specifies a thumbnail and the grid is rebuilt in a lower hierarchy with a smaller time interval around this 'anchor' thumbnail. This means that the viewed range shrinks (illustrated by the indicator bar) but the time interval between each thumbnail also becomes smaller, leading to more detailed information (finer granularity). Thus the previously disassociated thumbnails gain context as scenes as shown is Figure 2. Intuitively, the technique is similar to changing the scale of a map, affording more detail of a specific area while eschewing the general overview. The user can 'zoom in' multiple times, until the time interval between the thumbnails shrinks enough for a frame-by-frame representation. Inversely, they can also 'zoom out', backtracking through previous choices all the way to the top level or choosing a new thumbnail to use as an anchor point to go down a level. Parameters such as the number and size of the thumbnails are based on previous research in the area and were validated by a series of user studies. [1][2][3]

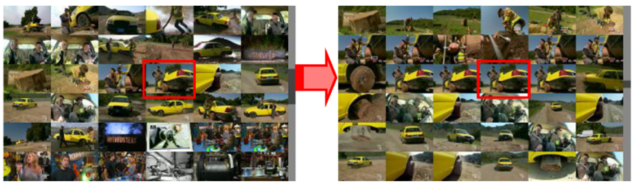


Fig. 2. The user has chosen a thumbnail, 'zooming in' and changing the granularity

3 Further Discussion

In the developing domain of mobile video browsing, our method builds upon in-depth research into the abilities of human perception in such circumstances; utilising them to their limit in order to overcome the inherent limitations of mobile devices. Based on promising results from user studies, we hope to further develop this method with feedback from the video browser showdown. Our initial endeavours suggest that the platform-born technical barriers such as small screen size and low processing power are surmountable and with the ever increasing trend for mobile computing continuing to hold centre stage, we hope to add the task of video retrieval as prime contender in the field.

As part of this research, an in-depth user study has been performed in order to gauge user performance, for both speed and accuracy, and to gather user feedback and preferences. The preliminary findings of this study have shown that, following a very short learning curve, users, from a wide age range and diverse levels of experience, were quick to understand and adopt the method in order to easily and accurately complete the given tasks. In our user study, users managed to rapidly and effectively determine the content of a video as well as accurately complete KIS tasks, demonstrating the apparent ease of use and intuitiveness of the presentation and interaction techniques of the method.

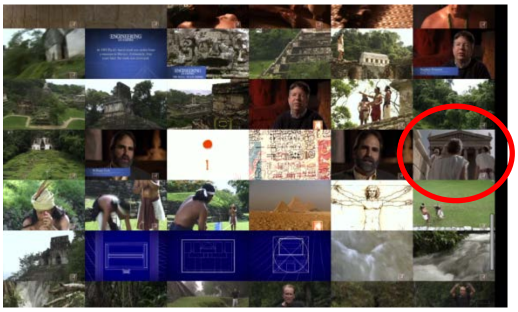
References

1. M. del Fabro, K. Schoeffmann, and L. Laszlo Böszörmenyi. Instant video browsing: A tool for fast non-sequential hierarchical video browsing. In G. Leitner, M. Hitz, and A. Holzinger, editors, HCI in Work and Learning, Life and Leisure, volume 6389 of Lecture Notes in Computer Science, pages 443-446. Springer Berlin / Heidelberg, 2010.
2. Hürst, W., Snoek, C. G., Spoel, W.-J., & Tomin, M. Keep moving!: revisiting thumbnails for mobile video retrieval. MM '10 Proc. of the International Conference on Multimedia (pp. 963-966), 2010.
3. Hürst, W., Snoek, C. G., Spoel, W.-J., & Tomin, M. Size matters! How thumbnail number, size, and motion influence mobile video retrieval. MMM'11 Proc. of the 17th International Conference on Advances in Multimedia Modeling (pp. 230-240), 2011.
4. Hürst, W. and Darzentas, D. Quantity versus quality: the role of layout and interaction complexity in thumbnail-based video retrieval interfaces. In *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval* (ICMR '12). 2012.
5. Video Browser Showdown 2013 (http://mmm2013.org/Video_browser_showdown.htm)

9.3 Text Questions for the first experiment and thumbnail answers

1. Find the image of the woman in the long green dress standing in front of a staircase
   Βρείτε την εικόνα με την γυναίκα που φορά ένα μακρύ πράσινο φόρεμα μπροστά σε σκάλες
   
