Electronic Gridlock: Reducing Download Latency For Video on the Web

Chris Johnson† and Jim Kavanagh††,

†*Department of Computing Science,*
††*Department of Geology,*
University of Glasgow, Glasgow, Scotland, G12 8QQ.
E-mail: johnson@dcs.glasgow.ac.uk,
Web: http://www.dcs.glasgow.ac.uk/∼ johnson
Phone: +41 330 6053 Fax: +41 330 4913

The introduction of interactive browsers, such as Mosaic and Netscape, has attracted new generations of users onto the Internet. This has reached a point where we are now faced with the threat of electronic gridlock. Many of the world's most popular sites are difficult or impossible to access during long periods of the day. In the short term, there is little prospect of resolving this problem by improving the underlying communications networks. Most users are restricted to the limited bandwidth offered by ethernet and modem connections. This paper, therefore, examines way of reducing the problems created by download latency. This is defined to be the period of time required to retrieve remote resources. In particular, it is argued that cinematic presentation techniques can be used to reduce the frustration that can arise when users attempt to access video documents over the World Wide Web. Finally, the paper argues that there is a pressing need for improved validation techniques if designers are to gain empirical support for human computer interfaces to distributed information systems.

Keywords:Electronic gridlock; World Wide Web; multimedia.

1 Introduction

World Wide Web browsers, such as Mosaic and Netscape, are providing new generations of users with access to a vast range of information resources [1]. There are World Wide Web pages devoted to Levi's jeans, to Guinness' beers, to Ford's cars and Sony's records. This diversity illustrates an important change in the use of the World's computer networks. In the past, the Internet was solely the preserve of computer professionals. Increasingly, these systems are being used as a means of mass market communication. This change has created a number of problems. In particular, we are facing the problem of 'electronic gridlock'. Many of the most popular sites are almost impossible to contact during peak periods. The situation is changing so rapidly that this problem has developed in under a year. Twelve or thirteen

months ago, it was relatively easy to download many of the pages provided by the National Centre for Supercomputer Applications. Today this is no longer the case.

The delays that are being introduced into the world's computer systems have important consequences for interface designers. Temporal properties have a critical impact upon the quality of human computer interaction. In particular, slow response times can lead to frustration and error [9]. This paper addresses one form of delay that causes tremendous problems for the current generation of World Wide Web browsers. In particular, we are concerned to reduce the problems that are created by download latency. This is the delay that occurs between a user requesting information from a remote site and the time that the information is actually presented to the operator [6]. This problem is particularly severe because users cannot determine the affect that network loadings will have upon their requests until they issue the command to begin retrieving information [4].

1.1 The Causes Of High Download Latency: Demand Push

In 1980 there were approximately one hundred computers attached to the Internet. In 1990 there were one hundred thousand. In 1994, the number of systems connected to the Internet exceeded one million. Recent estimates placed the number of Internet users at well over twenty-five million [12]. Two factors can be used to explain the tremendous increase in demand for distributed information resources. The first is that a series of innovative interactive applications, including Gopher, Mosaic and Netscape, has encouraged the active participation of new groups of users [1]. The Internet is no longer restricted to those computer professionals that understand the intricacies of ftp, telnet and the standard Unix mail facilities [5].

The increasing demand for Internet resources has not been mirrored by a similar revolution in the underlying communications technology. Only a small percentage of Web users are connected by high-bandwidth networks, such as the United Kingdom's SuperJANET. The majority still rely upon the relatively low bandwidth facilities provided by Ethernet and modem connections. There are further theoretical reasons why the situation will not improve. Electronic gridlock and high download latencies are symptoms of a fundamental problem for computing science that has become known as the Von Neumann bottleneck [6]. Communications facilities cannot keep pace with advances in storage and processing technology. This theoretical result has critical consequences for the future of Web based information systems. As more and more people require a greater and greater volume of information resources over finite bandwidths there will be an increase in the time that is required to download information from remote sites [6]. These delays may limit the continued expansion of the Internet as high download latencies increase the sense of frustration when using systems such as Mosaic and Netscape.

