Technical Report

UCAM-CL-TR-543
ISSN 1476-2986

Number 543

UNIVERSITY OF CAMBRIDGE

Computer Laboratory

Evaluating similarity-based visualisations as interfaces for image browsing

Kerry Rodden

September 2002

15 JJ Thomson Avenue
Cambridge CB3 0FD
United Kingdom
phone +44 1223 763500

http://www.cl.cam.ac.uk/

This technical report is based on a dissertation submitted 11 October 2001 by the author for the degree of Doctor of Philosophy to the University of Cambridge, Newnham College.

Some figures in this document are best viewed in colour. If you received a black-and-white copy, please consult the online version if necessary.

Technical reports published by the University of Cambridge Computer Laboratory are freely available via the Internet:

http://www.cl.cam.ac.uk/TechReports/

Series editor: Markus Kuhn

ISSN 1476-2986

Abstract

Large collections of digital images are becoming more and more common, and the users of these collections need computer-based systems to help them find the images they require. Digital images are easy to shrink to thumbnail size, allowing a large number of them to be presented to the user simultaneously. Generally, current image browsing interfaces display thumbnails in a two-dimensional grid, in some default order, and there has been little exploration of possible alternatives to this model.

With textual document collections, information visualisation techniques have been used to produce representations where the documents appear to be clustered according to their mutual similarity, which is based on the words they have in common. The same techniques can be applied to images, to arrange a set of thumbnails according to a defined measure of similarity. In many collections, the images are manually annotated with descriptive text, allowing their similarity to be measured in an analogous way to textual documents. Alternatively, research in content-based image retrieval has made it possible to measure similarity based on low-level visual features, such as colour.

The primary goal of this research was to investigate the usefulness of such similarity-based visualisations as interfaces for image browsing. We concentrated on visual similarity, because it is applicable to any image collection, regardless of the availability of annotations. Initially, we used conventional information retrieval evaluation methods to compare the relative performance of a number of different visual similarity measures, both for retrieval and for creating visualisations.

Thereafter, our approach to evaluation was influenced more by human–computer interaction: we carried out a series of user experiments where arrangements based on visual similarity were compared to random arrangements, for different image browsing tasks. These included finding a given target image, finding a group of images matching a generic requirement, and choosing subjectively suitable images for a particular purpose (from a short-listed set). As expected, we found that similarity-based arrangements are generally more helpful than random arrangements, especially when the user already has some idea of the type of image she is looking for.

Images are used in many different application domains; the ones we chose to study were stock photography and personal photography. We investigated the organisation and browsing of personal photographs in some depth, because of the inevitable future growth in usage of digital cameras, and a lack of previous research in this area.

Preface

Aspects of the work described in this dissertation feature in the following publications:

- K. Rodden. How do people organise their photographs? In *Proceedings of the BCS IRSG Colloquium*. British Computer Society Electronic Workshops in Computing, 1999.
- K. Rodden, W. Basalaj, D. Sinclair, and K. Wood. Evaluating a visualisation of image similarity as a tool for image browsing. In *Proceedings of the IEEE Symposium on Information Visualization (InfoVis'99)*. IEEE, 1999.
- K. Rodden, W. Basalaj, D. Sinclair, and K. Wood. Evaluating a visualisation of image similarity. In *Proceedings of SIGIR'99*. ACM, 1999. Poster.
- K. Rodden, W. Basalaj, D. Sinclair, and K. Wood. A comparison of measures for visualising image similarity. In *The Challenge of Image Retrieval*. British Computer Society Electronic Workshops in Computing, 2000.
- K. Rodden, W. Basalaj, D. Sinclair, and K. Wood. Does organisation by similarity assist image browsing? In *Proceedings of CHI 2001*. ACM, 2001.

Acknowledgements

Firstly, I would like to thank my supervisor, Jean Bacon, for giving me the opportunity and the freedom to work on whatever I was interested in, and for somehow managing not to lose patience with me while I figured out what that was.

I am especially grateful to four other people I have encountered in the course of this research (in a sort of chronological order):

- Matthew Chalmers, who succeeded in his stated aim of brainwashing me into becoming an HCI person, during my internship at UBS Ubilab in Zürich.
- Ken Wood, my supervisor at AT&T Laboratories Cambridge, who helped me to stay motivated by always being interested in my research ideas, and discussing them with me in detail; after every one of our meetings I felt more positive and encouraged.
- Wojciech Basalaj, for his MDS algorithms and his Visual C++ expertise, as well as many helpful discussions over tea.
- Rachel Hewson, who offered support and encouragement during the long months of writing up; there were many times when I had so much to explain that I didn't know where to start, and Rachel often helped me to decide.

I would have been unable to carry out any of my experiments without the assistance of the many participants, who of course must remain anonymous.

Alan Blackwell, Roz McCarthy, and Jonathan Pfautz all advised me on experiment design, and Gill Ward was my own personal statistics guru, giving me an initial tutorial on S-Plus, and patiently answering my many subsequent questions.

Will Hill of Anglia Polytechnic University provided insights into the use of images by graphic designers, and also recruited experiment participants for me from among his students. I am grateful to the organisers of the infodesign 99 conference for allowing me to set up an experiment there.

David Sinclair provided me with his image segmentation method and IRIS similarity measure, and gave me a quick tutorial on the implementation of image similarity measures. Yossi Rubner initially demonstrated to me how multidimensional scaling could be applied to image browsing, and subsequently provided me with image features to enable me to use his EMD measure.

Attending the Doctoral Consortium at ACM CHI'98 was the final part of my HCI brainwashing, and I would like to record my thanks to the organisers (Deborah Boehm-Davis, Clayton Lewis, William Newman, and Bonnie John) and the other participants, for their advice and feedback.

Jean Bacon, Wojciech Basalaj, Rachel DeWachter, Rachel Hewson, David Sinclair, and Ken Wood read and commented on all or part of this dissertation.

This work was financially supported by a studentship from the Engineering and Physical Sciences Research Council. AT&T Laboratories Cambridge provided generous additional funding under the CASE scheme, paid for my attendance at a number of conferences and workshops, and reimbursed me for the financial incentives I offered to my experiment participants. I received further support from the BFWG Charitable Foundation, and Newnham College.

Contents

Chapter 1

Introduction

Computer systems have been used successfully for many years for the storage, retrieval, and manipulation of textual and numerical data. Applications like word processors, databases, and spreadsheets are now widespread, simplifying tasks which were previously laborious and time-consuming. Massive improvements in speed and storage capacity mean that computers are now capable of manipulating far larger amounts of data, facilitating applications for the editing and browsing of new media such as digital images, video, and audio. Increasing use of these applications is resulting in large stored collections of multimedia, and the rapid growth of the World Wide Web means that such collections can be globally accessible.

For example, a graphic designer traditionally had to search for images by submitting a request to a trained intermediary at a photograph library, receiving a selection of hand-chosen prints via courier some time later. Now, many image libraries have digitised their collections and made them available via the Web, so that users can look through them directly. Some companies also provide smaller collections on CD-ROM, and newspapers and magazines often maintain their own digital archives. Users can simply download or copy digital images and incorporate them immediately into their work.

Image retrieval systems must support users in locating the images they want, quickly and easily. Many years of research in the field of information retrieval have resulted in effective techniques for automatically indexing text documents according to the words they contain, allowing users to retrieve them by entering text queries. Indexing images is more difficult, because they do not contain units (like words) that are both individually meaningful and easy to extract. The images in most collections are therefore annotated with descriptive text, such as captions and keywords, allowing the images to be indexed and retrieved using conventional information retrieval techniques. For example, a designer working on a tourist guide to Kenya might want to look for photographs of the country, to use as illustrations. If the image collection she is using has suitable annotations, she can simply look for the "Kenya" category, or enter the word "Kenya" as a textual query.