2. Find the image of the glass fronted skyscraper
   Βρείτε την εικόνα με τον ουρανοξύστη
   
3. Find the image of the hand turning the knob
   Βρείτε την εικόνα με το χέρι που γυρίζει ένα διακόπτη
   
4. Find the image of a car side view mirror
   Βρείτε τον καθρέφτη του αυτοκινήτου
   

5. Find the image of 3 men in front of a Greek temple
   Βρείτε την εικόνα όπου 3 άντρες στέκονται μπροστά σε ένα ελληνικό ναό



6. Find the image of a white horse
   Βρείτε την εικόνα ενός άσπρου αλόγου

7. Find the image of the car with the lit up headlights
   Βρείτε την εικόνα ενός αυτοκινήτου που έχει αναμμένα φώτα

8. Find the image of 4 men wearing red robes and red hats
   Βρείτε την εικόνα οπού υπάρχουν 4 άνδρες που φοράνε κόκκινες ρόμπες και καπέλα

9. Find the image of the hands holding playing cards
   Βρείτε την εικόνα όπου υπάρχουν δύο χέρια που κρατάνε τραπουλόχαρτα

10. Find the image of the woman wearing glasses
    Βρείτε την εικόνα μιας γυναίκας που φορά γυαλιά

11. Find the image of two hands wearing yellow washing up gloves
    Βρείτε την εικόνα όπου υπάρχουν δυο χέρια που φοράνε κίτρινα πλαστικά γάντια

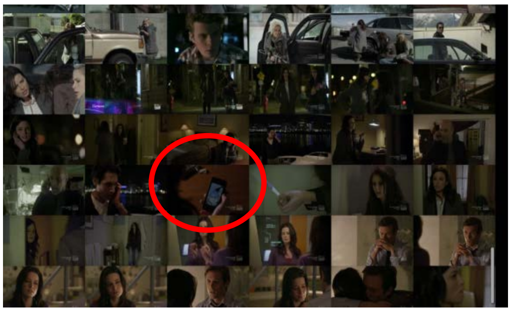
12. Find the image of the red double decker bus
    Βρείτε την εικόνα όπου υπάρχει ένα κόκκινο διώροφο λεωφορείο

13. Find the image of two hands in a handshake
    Βρείτε την εικόνα όπου υπάρχει μία χειραψία

14. Find the image of the car in the underground car park
    Βρείτε την εικόνα όπου υπάρχει ένα αυτοκίνητο σε ένα υπόγειο παρκινγκ

15. Find the image of the man putting on a white coat
    Βρείτε την εικόνα όπου ένας άντρας βάζει μία λευκή ιατρική μπλούζα

16. Find the image of the hand holding a smart phone.
    Βρείτε την εικόνα όπου υπάρχει ένα χέρι που κρατά ένα smartphone



9.4 Data

9.4.1 Scrolling Experiment t-test

t-Test: Two-Sample Assuming Unequal Variances		
	Continuous	*Paged*
Mean	38.69712	37.21635
Variance	795.6421	709.4747
Observations	208	208
Hypothesized Mean Difference	1	
df	413	
t Stat	0.178724	
P(T<=t) one-tail	0.429121	
t Critical one-tail	1.648551	
P(T<=t) two-tail	0.858242	
t Critical two-tail	1.965725	

9.4.2 HiStory Experiment t-test – Continuous/Paged

t-Test: Two-Sample Assuming Unequal Variances		
	Continuous	*Paged*
Mean	61.51041667	55.609375
Variance	2090.376854	1709.33876
Observations	192	192
Hypothesized Mean Difference	0	
df	378	
t Stat	1.326489293	
P(T<=t) one-tail	0.092739264	
t Critical one-tail	1.64889472	
P(T<=t) two-tail	0.185478527	
t Critical two-tail	1.966259636	

9.4.3 HiStory Experiment t-test – Continuous/HiStory

t-Test: Two-Sample Assuming Unequal Variances		
	Continuous	*HiStory*
Mean	61.51041667	71.68229167
Variance	2090.376854	4756.102721
Observations	192	192
Hypothesized Mean Difference	0	
df	332	
t Stat	-1.703405207	
P(T<=t) one-tail	0.044714032	

t Critical one-tail	1.649456205	
P(T<=t) two-tail	0.089428064	
t Critical two-tail	1.967135057	

9.4.4 HiStory Experiment t-test– Paged/HiStory

t-Test: Two-Sample Assuming Unequal Variances		
	Paged	*HiStory*
Mean	55.609375	71.68229167
Variance	1709.33876	4756.102721
Observations	192	192
Hypothesized Mean Difference	0	
df	313	
t Stat	- 2.769785993	
P(T<=t) one-tail	0.002972115	
t Critical one-tail	1.649736428	
P(T<=t) two-tail	0.00594423	
t Critical two-tail	1.967572019	