1.2 The Causes Of High Download Latency: Supply Pull

The quality of the resources that are offered over the Internet has changed over the last two or three years. Sites that used to provide textual information now also offer high quality photographs and video sequences. This change has been bought at a cost. In 1993, the author's home page required approximately 632 bytes of disk space. The shortest of the video clips that can be accessed through the 1995 page requires a megabyte. This trend towards the provision of resource intensive facilities is set to continue as companies develop novel means of interacting over the Internet. The benefits offered by distributed programming languages, such as Sun Microsystem's Hot Java, are bought at a cost in terms of the communications that they require. From the users' perspective, these might seem to be irrelevant engineering details. Disk space is cheap and processor technology is advancing. Unfortunately, the impact

of new presentation facilities is not simply felt in the amount of space required to store multimedia resources. Nor is it simply felt in the amount of processing power that is required to manipulate those resources. It also affects the amount of time that is required to access and retrieve information over distributed networks.

Faced with the irritation of high download latencies, user interface designers must find effective means of transferring large amounts of information over finite bandwidth communications networks. As mentioned, there are theoretical and practical limits on the network improvements that might be expected in the immediate future. This paper, therefore, explores alternative means of reducing the problems that are created by high download latencies. In particular, it is argued that improved display design can be used to reduce the frustration that often arises when users must retrieve video sequences over wide area networks.

1.3 Structure Of The Paper

Section 2 introduces the system that is used to illustrate our argument. The Hunterian Museum in Glasgow was anxious to provide people in the remoter areas of Scotland with access to a range of video exhibits. During the development of this system we became increasingly aware of the problems created by download latencies. In particular, it can be difficult for users to predict how long retrieval requests will take without a detailed knowledge of the underlying networks and the loadings on remote machines.

Section 3 argues that video clips must be long enough to convey significant information to the user and yet short enough so that they do not suffer from an excessive download latency. Unfortunately, the unpredictability identified in the previous section makes it difficult for designers to identify this optimal clip length. High loadings on particular machines can increase download latencies. This has the effect of reducing the optimal length of a video clip over the course of a day and between various locations on a network. It is, therefore, argued that designers must provide a range of clip lengths if they are to encourage users to access longer videos on heavily loaded networks.

Section 4 goes on to describe interface design techniques that have been developed to reduce the frustration caused by unpredictable download latencies. In particular, it is argued that users must be able to assess the benefits of downloading video material. Photographic icons and cinematic page design can be used to provide information about the quality of production and the content of the film. Running times and brief synopses can help users to determine whether the costs of retrieval, in terms of download latencies, are worth the benefits of viewing the video.

Section 5 presents the results of a preliminary evaluation that has been conducted to investigate the utility of these techniques. Section 6 presents the conclusions that can be drawn from the work in this paper. Directions for further work are also discussed.

2 The Example: A Video Exhibition For The Hunterian

Early in 1995, a multi-disciplinary team was formed between the Department of Computing Science, Glasgow University and the Schools Liason Officer at the Hunterian Museum. Our intention was to develop a 'leading edge' web site for Scotland's oldest Museum. Our work faced a number of problems that forced us to focus upon the issues of download latency. The Museum was anxious to provide people in the remoter regions of Scotland with some of the resources that are available in the major population centres. Any electronic facilities that we provided had to be accessible using the low bandwidth networks and limited processing

capabilities that are available to these remote users. At first, our work focussed upon the presentation of text and photographs. The results can be viewed on:

`http://www.gla.ac.uk/Museum`

However, we quickly became aware that many exhibits could not be presented using static media. For example, Figure 1 presents a video clip about the flexibility and strength of Roman armour. This material makes little sense unless users can view the way that various layers of

Figure 1: Frame from Hunterian Video On Roman Armour.

clothing moved to protect the soldiers.

3 Optimal Clip Lengths For Video Retrieval

The provision of video resources over the Internet creates many problems for interface designers. The first issue is how to produce a high-quality film. Too often, users are disappointed by the production and direction of multimedia resources on the web. We were extremely fortunate in this respect. Scottish Television provided professional support in shooting a series of clips for a programme about the Museum. Our next problem was that the existing client-server architecture for World Wide Web document delivery does not support multiple 'live' video broadcasts. Designers cannot simply open a connection and play the material direct to the users' local machines. Instead, video files have to be sent over the network. These are then played by helper applications that are controlled through the users' browser. This creates problems because video files have to be of a fixed size. Standard interface guidelines provide little or no advice about optimal file sizes for video clips. This decision has a critical impact upon download latency. The bigger a file is, the longer it takes to be downloaded. We, therefore, conducted a series of tests to measure the download latency of various lengths of film under optimal conditions. The intention was to determine how long it would take to retrieve video material before considering the unpredictable delays that would affect the final

system. The results are shown in Figure 2. The file sizes are not at regular intervals because this could not be achieved without cutting the video segment in an arbitrary fashion.