Creating annotations is very time-consuming, and the results are often highly subjective. Research in image retrieval has therefore concentrated on

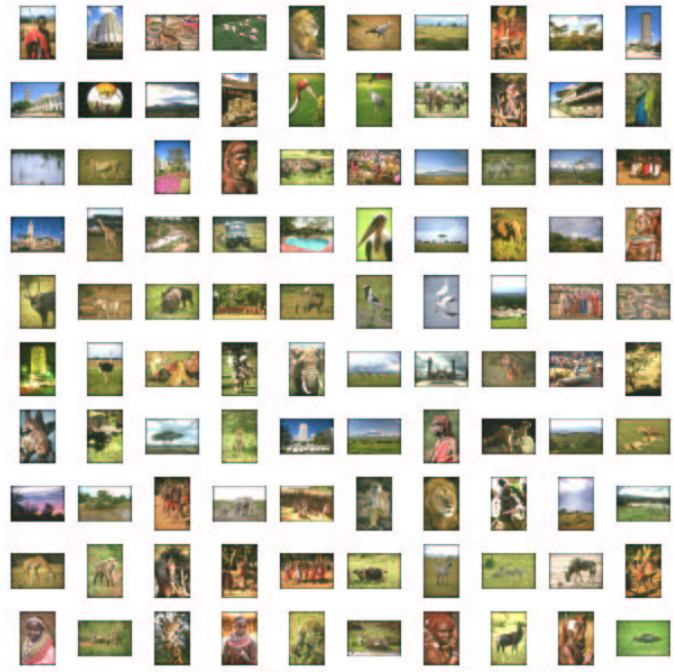
developing fully automatic, objective methods of indexing images. At present, however, these can extract only low-level visual features, such as colour and texture. In image retrieval systems that rely on these techniques, the user is expected to express her requirement in a visual form: this may involve selecting an indicative image, drawing a sketch, or specifying a colour distribution. Such queries only yield useful results if the desired photographs are very similar in appearance, which is not usually the case for semantic requirements. Consider photographs of birds, for example: there are many different species of bird, of different colours, shapes, and sizes, and they may appear against a wide variety of backgrounds.

In the results returned by a text retrieval system, a document is usually represented by its title and some sample text. The user can read this and decide whether the corresponding document is likely to be relevant to her requirement; if so, she can open and read the document itself. Images, on the other hand, are usually represented by miniature versions of themselves, called **thumbnails**, which can often be judged for relevance at a glance. A large number of thumbnails can be effectively presented to the user at once, commonly in a two-dimensional grid like Figure 1.1(a), which shows the contents of the "Kenya" category from a CD-ROM stock photograph collection. Having opened this category, the designer could simply browse through the thumbnails, recognising suitable images for the travel guide when she sees them.

Current image retrieval systems tend to present a grid of thumbnails in some default order, with no other support for browsing. For example, the images in Figure 1.1(a) are placed in the order in which they appear on the CD. It has been suggested that information visualisation techniques can support browsing of text document collections, by imposing some structure on them. In a visualisation, each document is usually represented by a point on the screen, and the points are arranged such that those representing documents that are similar (based on the words they have in common) are placed together. The same techniques can be applied to a set of thumbnails; for example, in Figure 1.1(b), the images of Kenya are arranged according to their visual similarity.

Compared to the default ordering, such an arrangement might give the designer a better overview of all of the different types of photographs of Kenya that are available in the collection. She could also add further criteria to her requirement without having to specify them explicitly, by simply shifting her gaze to the area of the screen where the images look suitable. In Figure 1.1(b), for example, the images of native people are mostly at the top right, with wildlife at the bottom right, and landscapes and buildings on the left. Of course, such arrangements are subject to the same limitations as visual queries, in that their definition of similarity is visual, not semantic (for example, not all of the images of birds are grouped together).

In the research documented in this dissertation, we set out to investigate the usefulness of similarity-based arrangements of images, via a series of evaluations. We concentrated on visual similarity, because these measures can be applied regardless of the availability of annotations, although in one exper-



(a) Default ordering, from the original category.

(b) Visualisation, based on visual similarity.

Figure 1.1: Two arrangements of the same set of 100 images of Kenya.

iment caption-based similarity was also considered. We employed a variety of evaluation methods, drawn from information retrieval, human–computer interaction, and more recent proposals that combine ideas from both fields.

In most cases our assumed image collection was a general-purpose photograph library (as used in publishing and advertising, for example); a number of previous studies have investigated how such collections are organised and searched in practice. We also chose to consider whether similarity-based visualisations might be useful for arranging personal photographs, especially given the inevitable future growth in the use of digital cameras. Very little research attention, however, has been devoted to investigating the organisation and browsing of personal photograph collections, and we therefore studied this area in some depth.

1.1 Dissertation outline

Chapter 2 provides an explanation of the background to this research, situating it in the context of information retrieval, image retrieval, and information visualisation. We discuss relevant work by others in these areas, and describe possible approaches to evaluation.

Chapter 3 describes the results of a theoretical investigation that was intended to establish whether arranging a set of images based on visual similarity tends to group together those of the same generic type, as well as compare a number of different visual similarity measures for the creation of such arrangements.

Chapter 4 presents the findings of two user experiments that aimed to discover whether people can locate required images more quickly in an arrangement based on visual similarity than in a random arrangement. In the first experiment, the task was finding a given target image within a set, and in the second experiment the task was finding a set of images to match a given textual description.

Chapter 5 describes two further experiments, this time aiming to find out if similarity-based arrangements were useful for a reasonably realistic task. Designers were used as the participants in both of these studies, and were asked to select images to accompany a given passage of text. Both visual and caption-based similarity are considered in this chapter.

Chapter 6 focuses on the application area of personal photography, considering both how people organise and browse their photograph collections at present, and how they will do so when the use of digital cameras is widespread. A group of volunteers were given digital cameras and prototype photograph organisation software, and were interviewed at the beginning and end of a usage period of six months. The participants were also asked for their opinions of similarity-based arrangements of their own photographs.

Chapter 7 summarises the main findings of this research, bringing together themes common to more than one chapter, and then discusses possibilities for further work.

Chapter 2

Background and related work

In this chapter, we introduce the three areas of computer science to which the research described in this dissertation relates — information retrieval, information visualisation, and image retrieval — and situate it in the context of these. Our focus is on evaluation, and we therefore pay particular attention to previous work that has been supported by user experiments, and examine different approaches to evaluation, including those used in human–computer interaction. We close the chapter by outlining the framework within which our own research was carried out.

2.1 Information retrieval

Information retrieval systems are used to automatically index collections of text documents: the individual words are extracted from each document, and a list of every unique term in the collection is created, where each term has pointers to all of the documents that contain it. The user issues a query by entering a few terms that characterise her requirement, and the system matches these against its list, returning a set of documents containing the user's terms. The system may weight the terms according to their frequency (within the individual documents or the collection as a whole), and this allows the results of the query to be ranked in order of their estimated relevance. The most widely visible application of these systems is in World Wide Web search engines, which index the full text of billions of documents.

Users can also browse a collection, or a subset such as the results of a query, by looking through it in whichever form it is presented; for example, by following links between Web pages. It is usually assumed that if the user has a requirement in mind, she will specify it with a query, and if she does not, she will browse, looking through the collection and recognising interesting items when she sees them. However, this linkage is not inevitable [9], and it is easy to imagine situations where the user might choose to browse through documents in order to satisfy a requirement, as an alternative to issuing a query: for example, her requirement may be vague, or difficult to express.

Researchers have proposed detailed taxonomies of browsing strategies [22] and information seeking strategies [9, 76] in general, but in this dissertation a

simpler classification will suffice. In information retrieval, the term **searching** is conventionally used synonymously with querying. However, we shall use it with its broader meaning of looking with a requirement in mind, which may be accomplished either by issuing a query or by **directed browsing**. We shall specifically refer to browsing with no conscious requirement as **undirected browsing**. Of course, browsing may become more or less directed as the user proceeds.

If a collection has a noticeable structure, where related items are grouped together, this should make it easier for the user to decide where to start browsing, and also help her to find other relevant items after locating the first one. For example, the non-fiction books in a library are classified according to subject matter. Information foraging theory [89] suggests that, when searching, humans are drawn towards groups of relevant items, in an analogous way to animals attempting to locate dense patches of food in the wild.