As expected, the time taken to download a video is proportional to the size of the file. It is also worth noting that file size remains roughly proportional to the length of the video

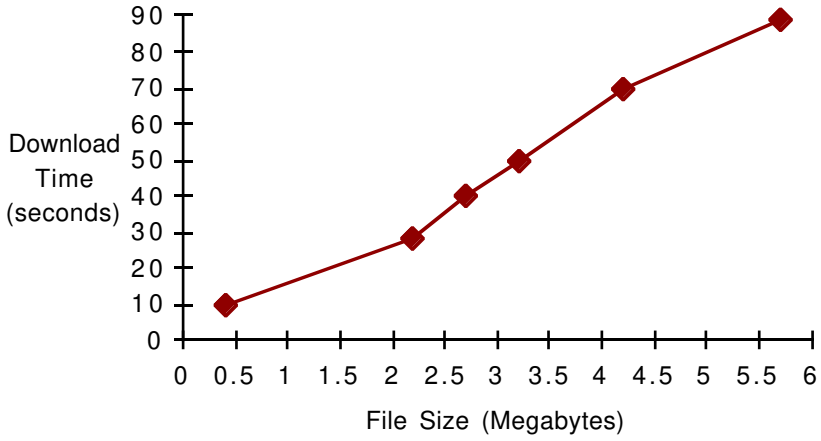

Figure 2: Variable Download Latency With File Size.

under the major MPEG and QuickTime video formats. This would have been significant to both users and information providers if, for instance, the standard formats involved a disproportionate overhead for small files. Figure 3 provides data for the file size that we

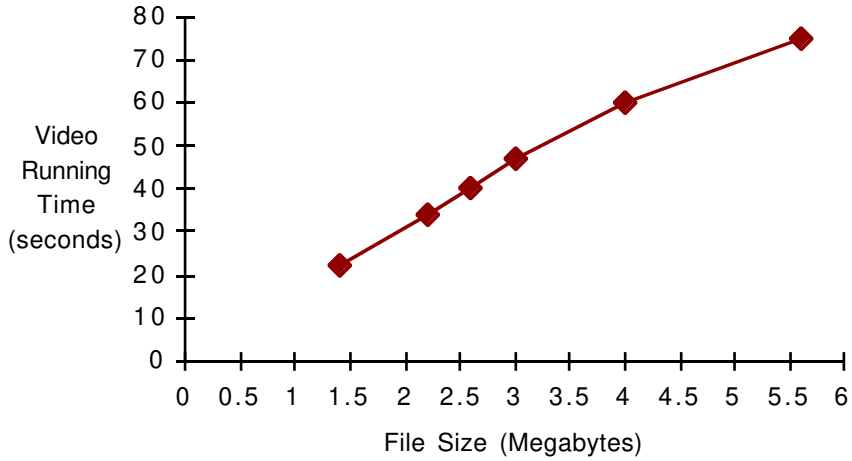

Figure 3: Variable File Size With Length of Video (QuickTime files).

obtained on our example clips of video . The previous two figures, therefore illustrate that the longer the video, the bigger the file needed to store the resource and the bigger the file needed to store the resource, the longer it takes users to download the file in questions. The extent of download latency is, therefore, directly related to the length of the video segment.

The task of identifying an optimal clip length for our video material was exacerbated by the fact that download latencies differ in various parts of the globe. Figure 4 presents the results of a trial that we conducted to illustrate this point. A number of users were asked to download one of our video files at approximately the same time of day. There are a number of interesting anomalies in this data. For instance, Dublin took longer than Montpellier. This

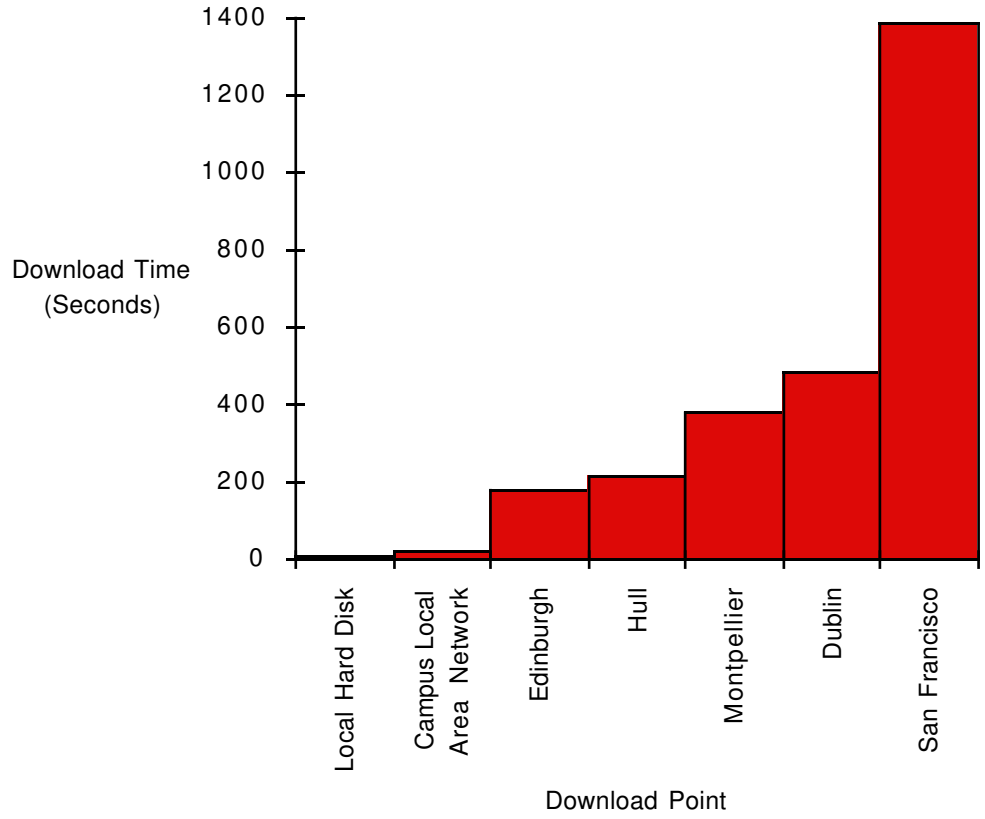

Figure 4: Variable Download Latency With Users' Location.

is significant because Dublin is 200 miles away from Glasgow while Montpellier is approximately 750 miles further away. If video resources were to be provided for a group of users in Dublin, therefore, the recommendations on file size and clip length would have to be based upon network properties rather than geographical location. For completeness sake, the delay between Dublin and the United Kingdom arises from a low bandwidth 256 kbit line. The Irish government have recently attempted to reduce the delays in their Internet access by installing a 512 kbit line across the Atlantic to the United States. It is important to note, however, that such initiatives do not remove the problems of download latency. Many users still rely upon low bandwidth ethernet and modem connections at a local level. Those users that do have access to high performance resources are increasingly tempted to load their systems with more advanced multimedia resources [12].

Identifying an optimal clip length is further complicated by the fact that network traffic and hence download latencies vary over time. Anyone using the Internet will have experienced the consequences of this. It is surprising, however, that the Human Computer Interaction literature has ignored the impact that this is having on operator behaviour. For instance, many users now schedule their Internet activities to maximise the 'fallow' periods before the United States come on-line [2]. The temporal characteristics of network loading also have an impact on download latency at a more local level. Figure 5 provides the timings for retrieving a video file over the Glasgow University network at different times during the day.

The previous paragraphs have presented a bleak picture for interface designers and information providers. Download latences are affected by clip length. Unfortunately, optimal clip lengths vary according to the network location of the user and the time of day that the infor-