Information retrieval researchers have experimented with automatic clustering of entire document collections, grouping together those which are similar, according to the words they have in common. It was once thought that it would be more effective to match a query against whole clusters, instead of each individual document in the collection, but this has never actually been proved [125]. The **cluster hypothesis** states that "closely associated documents tend to be relevant to the same requests" [118, Ch.3], or alternatively "relevant documents tend to be more similar to each other than to non-relevant documents" [52].

The Scatter/Gather system [32] has shown that clustering can be employed to support browsing. It partitions a document collection into clusters, and then presents a list of these to the user, with representative terms and document titles from each. The user can select any number of the clusters, and the system then dynamically re-clusters their contents, presenting her with a new list. Pirolli and his colleagues [90] found that this system helped its users to get an idea of what the collection contained, although when searching, querying was more effective than pure directed browsing as a means of locating relevant documents. A later study by Hearst and Pedersen [52] found that, in theory, clustering the results of a query (rather than the whole collection) using Scatter/Gather should be helpful for locating the relevant documents, as most of them are usually placed in the same cluster, in accordance with the cluster hypothesis.

Kural, Robertson, and Jones [63] noted that the usefulness of the clustering of query results is dependent on how well the clusters are represented to the user, and they described the trade-offs involved in producing a representation that is both compact and understandable. Hearst and Pedersen's experiment participants tended to make the cluster with the highest proportion of relevant documents their first selection, suggesting that their chosen representation (a list of terms that appeared frequently in the cluster and relatively infrequently in the rest of the collection, plus the titles of documents near the centroid of the cluster) was effective. Kural and her colleagues used a similar representation, and studied it in more depth, finding that it was "not always adequately informative", varying in quality between different clusters, and sometimes mis-

leading.

In terms of information foraging theory, the quality of a cluster representative determines the amount of **information scent** it provides; this is defined as "the strength of local cues [...] in providing an indication of the utility or relevance of a navigational path leading to some distal information source" [89]. Furnas [45] mentions that it is difficult to ensure that any large collection has good information scent, meaning that the best search strategy is likely to be using a query to find a good starting point for browsing.

2.2 Information visualisation

Card, Mackinlay, and Shneiderman [19, p6] define information visualisation as "the use of computer-supported, interactive, visual representations of abstract data to amplify cognition". From their introductory text, and that of Spence [113], it is clear that the most common use of information visualisation techniques is producing graphical representations of tables of numeric data. Different variables can be plotted against each other to allow analysts to identify relationships between them, facilitating data mining.

These techniques can be combined with those of information retrieval to produce visualisations of document collections, as in the Bead [15, 21] and SPIRE [126] systems. Text documents have no obvious coordinates for convenient representation on a screen, and so these systems attempt to place points in a two- or three-dimensional space such that their relative proximity reflects the similarity of the corresponding documents. Usually, clusters are not explicitly formed or identified, but users can detect from the arrangement of points whether clusters exist, how they relate to each other, and how closely placed (and therefore similar) the documents are. In contrast, the Scatter/Gather system creates a fixed number of clusters, with no representation of how they relate to each other, or how closely related the documents within a particular cluster actually are.

A common method of measuring the similarity between documents involves representing each document as a vector [104] with one entry for each unique term in the whole collection. The value of each entry in the vector is greater than zero if the corresponding term is present in the document, and zero if it is absent (the frequency of the term within the document may be used, for example). There are a number of ways of measuring the similarity between these vectors in their high-dimensional space, such as taking the dot product (see Appendix B for an example). Some method of dimensionality reduction must be used, to produce a low-dimensional configuration of points where the inter-point distances reflect the original high-dimensional similarities. One such method is **multidimensional scaling (MDS)** [10], which is explained further in Appendix A. Basalaj [7] implemented a number of MDS algorithms and compared them to related techniques such as principal components analysis (PCA), finding that visualisations created with MDS were generally quantitatively and qualitatively superior.

As Wise and his colleagues [126] have noted, a visualisation may help peo-

ple to make sense of a large collection of documents while having to read fewer of them, thereby ameliorating the problem of information overload. In this research, however, we were primarily interested in the use of visualisation techniques as a means of assisting directed browsing, to help people locate items that are relevant to a requirement. A number of different document visualisation systems have been proposed, but few have been evaluated; as Hearst [51] has stated, "although intuitively appealing, graphical overviews of large document spaces have yet to be shown to be useful and understandable for users." It is being increasingly acknowledged that in some cases, a visualisation may not be any more useful than a default representation, such as a simple list of the items [107]. As a result, interest in evaluation is growing [24].

Leuski and Allan [68] found in a simple experiment that "users have no difficulty grasping the idea of spatial proximity as the metaphor for inter-object similarity". They created visualisations of sets of text documents, where each set contained the results retrieved for a particular query, and the relevance of each document to the query was already known. The documents were represented as white spheres that (after a short delay) turned green or red when clicked, depending on whether the original document was relevant or non-relevant. The top-ranked relevant document was already coloured green, and the participants were simply instructed to try to find all of the other green spheres, as quickly as possible, while avoiding the red spheres. They were also told that spheres of the same colour tended to be placed in close proximity to each other, but not always. This allowed the experimenters to isolate the task of understanding a proximity-based visualisation from that of making relevance judgements. They produced a model of the optimal strategy to use when searching a visualisation, and a model of simple traversal of a ranked list, and found that the participants' actual performance was only slightly worse than the former, and better than the latter. They proposed that the results of a query should be presented as both a ranked list and a visualisation, displayed simultaneously and coupled together, so that, for example, when an item is selected from the ranked list, it is highlighted in the visualisation. Later, they incorporated this idea into a system called *Lighthouse* [69]. When users make relevance judgements, the system reflects this feedback by colouring the other query results (in both views) in shades of green or red, to indicate its prediction of their relevance, based on their proximity to those items already judged. Conventional information retrieval systems react to relevance feedback by re-ordering the ranked list, or presenting a new set of results.

Commonly, most of the screen space in a visualisation is devoted to graphics, and "often labels are entirely missing and users have to peck at graphical objects one at a time" [40] in order to pop up information about them. Such neglect of text labels seems odd, given that the visualisation is supposed to be representing the underlying data. For large collections, it is impossible to display labels for every item simultaneously, especially as the labels are larger than the points, which is why many systems require the user to click on an item or hover over it with the mouse pointer in order to see its label. For example, Chen and Czerwinski [23] created a VRML visualisation of 169 abstracts of academic papers, where each point was permanently labelled only

with the initials of the corresponding paper's author; running the mouse over a point displayed the paper's title, and clicking on it opened the abstract in a frame next to the visualisation. They carried out an experiment where participants were asked to locate documents on a given topic, and found that groups of relevant documents were often missed, especially those at outlying points in the visualisation. The results of their post-experiment interviews suggested that a text-based query facility (like that provided in Bead [15, 21]) would have been helpful, to highlight the documents containing a particular word and thus provide possible starting points for browsing. Extra navigational cues such as cluster labels would also have helped the participants to understand the structure of the visualisation. Bead [15, 21] keeps track of commonly used query terms, and labels the relevant clusters of points with them. It also randomly pops up object labels for a few seconds each, to help reveal the structure of the visualisation without any effort on the part of the user. Fekete and Plaisant [40] have proposed a space-efficient method of labelling all of the points in the user's current region of interest.

More discrete visualisations can be created with self-organising map (SOM) algorithms; these use a neural network to map objects into a lattice, which is usually a two-dimensional grid, for easy display. More than one object can be mapped to the same grid cell. Lin [71] has demonstrated the use of a SOM to represent a set of text documents. His method uses the vector model, and each grid cell has a combined vector that is adjusted as new documents are mapped to it. The highest-weighted term in the vector is chosen to represent the cell, and cells with the same representative term are connected, forming explicit clusters, which are then labelled with their common term. Other terms can be mapped onto the grid by finding the cell where the chosen term has the highest weight. The resulting map is thus a thematic overview of the set of documents, without explicitly showing their relationships; the user can choose to display the document titles within a cluster by selecting it. Lin conducted an experiment [70] comparing four arrangements of the same set of 133 documents. One was randomly ordered, and the other three were based on similarity: one created using a SOM, and two created manually (by different people). In each trial, the participant was asked to locate a given document title from within the set. The results showed that the participants were faster with each of the similarity-based arrangements than they were with the random arrangement, and that there was no significant difference between the SOM arrangement and those created by hand.

It is also possible to construct a SOM with multiple levels; Chen and his colleagues [26] conducted a simple experiment to compare the Yahoo! Web directory's "Entertainment" category to a multi-level SOM of the 110,000 documents in the category. They found that for an undirected browsing task (where the participant was simply asked to look for a site that she found interesting), the success rates were roughly equivalent, but for a directed browsing task (locating a particular site), Yahoo! was definitely superior. This was probably because its organisation is more predictable: a conventional hierarchy, with the entries going from generic to specific as the user descends through the levels (for example, "TV" at the top level, with individual programme names below

it), and the entries within a level are ordered alphabetically. However, in the SOM, generic and specific items could appear at the same level (for example, "Star Trek" was at the top level, with more generic labels like "Music" and "Film") and each level was arranged associatively, with similar items grouped together. In Lin's experiment, this type of organisation was more helpful than a random arrangement when looking for a particular item, but he did not compare it to an alphabetical ordering.

2.3 Image retrieval

As Eakins and Graham mentioned in their survey [36], images are used in a large number of application domains, including medicine, architecture, satellite imaging, art history, graphic design, publishing, advertising, and home entertainment. In the work described in this dissertation we shall consider general photographic images, which are mainly used in the last four of these domains. Before it was possible to store photographs in digital form, they would be held as prints, negatives, or slides, perhaps with a card index (or more recently, a conventional database) containing associated keywords or classification. Some collections still rely on this method of organisation, but increases in the speed and storage capacity of computers, and the availability of high quality scanners and digital cameras at reasonable prices, have meant that, in many cases, digital photographs are now being used in these domains. It is important to have effective methods of accessing them, and therefore fully computer-based image retrieval systems are being used more widely.

2.3.1 Image content

Automatic indexing of text documents is reasonably straightforward, because the words they contain are both individually meaningful and easy to extract. The content of images, however, cannot be divided into such convenient units. Systems that can index and retrieve images using only whatever can be automatically extracted from their pixels are known as **content-based** image retrieval systems.

An image's content may be described on several different levels, and the classification we use in this dissertation is a simplified version of that defined by Eakins and Graham [36]. Our description of the levels is made with reference to the photograph in Figure 2.1.

- The **primitive** level: the colours, textures, composition, and the simple shapes present. Our example photograph contains lots of green, with some yellow and blue shapes, a small white circle, and so on.

- The **generic** level: the types of objects and activities depicted. In our example, the objects are people, grass, and a ball, and the people are playing in a football match (it is interesting to note that the original caption does not mention this). Shatford Layne [108] has noted that several generic terms may apply to the same object: a mallard is also a duck,

Figure 2.1: A picture from a football match. The original caption assigned by the photograph agency was "12 Jul 1998: Zinedine Zidane of France heads the ball past Ronaldo #9 and Dunga #8 of Brazil to score the second goal as France defeats Brazil 3-0 to win the 1998 World Cup final at Stade de France, St Denis, France. Photo by Allsport."

which is also a bird, for example. All of these should be included at this level.

- The **specific** level: the names of people, places, landmarks, events, and so on. The names of the players in our photograph, such as Zinedine Zidane and Ronaldo, would be included at this level, as well as the fact that the match is the 1998 World Cup Final, where France played Brazil.

- The **abstract** level: this involves interpreting the overall meaning of the image, such as the mood or feeling that it evokes. Our photograph might be described as depicting action and excitement; this goal gave France a decisive lead in the match.

The user may have a requirement at any of these levels, and so an ideal image retrieval system would be capable of automatically identifying content at all of them.

Primitives are relatively easy to extract, and are used by all of the current content-based image retrieval systems (such as QBIC [42], Virage [48], and VisualSEEk [111]). A query at this level is purely visual; for example, the user may attempt to sketch an image that is representative of her requirement, or specify what proportions of particular colours she would like the results to contain. The most common paradigm, however, is **query-by-example**, where the user chooses one of the images from the collection, and the system returns a set of images that are similar (at the primitive level). Colour is normally used as the main determinant of similarity, and so two images may be defined to be similar if they contain a similar distribution of colours. More advanced image processing techniques allow an image to be segmented into a number of coherent regions, enabling users to construct queries by selecting individual regions, rather than a whole image. Region outlines can be superimposed on an image to assist selection, as in the Blobworld system [20].

Researchers in the fields of computer vision and artificial intelligence have worked for many years on developing automatic image understanding techniques, many of which have been applied to image retrieval. At present, it is possible to identify features at the generic level (and sometimes the specific level), but this depends on whether they have distinctive primitive characteristics: for example, sunsets are easy to identify, because the combination of bright orange and black in a particular composition does not tend to occur in other types of photograph. Meaningful feature recognition can only be done reliably within very limited domain areas, where domain-specific heuristics can be applied (such as flower patents [33] and aerial photographs [94]). Reliable domain-independent recognition and interpretation has been described as an *AI-complete* problem [97], meaning that finding a solution requires solving strong artificial intelligence in general.

The problems are readily apparent upon looking at the example image. At the generic level, current systems would be able to identify that there are several faces (and therefore people) present, and the grass could be identified by its colour and texture. It is possible that a system could guess (from the presence of these features, plus a ball-shaped object) that some sort of sport is being played. At the specific level, face recognition may allow some of the people to be identified by name, but only if the system already has a record of their characteristic features; many of them are not facing the camera directly, which would make recognition more difficult. Even a human would not be able to infer some of the details from the image alone: she would need a good knowledge of football to tell that this photograph was taken at the 1998 World Cup Final, and it would be impossible for her to deduce contextual information (for example, the fact that this header resulted in a goal, France's second of the match), which can only be obtained through annotation at the source.

In research papers about content-based image retrieval systems, the typical sample query is for sunsets (such as [48, 111]), which tends to produce impressive results because (as we have already noted) sunset images are usually similar at the primitive level. However, in many other cases, the primitive similarity between two images does not correspond to their semantic similarity, at the generic, specific, or abstract levels. Users may therefore be disappointed in the results of a query-by-example (whether their exemplar is an existing image or a sketch), if they are expecting that the system's definition of similarity will be like their own. Even if a system was capable of measuring similarity at higher levels, it would be difficult to automatically infer the user's intended level. For example, a user may select the image in Figure 2.1 as her exemplar because she wants images that contain a combination of green, blue, and yellow (primitive), people (generic), Ronaldo (specific), and so on.

There have been very few user evaluations of content-based image retrieval systems. Jacobs, Finkelstein, and Salesin [59] tested their query-by-sketch system, and showed that their participants' sketches from short-term memory of given images were good enough to place the targets in the top 1% of the ranking more than half of the time. However, it is not clear whether users will actually be willing to make the effort to create such sketches in real situations (particularly if they are not confident about their artistic ability), or whether

query-by-sketch can be effective for anything other than a requirement for a (recently seen) known image. With a generic requirement, for example, it is likely to be easier for the user to browse the collection and recognise relevant images, rather than attempt to draw a representative sketch, as we shall see in Section 2.4.2.

2.3.2 Image annotation

Because of the difficulty of automatically extracting image content at anything other than the primitive level, it is clear that, for the foreseeable future, users will have to rely on manually assigned annotations. These may take the form of a free text caption, or keywords from a controlled vocabulary [95], and often both are used. Conventional text retrieval techniques can then be used to index the images, allowing the user to issue text queries. A controlled vocabulary can also be used as a basis for browsing: the user can select a keyword and see which images it has been assigned to. Applying such keywords is thus equivalent to categorising the collection.