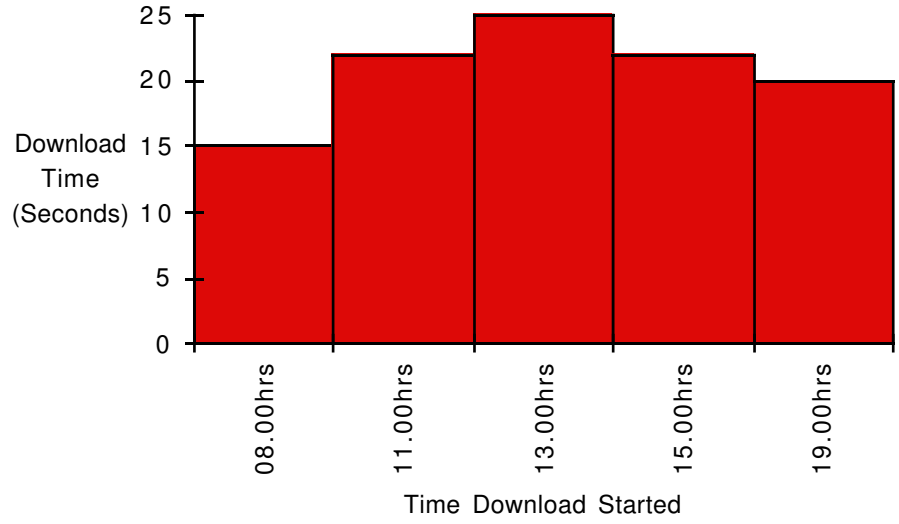

Figure 5: Variable Download Latency With Time Of Day.

mation is requested. With this in mind, designers must offer a range of different video clips with different running times. This might seem to be a trivial guideline and yet the authors could find no other web site that exploited this approach to the provision of video material. A survey of fourty-three sites revealed average clip lengths in the region of thirty-five seconds. None of these sites offered clips that were longer than fifty seconds, even though fallow periods on regional networks could support the resulting download latencies. Conversely, only three sites provided clips of less than thirty seconds in length. Such short 'snippets' are extremely important. They provide users with a low cost means of determining whether the longer videos are worth viewing in spite of retrieval delays.

4 Improved Interface Design For Video Retrieval

Our first attempts to develop video pages for the Hunterian quickly made it apparent that users needed detailed information about the content of the films before they downloaded them. The poor quality of many multimedia resources on the web has made users reluctant to download large files if there is a chance that the content will not be worth the trouble. Our first page, exploited the standard web approach of using selectable text to initiate the retrieval. In addition, the running times for each film were presented to provide users with a rough idea of the anticipated retrieval delay. It is important to note that the duration of the clip was more significant in terms of the users retrieval task than the file size information presented by Mosaic and Netscape. Many of the 'visitors' to the Museum web site had only a vague notion of communications issues and video file formats. For example, our first design contained lines such as:

```
Clip A: Introduction To The Museum's Eductional Activities (Clip length 22 secs).
```

In retrospect, it is hardly surprising that few people were enticed to access the contents of this file! The disappointing response from visitors to the pages and from people actually entering the Museum persuaded us to revise this screen layout. Our initial design had been based around the sorts of labels that interface designers traditionally use to identify menu

options. They were supposed to be terse but informative. Our second attempt, was based around the sorts of descriptions that are frequently found in the television pages of magazines and newspapers [13]. This brought us further away from the traditional domain of Human Computer Interaction and moved us closer to the worlds of cinematography and advertising:

> `Romans In The Bathhouse. Archaologists have unearthed various objects which tell us about the daily life of the Romans. At the Bearsden Bathhouse we can see how the Romans washed.`

These textual descriptions were combined with photographic icons taken directly from the video. Users accessed the video file by selecting pictures such as those shown in Figure 6. Following May and Barnard [10], the intention was to exploit the techniques that are used by the distribution companies that are responsible for marketing feature films. Excerpts from the main action are used on posters and in display cases to encourage people to pay for their product. Figure 6 shows one of the pages that were developed for the Hunterian. A side-effect

Figure 6: The Eventual Screen Design For The Hunterian Video Page.

of introducing these images into our exhibits was that the retrieval time for the main page was increased. We were, therefore, anxious to gain evidence in support of our design decisions.

5 Validating Our Approach

In validating our work, we were concerned to support two different tasks that were being performed by our user population. Some people were, typically, accessing the Web pages to answer specific questions. For example, students were being set exercises by their lecturers; 'use the pages to find the nearest point on the Antonine wall'. Other users were simply browsing the pages in a less directed manner. Our investigations, therefore, focussed upon four different conditions. In the first, users were presented with the terse textual links to the videos. They were encouraged to browse the material and were then asked to give a qualitative reaction to the time it took for the video to be retrieved. They were free to select the videos

that they viewed and so only reported on those films that they actually downloaded. In the second condition, users were presented with the same textual interface but were given the task of answering a number of comprehension questions. The intention was to simulate the sorts of tasks that lecturers might set their students. The third condition was one in which users were browsing but had the cinematic layout illustrated in Figure 6. The final condition was one in which the users had the cinematic layout and had a series of questions to answer.