However, the annotation process is time-consuming and difficult, especially if the images are complex, ambiguous, or abstract; it is impossible for the annotator to be completely comprehensive, and many photographs are just hard to describe in words. The annotations will always reflect the subjective judgement and prior knowledge of the annotator, which may not be shared by those searching the collection. In particular, interpretation of the mood or feeling evoked by the image (the abstract level) is likely to be highly subjective. Different people will notice different things about the content of an image, and have different opinions about what the most important details are. For example, only three of the players have been named in the caption of Figure 2.1, so a query for the name of one of the other players visible would not return this image in the results. The annotator simply cannot anticipate all of the potential current (and future) requirements for which a photograph may be relevant.

Furnas and his colleagues [46] identified what they called *the vocabulary problem* in the context of command naming: they found that, in five different application domains, it was unlikely that two people would spontaneously use the same term to describe a given concept. There also tend to be "low levels of interindexer consistency" [108] with images. For example, Enser [39] asked 18 people to supply annotations for a photograph of the Eiffel Tower, and found that only 14 had used "Paris" and only 12 had used "Eiffel Tower"; no other term was used by more than half of the participants. It is also quite likely that annotations made by the same person will be inconsistent between images. Tools to support the annotation process may help to enforce consistency; a controlled vocabulary is particularly useful, as it allows the annotator to tag images with appropriate keywords from a checklist, rather than having to enter all of the information in a free text caption.

Despite these problems, the use of annotation is still much more likely to provide users with useful semantic information than any automated approach, and it is therefore usually worth the effort for the owners of the collection.

News agencies, for example, will record names, dates, places, and events, as in the caption of Figure 2.1, as soon as possible, because this contextual information is vital to allow the photographs to be used by journalists.

Other types of metadata may be available, depending on the origin of the images. Much of it is not content-descriptive, for example the date and time that the photograph was taken (which can be recorded automatically by the camera), or the name of the photographer.

2.4 Supporting image browsing

The user of an image retrieval system can search for images matching a requirement by issuing a visual query, or a text query when annotations are available. We have already noted that the results of visual queries are often disappointing to the user, because they rely only on primitive image attributes. Text queries are not always appropriate because, as we discussed at the beginning of this chapter, the user's requirement may be vague, or difficult to express in words. In such situations, directed browsing is a preferable search strategy. This allows the user to recognise a relevant image when she sees it, rather than having to specify which objects it contains, or what it looks like.

The content of an image can be taken in at a glance, and a large number of them can be presented to the user at once by using thumbnails, which are very good representations of the original images because they are simply miniature versions of them. On the other hand, text documents have to be read in order to judge their relevance, which takes some time, and because of their length they are usually represented in query results by a surrogate (perhaps a title and the first few words), which may not provide enough information for the user to decide whether to open the full document. Thus, directed browsing is probably an easier and more efficient strategy to use in image collections than in text collections, and should be supported by image retrieval systems.

Shatford Layne [108] has noted that, because abstract attributes (like mood) are so subjective, the best image search strategy is probably to issue a broad text query and then browse through a large number of results, rather than attempt to construct a focused query that returns very few results. Any secondary criteria can be applied while browsing the thumbnails, instead of being specified in the query.

So far, a number of researchers have described systems which attempt to provide explicit support for image browsing. We have roughly divided their approaches into three categories:

Relevance feedback includes those systems that have adapted the query-by-example approach, so that when the user selects an image, it is not regarded as a one-off query, but as further evidence about her requirement, in combination with her earlier selections.

Clustering includes those systems that have applied clustering techniques to image collections, to impose a top-level structure; we have already noted

that this kind of automatic organisation can be helpful to users when browsing text collections.

Visualisation includes those systems that have applied information visualisation techniques to images, usually assuming that the user has already restricted the collection in some way (for example, by selecting a category, or issuing a query) so that only a subset remains. All of the images in the subset are displayed simultaneously.

2.4.1 Relevance feedback

In both PicHunter [31, 87] and the system described by Campbell [18], the system invites the user to select from a presented set of images, and treats her response as an indication of her requirement. In the next iteration, the system displays the set of images that it regards as most relevant, based on all of the selections the user has made since the start of the search.

Both systems hide the underlying information retrieval system from the user, and measure similarity according to the image annotations (although PicHunter can also be set up to use visual similarity, or a combination of the two). The collection used by Campbell has annotations written in French, but the user can still make use of the similarity between them, without having to understand French herself.

PicHunter is designed specifically for searches where the user is looking for a particular target image, and evolution of a search is not well supported: the user has to explicitly terminate her search and start a new one, in order to reset her relevance feedback. Campbell's system is expressly designed to support evolving requirements, by adjusting the weight assigned to each selection, so that the most recent selections have the highest weight (this is known as the Ostensive Model). It also has a more innovative interface design, which provides an explicit graphical representation of the user's browsing history, allowing her to see her path through all of the images she has selected so far, and go back to any of them.

Both systems present only a small number of new images to the user after each selection; PicHunter always shows nine, in a simple 3×3 grid. It is likely that in many cases none of the new images will seem especially relevant or interesting (although Campbell's system at least allows the user to retrace her steps).

PicHunter has been subject to several experiments. One found that participants could reach a given target in fewer iterations when the system based its selections on annotation similarity alone, suggesting that their participants' definitions of image similarity were based more on semantics than appearance. In another experiment, the participants were asked to find an image "very similar" to the given target, where a version of PicHunter using visual similarity was compared to a version which displayed nine images at random. The participants took fewer iterations with PicHunter, as expected, but the difference was not overwhelmingly large. The collection used in the experiment

contained large groups of semantically similar images, so this result is not especially surprising.

Campbell conducted an experiment comparing his Ostensive Model to one where all selections were weighted equally, and did not find a significant difference between them. However, the participants were searching for images matching a fixed requirement, when his model is designed to support evolving requirements.

It is difficult to judge the potential effectiveness of these systems, because neither was compared to a more conventional system, where users could simply enter a text query (or a visual query, if assuming that no annotations are available) and browse through the results as thumbnails.

2.4.2 Clustering

Two systems which are alike from a user's point of view are Similarity Pyramids [27] and Filter Image Browsing [121]. Both of them cluster the collection according to visual similarity, and at each iteration they present the user with a fixed number of thumbnail images, each of which is a representative of a cluster. The first system uses hierarchical clustering, and the user must select one image from the display to descend to the next level of the fixed hierarchy, where she is presented with a new set of representative images. The second system applies the Scatter/Gather principle to image collections, so the user can select more than one of the clusters, and the system will dynamically recluster the images they contain, again presenting her with a new set of cluster representatives. In both systems, the collection is progressively filtered down, because the next selection can be made only from within the selected clusters, rather than the whole collection. This means that if the user's requirement evolves while searching, she may have to move backwards in order to bring images from rejected clusters back into consideration.

Both of these systems cluster the collection according to visual similarity (at the primitive level), which does not always relate very closely to semantic similarity (at the generic, specific, or abstract levels). This may make it difficult for the user to decide which cluster contains images matching her requirement; depending on her expectations, the image chosen to represent a cluster may not be a good indicator of its content, providing a misleading scent [45].

The approach taken by CHROMA [64] is perhaps the most realistic for a system that is based only on visual similarity: it simply uses colours as cluster representatives, which is less likely to give the user a misleading scent. Images are clustered into a two-level hierarchy according to ten basic colours; the system's interface looks like the Windows file explorer, with the colours displayed in a tree on the left. When the user selects one, the images with that dominant colour are displayed as thumbnails on the right, and its part of the tree expands to show the second level of the hierarchy, which is a subset of the same group of ten colours. The system's creators specifically chose to use only two levels, to prevent the user becoming lost in the hierarchy. They conducted an experiment comparing this browsing tool to a query tool that allowed users to construct a sketch using squares of the ten basic colours. When

the participants were searching for a given image, there was no difference in search time between the tools (about 4 minutes on average). When searching for an image "on a general theme, for example, the countryside, or a desert landscape" (generic, according to our classification), the query tool was significantly slower (taking about 7 minutes on average). Thus, it was more difficult for the participants to use the query tool when they did not know precisely what the image they wanted looked like. The participants rated the browsing tool significantly more highly than the query tool in terms of "fast and efficient access to the images" and finding the required image without difficulty. The collection used in the experiment, however, contained only 1000 images, and it would be interesting to know how quickly the participants would have completed the tasks with some baseline interface, such as flipping through 10 screens of 100 thumbnails.

This result suggests that, when there are no annotations available, directed browsing in an image collection that has been clustered according to visual similarity should be more effective than issuing a visual query. In contrast, text queries were more effective than cluster-based directed browsing in the original Scatter/Gather system, as we noted in Section 2.1. Chen and her colleagues [25] have created a "multi-modal" version of Scatter/Gather, which they applied to a collection of images from the World Wide Web, and any text associated with them from the original pages. The collection is initially clustered according to the text, as in the original Scatter/Gather, and for the subsequent clustering steps, the user can choose to use either textual or visual similarity. The user can also expand the current working set, which adds images that are (either textually or visually) similar to those that it already contains, bringing back images previously discarded. This is an interesting approach, but the examples given in their paper are not particularly compelling, perhaps because of the nature of the collection.

2.4.3 Visualisation

Combs and Bederson [28] reviewed a number of commercial image browsing systems, and found that a two-dimensional grid of thumbnails was the most common way of presenting a set of images. Then, they conducted an experiment comparing one of these systems to three alternative designs: one that had a zooming facility instead of a scrollbar, and two with new three-dimensional interfaces (built using VRML). The participants were asked to locate a given image within a set of 25, 75, or 225, and were both faster and more accurate with the two-dimensional browsers than the three-dimensional ones. The authors noted that the zoomable system, which allowed more images to be simultaneously visible than the other three, "received many comments suggesting the ability to group images in clusters by content".

Depending on the availability of annotations, and how the user wishes to search, the set of images may be

- the results of a text query
- the results of a visual query

- the contents of a category
- the contents of some default subset

To issue a text query, or select a category, the collection must first be suitably annotated. Without annotations, the user must rely on visual queries, or if she simply wishes to browse the collection, she can request that the system presents her with some subset of it, perhaps randomly chosen. Query results would normally be ranked in order of their estimated relevance to the query, and in the other two cases the the thumbnails would be presented in some default order. Information visualisation techniques (as introduced in Section 2.2) can be used to arrange a set of images according to their similarity, either visual or text-based (if annotations are available). This might help the user to narrow the set down to whichever images are relevant to her current requirement.

Most of the systems we shall describe in this section use MDS, or one of its close relations, such as PCA. As with other similarity-based visualisations, the goal is to create a low-dimensional configuration that accurately reflects the relative similarities of the images. These can represented on a screen as dots, as in a conventional visualisation [73, 101], where the user can run the mouse over a dot in order to see the corresponding image. However, this is like displaying a visualisation of a textual document collection without using any labels; thumbnails are excellent representatives of the original images, and can be displayed instead of dots (as in Figure 1.1(b) on page 15).

We are aware of four systems that use these techniques to arrange thumbnails according to visual similarity:

- In the system described by Rubner, Tomasi, and Guibas [103], the user can issue an initial query by specifying approximate colour percentages, and the system arranges the results according to visual similarity, using MDS. Their paper explains that "while more traditional displays list images in order of similarity to the query, thereby representing n distances if n images are returned, our display conveys information about all the $\binom{n}{2}$ distances between images." The user can refine her query by selecting an individual thumbnail to use as an example, or by indicating a whole area of the visualisation.

- Hiroike and his colleagues [53] describe a system in which the user first selects a small set of example images, and the system arranges these in a three-dimensional space according to visual similarity, using PCA. It then retrieves a set of similar images for each, and clusters them around their corresponding example images. The system can automatically cycle between different possible projections, to give the user different perspectives on the set.

- The El Niño system [105], described by Santini and Jain, arranges a set of images according to visual similarity, in two or three dimensions. The user can give relevance feedback by marking images in the display, and

also by moving images around to demonstrate her definition of similarity in the context of her current requirement. The system can then retrieve more images, and create a new arrangement that reflects the user's feedback. It is difficult to know how effective this approach is likely to be in practice.

- The most recently developed (and probably the most complete) system is CIRCUS [88], which aims to closely integrate browsing and querying facilities. The system arranges the whole collection according to visual similarity, using PCA, and also hierarchically clusters it. This allows a semantic zoom facility to be implemented: as the user descends the hierarchy, only the cluster representatives at the current level are shown in the arrangement (the others are hidden), so when she reaches the leaf level, all of the images are visible. The user can adjust the current view by panning and zooming around the arrangement. When she issues a query, she can view the results in a conventional ranked list, or as part of the visual arrangement (again, non-matching images are simply hidden). Its creators carried out an exploratory user evaluation, but did not report the results in any detail.

The Design Galleries system [79] uses MDS to arrange different versions of the same image, rather than a pre-existing set of heterogeneous images: it is intended to support the user in choosing parameters when rendering computer graphics, by automatically generating possible variations of the same scene, and organising them according to the similarity of their parameter values.

Some systems use a self-organising map (such as [49, 61]) to arrange an image set in a grid, according to visual similarity. PicSOM [61] uses a hierarchical variant of the SOM algorithm, and thus its interface superficially resembles that of the Similarity Pyramids system described in the previous section.

ANVIL [102] is the only system (of which we are aware) that uses annotation similarity to arrange a set of images. The user issues a text query, and the top-ranked image is shown in the centre of the screen, with other results surrounding it, at distances that reflect their estimated relevance to the query. These other images are also clustered according to the mutual similarity of their captions. In a pilot study, this type of arrangement was compared to a conventional ranked list. Participants completed eight tasks, in each of which they were required to find two images suitable for a given requirement, which was either generic (such as "countryside") or abstract (such as "friendliness"). The system's creators expected that the similarity-based arrangement would provide better support for browsing of query results. They found that participants took less time (on average) to complete a task using the similarity-based arrangement, and also entered fewer text-based queries while using it, but in terms of participants' satisfaction with their choices, there were no significant differences between the interfaces.

2.5 Approaches to evaluation

As we have noted, very few researchers have attempted to evaluate information visualisations or image browsing systems. No specialised evaluation methods exist in either of these areas, and so when evaluations have been carried out, methods have usually been borrowed from information retrieval (to test the underlying system) and human–computer interaction (to test the user interface aspects). We examine both of these approaches in this section, as well as recent proposals that the two should be combined, to facilitate comprehensive evaluation of *interactive* information retrieval systems.

McGrath [80] has classified research strategies into four main categories:

Theoretical strategies include the formulation of general theories about existing empirical evidence, providing a basis for further work. This category also covers computer-based simulations that have no user involvement, with all of the parameters specified in advance.

Experimental strategies include controlled experiments, usually carried out in a laboratory. These experiments must be carefully designed to ensure the validity of the results. The researcher has to decide what to control, what to measure, and how to measure it, and the results can be statistically analysed to establish the size and significance of any effects. An *experimental simulation* is a laboratory experiment where the researcher has attempted to make the setting and context as realistic as possible.

Respondent strategies involve using surveys and questionnaires; such studies may be large scale, with a sample of people carefully selected to be representative of the target user group, or they may be smaller exploratory studies resembling informal experiments. Participants may be asked questions about themselves, or about their opinions of presented stimuli.

Field strategies involve studying behaviour in a natural setting, such as within an organisation; this may take the form of simple observation, as in ethnography, or something more obtrusive like manipulating an element of the system in use and recording what effect it has.

Because all of these have their own strengths and weaknesses, McGrath advocates the use of multiple strategies: “credible empirical knowledge requires consistency or convergence of evidence across studies based on different methods”. He also discusses the trade-offs that must be made between generalisability, control, and realism when selecting a strategy. For example, laboratory experiments give the researcher a high degree of control, but are not as realistic as field studies, and neither are as generalisable as a survey with a large and representative sample of respondents.