The evaluation of human computer interfaces to distributed systems raises a large number of methodological problems. For instance, the results of Figures 4 and 5 show that geographical distances and local network conditions have a profound impact upon download latency. This makes it difficult to replicate experimental conditions. The same user accessing the same video on the same machine might experience very different delays if one request were started a few minutes after another. This problem can be avoided by simulating retrieval delays on a local machine. With this in mind, we then had the problem of selecting users for the evaluation. We were concerned not just to recruit University undergraduates. Our evaluation was, therefore, conducted with sixteen users from a range of backgrounds including secretaries, technical staff and casual visitors to the Museum.

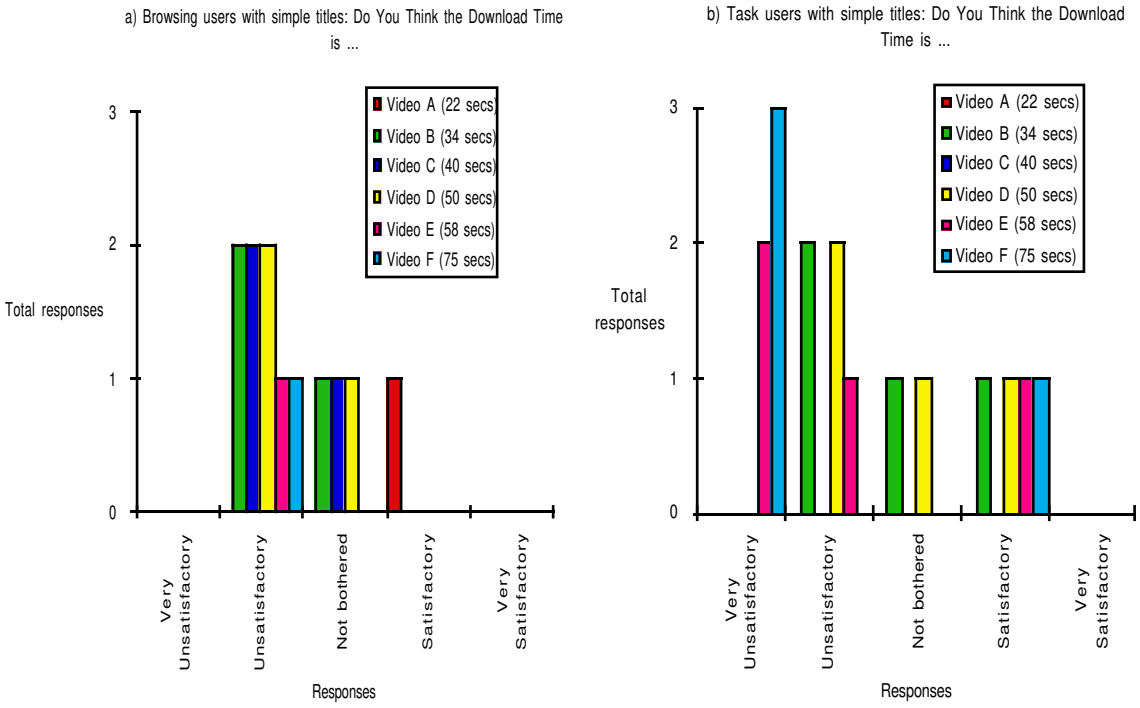

Figure 7: Results For The Textual Interface

Our evaluation faced a number of additional problems. For instance, O'Donnell and Draper argue that many users will adopt coping strategies to off-set the effects of machine delays [11]. There is considerable evidence for this in the use of our web pages. Many users coped with download latency by 'pacing' their requests. This strategy was based on the assumption that the shortest clip length involved the lowest cost in retrieving a video file. These users, therefore, decided to download the short clips first. Longer videos were only retrieved once they were satisfied that it would be worth the wait. In contrast, other users downloaded the videos in an apparently random order. The existence of these different coping strategies

makes it difficult to compare users' attitudes between our different experimental conditions. It is important to emphasise, therefore, that the development of appropriate methodologies for the empirical validation of interfaces to distributed systems remains an area of research in its own right. We are currently investigating these issues together with Steve Draper and Paddy O'Donnell of the Department of Psychology, Glasgow University in the Temporal Aspects of Usability (TAU) project.

Figures 7 a) and 7 b) show the qualitative results for the impact of download latency with the textual interface. In both the question and browsing conditions, users expressed negative attitudes towards the download latencies. Just over 60% of all retrieval requests were considered to be either unsatisfactory or very unsatisfactory. This finding could have been influenced by one or two users expressing consistently negative attitudes towards the interface in each condition. This analysis is not, however, supported by the data. All users expressed irritation with the delays.