2.5.1 Traditional information retrieval

Information retrieval evaluations are normally based on **test collections**, the best known of which are those compiled as part of the TREC project [50]. A test collection consists of a large number of documents, a set of queries, and relevance judgements indicating which of the documents are relevant to each query. In an evaluation, a query is submitted to all of the systems being compared, each of which returns a list of results ranked according to their estimated relevance to the query, up to a cut-off point. The performance of each system for this query can then be assessed by calculating **recall** (the proportion of the relevant documents in the collection that the system has actually retrieved) and **precision** (the proportion of the documents retrieved that are relevant). Commonly, these measures are aggregated across all queries, and then plotted against each other on a **recall–precision graph**, with one line for each of the systems being compared.

It is often thought that the results of these evaluations cannot be subjected to statistical tests, because recall and precision are fundamentally discrete, not continuous. Hull [55, 56] has shown, however, that aggregated versions of recall and precision approximate continuous variables, and can fit the assumptions of either parametric tests or the less powerful non-parametric tests. Tague-Sutcliffe [117] also discusses the use of statistical testing, as part of a wider description of the different stages involved in conducting an information retrieval experiment.

The test collection approach would be classified as a theoretical strategy in McGrath's taxonomy. It allows evaluation to be done batch-style, and is therefore fast and cheap. It enables researchers to study the effect of varying some part of an information retrieval system's underlying engine, and is extremely valuable during development and fine-tuning. There are no established test collections for image retrieval, however. Leuski and Allan [67, 68] and Rorvig [100] have experimented with using TREC test collections to evaluate information visualisations; we consider their work in Chapter 3.

2.5.2 Human–computer interaction

Evaluation in human–computer interaction (HCI) involves assessing the interface of a system via experiments with users. Much of the literature on HCI evaluation (such as Nielsen's book [82]) focuses on how it can be used in the development of a real system; experts recommend an iterative design process, where future users of the system are asked to carry out typical tasks using prototype versions, to uncover problems and gain insight at an early stage. Research in HCI involves designing and evaluating new interaction techniques, and may use any of the four strategies mentioned by McGrath, although laboratory-based experiments are perhaps the most common.

In an experiment, the **efficiency** of the participants' interactions can be measured by considering the time taken to perform a set task. The **effectiveness** of their solutions to the task may also be assessed, in terms of quality or accuracy. Experiment participants are also usually asked to fill in question-

naires to indicate their **satisfaction** with various aspects of the system, or their preference for different versions. In this way, different interfaces (or different versions of the same interface) can be rigorously compared to each other. Frøkjær, Hertzum, and Hornbæk [43] found that effectiveness, efficiency, and satisfaction are not necessarily related to each other, meaning that all three should be considered in any evaluation. They also make the distinction between routine and complex tasks, stating that efficiency is a far better indicator of usability for the former than the latter, where a fast completion time may simply mean that the solution is of low quality.

For a deeper, more qualitative understanding of how users are interacting with a system, they may be asked to think aloud while working; their comments can be recorded, transcribed, and analysed. Thinking out loud may affect users' performance, perhaps distracting them or slowing them down, and so this method is not normally used in a formal experiment. It also generates a lot of data, which can take a long time to analyse, meaning that it may only be practical for a small number of participants.

Egan [37] has described how individual differences between the participants in an experiment can affect their performance: age, experience, and spatial and verbal ability are among the characteristics which have been found to be correlated with performance on certain tasks. Questionnaires and psychological tests can provide appropriate data to test whether some classes of people perform better than others for the task being studied. Borgman [11] considered individual differences in information retrieval.

2.5.3 Interactive information retrieval

The test collection approach was developed in the 1960s, when query execution was slow, and trained intermediaries (such as librarians) usually performed searches on behalf of clients. In current information retrieval systems, results are returned almost immediately, and it is much more common for end users to perform their own searches. Many of the assumptions of the test collection model are no longer valid, and a number of researchers have pointed out its limitations for evaluating interactive information retrieval systems [34].

Firstly, it assumes that a search consists only of a detailed query, issued as an expression of a fixed requirement. Bates [8] proposed an alternative model of searching, which she called *berrypicking*: she suggested that requirements evolve rather than remain static, and that users "gather information in bits and pieces instead of in one grand retrieved set", using a number of different search strategies. The user may be unwilling or unable to express her requirement in a single detailed query, and may prefer to browse the collection, or issue a series of simple queries. Subsequently, evidence for this model was provided by O'Day and Jeffries [83] who found that, based on what a user learned from the documents returned by an initial query, her requirement would often shift, resulting in a series of interconnected requests. Relevant information was progressively gathered from these, rather than entirely from the initial query. The test collection approach counts the number of relevant documents retrieved for a single query, not the amount of useful information gathered

from a whole search process.

Secondly, evaluation using a test collection relies on the relevance judgements, which are usually based solely on relatedness to the topic of the query. Su's study [115] found that precision was not correlated with the user's overall rating of the success of a query. Cooper [30] has argued that users do not just want relevant documents, they want documents with high *utility*, which he described as "a catch-all concept involving not only topic-relatedness but also quality, novelty, importance, credibility, and many other things." Subsequent user studies (summarised by Schamber [106]) found that in practice, users do apply many criteria other than topic-relatedness (such as those suggested by Cooper, as well as recency, clarity, cost, availability, and so on); these are not taken into account in a test collection's relevance judgements. In addition, relevance is subjective and dynamic, and can only truly be judged in context, but relevance judgements are objective and static.

Finally, only the engine of the system is considered in test collection evaluations, not its user interface. The engine might be capable of achieving high recall and precision, but weaknesses in the interface will greatly reduce the system's overall effectiveness. More recent work in information retrieval evaluation has therefore concentrated on developing methods that address these limitations, by involving user experiments. These methods are hybrids of information retrieval and HCI evaluation, and are more qualitative and more time-consuming than the test collection approach.

For example, the TREC series now has an interactive track [86] where experiment participants are asked to search for documents matching a given requirement, which corresponds to a query in the TREC test collection. The evaluation is still primarily based on measuring the recall and precision of the results saved by the participants, whose choice of query terms is likely to be greatly influenced by those used in the given specifications. As Borlund [12] has noted, the participants in these experiments are simulating intermediaries rather than end users, because they are searching for given requests, out of context, rather than realistic requirements.

In an attempt to bring more realism into information retrieval evaluation, Borlund and Ingwersen [13] have proposed the **simulated work task situation**, which would be called an experimental simulation in McGrath's taxonomy. Potential users of the system being evaluated are selected as participants, and they are given a scenario that is similar to one they might meet in real life (for example, imagining that they have to write a report for their manager on a given subject), including a described role and situation, and can then develop their own subjective and evolving requirements within this context. The participants save only the documents which help to satisfy their requirements, rather than saving as many topic-related documents as they can. To test this method, Borlund [12] carried out an experiment where the participants were given four simulated work tasks, and were also asked to search the collection for a real requirement of their own. She found that they tended to put as much effort into the simulated tasks as the real ones (in terms of the amount of time spent, for example, and the number of queries issued), and therefore concluded that they could be adequate substitutes. A good simulated work

task was one where the participant could relate to the described situation and found it interesting.

As far as we are aware, Jose, Furner, and Harper [60] are the only researchers to have applied simulated work tasks to the evaluation of an image retrieval system. They used graphic designers as experiment participants, and set them a task that involved selecting images to illustrate leaflets, as part of an imaginary freelance job. The experiment compared a spatial query interface to one where the user could only issue text queries. The designers rated the effectiveness of their selections, and their satisfaction with each of the interfaces; neither system effectiveness nor user efficiency were measured. The results showed that the participants preferred to issue spatial queries, as it allowed them to express their mental images. Neither of the two versions of the system had much support for image browsing, other than to scroll through query results; the example interface shows only three thumbnails at a time.