Figures 8 a) and 8 b) show the results for the cinematic interface. As can be seen, the 'peaks' are now well to the right of those for the textual system. Only one request was considered to be very unsatisfactory in the browsing condition. All retrieval requests

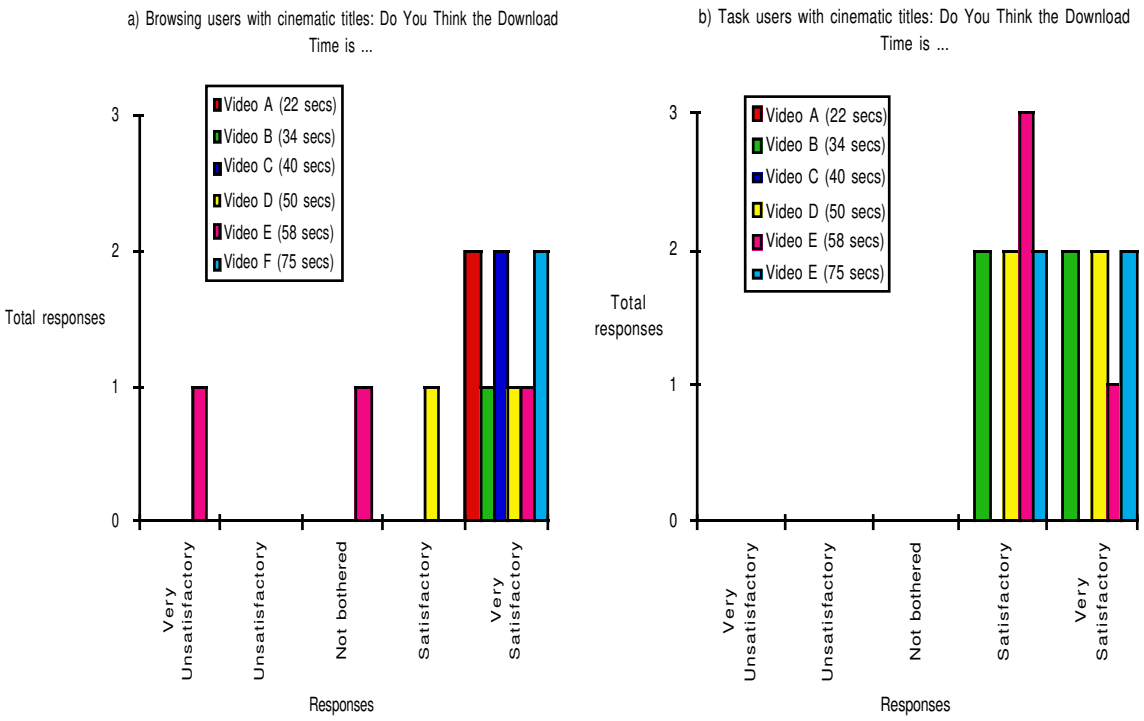

Figure 8: Results For The Cinematic Interface

for the users who were answering questions was considered either to be satisfactory or very satisfactory.

Brevity prevents a more detailed presentation of our results in this paper. Further details are provided in [7]. The previous Figures suggest that the cinematic approach to video retrieval creates a more positive attitude to download latency than the textual interface. A number of factors can be used to explain this finding. For instance, the information content in the revised Web page was greater than that in the textual labels. The pictures provided information about the content of the video: where it was shot; who appeared in the film etc.

This was clearly a significant issue for those users who were presented with the comprehension questions. The information in the icons provided important cues about the content of the video file. More importantly, perhaps, the picture icons also provided information about the production and reproduction quality of the multimedia resources. Several users commented that they had been disappointed when downloading poor quality video sources from other sites. The cinematic interface helped to demonstrate that the film had been produced to a professional standard.

6 Conclusion

This paper has argued that we are faced with the increasing problem of 'electronic gridlock'. The world's computer networks are filling with more and more users accessing more and more resources. Requests that might have taken a few seconds are now taking minutes and even hours. The underlying networks are slowly being improved. At the same time, however, users are demanding more sophisticated, multimedia resources. This adds to the burdens that are being placed on the communications infrastructure. A consequence of this is that download latency is an increasing problem for human computer interaction to distributed systems. Delays in retrieval times can lead to frustration and error [9]. Users will abandon requests when the delays in retrieving data are perceived to outweight the benefits of the final information [6].

We have argued that the problems of download latency can be addressed by providing users with better information about the resources that they are requesting. In particular, it has been argued that the standard 'web' approach of hypertext links provides insufficient detail about the content and quality of video resources. In contrast, we have developed a web site that exploits cinematic presentation techniques to avoid these problems. Thumb-nail pictures provide information about the location and people in the films. More importantly, perhaps, they also provide information about the production quality of the video. The intention is to help the user to accurately assess whether the benefits of retrieving the film will be worth the delays that are imposed by high download latencies. We have conducted an initial qualitative investigation that supports our argument.

It is important to emphasise that much work remains to be done. Interfaces to distributes systems, such as Netscape and Mosaic, pose unique challenges for psychologists, computer scientists and sociologists. From a computing science perspective, network characteristics and the variable loadings on remote machines frustrate attempts to replicate or predict the temporal behaviour of underlying applications. From a sociological point of view, there has been little work to determine the impact that retrieval delays have upon group working practices [3, 8]. Finally, our investigations together with O'Donnell and Draper [11] suggest that more research needs to be done into the psychology of coping strategies for machine delays. Until these issues have been addressed it will be difficult for interface designers to address the threats posed by electronic gridlock.

Acknowledgements

Thanks are due to Scottish Televison for their help in producing the video footage that was used in the web site. Thanks are also due to the staff at the Hunterian Museum in Glasgow for their help and encouragement in setting up the project. In particular, Jim Devine, the Schools Liason Officer, provided many hours of advice about the content and design of the pages.

References

[1] T. Berners-Lee, R. Cailliau, and J.-F. Groff. The world wide web. *Computer Networks And ISDN Systems*, 25:454–459, 1992.

[2] S. Folkard. Circadian performance rhythms: Some practical and theoretical implications. In D.E. Broadbent, J. Reason, and A. Baddeley, editors, *Human Factors In Hazardous Situations*, pages 95–105. Clarendon Press, Oxford, United Kingdom, 1990.

[3] C.W. Johnson. Applying temporal logic to support the specification and prototyping of concurrent multi-user interfaces. In D. Diaper and N. Hammond, editors, *People And Computers VI: Usability Now*, pages 145–156. Cambridge University Press, Cambridge, United Kingdom, 1991.

[4] C.W. Johnson. The challenge of time. In P. Palanque and R. Bastide, editors, *The Second Eurographics Workshop On The Design, Specification And Verification of Interactive Systems*, pages 345–357. Springer Verlag, Berlin, Germany, 1995.

[5] C.W. Johnson. The economics of interface design. In K. Nordby, P.H. Helmersen, D. Gilmore, and S.A. Arnesen, editors, *Human Computer Interaction - Interact '95*, pages 19–25, London, United Kingdom, 1995. Chapman And Hall.

[6] C.W. Johnson. Time and the web: Representing temporal properties of interaction with distributed systems. In M.A.R. Kirby, A.J. Dix, and J.E. Finlay, editors, *People And Computers X*, pages 39–50, Glasgow, Scotland, G12 8QQ, 1995. Cambridge University Press.

[7] J. Kavanagh and C. Johnson. An analysis of download latency for video retrieval. Technical report, Department of Computing Science, University of Glasgow, Glasgow, United Kingdom, 1995.

[8] M. Klein. Capturing design rationale in concurrent engineering teams. *IEEE Computer*, 26(1):39–47, 1993.

[9] W. Kuhmann. Stress inducing properties of system response times. *Ergonomics*, 32(3):271 – 280, 1989.

[10] J. May and P. Barnard. Cinematography and interface design. In K. Nordby, P.H. Helmersen, D.J. Gilmore, and S.A. Arnesen, editors, *Interact'95*, pages 26–31. Chapman And Hall, London, United Kingdom, 1995.

[11] P. O'Donnell and S.Draper. How machine delays change user strategies. In C. Johnson, editor, *The Challenge Of Time*. Glasgow Interactive Systems Group, Glasgow, United Kingdom, 1995. G-95.1.

[12] S. Schofield. *The UK Internet Book*. Addison-Wesley, 1994.

[13] E. Young and C. Clanton. Film craft in user interface design. Technical report, ACM Press, Amsterdam, 1993. Tutorial notes from InterCHI'93.