Different strategies are appropriate for evaluation of the different levels of an information retrieval system: for example, the underlying engine can be evaluated with a theoretical strategy (the test collection approach), and the system as a whole, as well as individual features of its user interface, can be evaluated with an experimental strategy. Dunlop [35], in a review of the research carried out in a three-year working group on evaluation of interactive information retrieval systems, concluded that "we need to consider more varied forms of evaluation to complement engine evaluation", including "importing and adapting evaluation techniques from HCI", "evaluating at different levels as appropriate" and "evaluating against different types of relevance". His recommendations are in accordance with those of McGrath, which we described at the very beginning of this section.

2.6 Studies of image searching

We have noted that the participants in evaluations of interactive information retrieval systems should be given realistic and representative tasks. In this section we discuss previous studies that aimed to establish how people search for images, and what types of requirements they tend to have. As we mentioned at the beginning of Section 2.3, our primary interest is in general photograph collections, and we therefore concentrate on studies covering the use of such collections (in the domains of publishing and advertising).

Other studies have considered more specialised collections [44, 66], and at present the VISOR project [3] is undertaking a wide-ranging investigation of image seeking behaviour in medicine, journalism, art history, museums, picture libraries, broadcasting, the police force, and architecture.

Fidel [41] has noted a distinction between seeking an image as an object, and seeking it as a source of information. In the applications considered in this section, the former type of requirement is far more common than the latter, and therefore an image's aesthetic attributes may be as important to the user as its content.

2.6.1 Markkula and Sormunen

Markkula and Sormunen [77, 78] observed and interviewed journalists (including sub-editors and layout designers) at a Finnish newspaper, to investigate their illustration requirements, and study how they searched for and selected appropriate photographs. Ornager [85] also considered the newspaper domain, but she concentrated on how the images were indexed and archived. Markkula and Sormunen's study is the only one carried out to date that involved the users of a *digital* photograph collection, who were able to search through it directly via a computer system. We shall therefore discuss their findings in some detail.

For a current news story, the illustrations had to be chosen quickly, usually from the group of agency photographs available for that event. With feature articles there was more time, and a wider scope for creative selection. The search process involved thinking of an initial idea, and then developing it while browsing through photographs, usually selecting a candidate set, and then making the final choice from among those.

The journalists had the option of searching through the collection themselves, or sending their request to the newspaper's archivists. The authors sampled 108 requests sent to the archive, and found that 22% were for "common nouns" (equivalent to the generic level) and 59% were for "proper names" (the specific level), with 8% for "themes" (the abstract level). A further 4% of requests were for "news events", 4% were for photographs of films or television programmes, and 3% were for a known photograph or series of photographs. About half of the requests specified further criteria, such as shooting distance.

Most of the journalists usually chose to do their own searching, rather than use the archivists, especially as they became more familiar with the system. The authors analysed the topics of 27 searches performed directly by the journalists, and found that requests for abstract content were relatively more common in that sample. Such requirements were difficult to express to the archivists, and doing so usually gave disappointing results, so the journalists wanted to browse through the collection themselves, to apply their own subjective judgement. This finding is in accordance with Shatford Layne's prediction [108], discussed in Section 2.4.

Directly entered queries tended to consist of a single word or phrase, and were often restricted by date or source. The journalists would usually have a more precise requirement in mind, but they believed that detailed queries often excluded the best photographs from the results, and felt that it was easier to issue a simple query and then browse the results as thumbnails. Browsing also supported the development of illustration ideas. The number of photographs they were willing to browse varied according to the person, the task and the time available: most of the journalists would reformulate a query if it had more than 100 results, but in some circumstances they would browse through far more, if they were motivated enough to find "the perfect photo for a particular article" (a case of browsing through 2000 photographs was noted).

The journalists verbally expressed a wide range of selection criteria while

searching, with the most important being relevance to the topic, which they established by checking the captions. Then, they would consider technical quality, recency, any previous usage of the photograph, its cost, the impression conveyed, and visual attributes and aesthetics (such as colour and composition). They would also reject posed or stereotypical photographs (such as politicians shaking hands) where possible. The importance of the different criteria varied according to the task. Once they had found a set of candidate photographs, the final selection would usually be based on visual attributes and aesthetics. It was also very important that the photograph should work well with those already chosen for the same page; the journalists wanted to have a variety of styles, orientations, and visual features.

The authors suggested that systems could provide more support for browsing by structuring the set of retrieved images, to make it easier for the user to see which types of photographs are present in the results, and apply secondary selection criteria. They noted that this structuring could be done using the annotations, or automatically extracted visual features; although the journalists did not seem to want to issue visual queries, their secondary selection criteria were often visual. In a later paper [112], the authors specifically investigated which features a journalist would use to classify a set of topic-related images. Of course, most of these were based on higher-level content, such as the number of people present, shooting distance, gestures, facial expressions, mood, and action.

2.6.2 Garber and Grunes

Stock photograph agencies own the rights to thousands of images, and these are used by clients who have decided against commissioning original photographs for a certain job. Garber and Grunes [47] modelled the picture selection process of art directors at advertising agencies, with the aim of establishing how best to support them when browsing a stock photograph collection. The art director creates an overall artistic concept, of which the image concept (the photograph or photographs used in the advertisement) is one component. She may wish to look through photographs to help her get some initial ideas for the artistic concept, or to find those that match a particular image concept. Both types of concept are liable to evolve in the course of working on the project. While browsing through the available images, the art directors often changed their stated search criteria.

2.6.3 Enser

Picture archives are distinct from stock photograph collections in that their content is mostly historical rather than contemporary, although they tend to have similar clients, such as publishers. Enser [38, 39] studied the Hulton Deutsch collection, which at the time was the largest such archive in Europe, with approximately ten million items (mostly black and white negatives and prints). A client usually contacted the archive by telephone; a picture researcher would help her to formulate her request, and then write it on a stan-

dard form. Enser analysed a sample of 1000 of these forms, which contained 2722 individual requests. He classified them into two categories: *unique* (equivalent to the specific level) and *non-unique* (which appears to encompass both the generic and abstract levels). He also noted that within these basic categories, the requirement could be subject to refinement (by time, location, action, event, or technical specification). Unique requirements were the most common overall; those without refinements made up 42% of the total sample, and those with refinements made up 27%. These were followed by non-unique requirements with refinements (25%) and those without (6%). In Enser's later work with Armitage [2] they studied seven image libraries with more specialised content, and classified a sample of 1749 written requests. The relative frequencies of unique and non-unique requests varied widely between libraries.

Unlike Markkula and Sormunen, Enser was able to analyse only the artefacts of the search process, not the process itself. It is likely that in both of his studies, the nature of the written requests was heavily influenced by the mediation of the picture researcher, and it is unclear how well they represented the clients' actual requirements. In online digital image collections, clients can look through images directly, without having to specify their requirements in detail.

2.7 Summary and research framework

In this chapter, we have introduced the fields of information retrieval, information visualisation, and image retrieval. We have noted that it is essential for image retrieval systems to provide support for browsing, and that although a number of researchers have developed systems which apply clustering or visualisation techniques to image collections, few have attempted to evaluate them. Possible approaches to evaluation can be borrowed from information retrieval and human–computer interaction; more recently, methods have been proposed that combine the two. The usage of multiple evaluation strategies has been advocated, to combine evidence gathered through different methods. In this research we specifically chose to evaluate visualisations created with MDS, to investigate whether they are useful for directed browsing of a set of images.

In the work described in Chapter 3, we used a theoretical strategy, based on the conventional test collection approach from information retrieval, to establish whether such visualisations should be useful in theory, and to compare the quality of versions created using different measures of visual similarity. Our indicator of quality was how closely the images with similar generic content were grouped (and separated from images with dissimilar generic content).

Then, in Chapters 4 and 5, we moved to an experimental strategy, conducting user experiments where we measured effectiveness, efficiency, and satisfaction. In these experiments, we assumed that the user had already carried out some restriction of the collection (either by querying or browsing), to reach approximately 100 images. As we have already noted, Markkula and

Sormunen [77, 78] estimated that the journalists they studied were generally willing to browse up to 100 images. Another consideration is the time taken to perform MDS, because the algorithms tend to be $O(n^2)$, requiring only about one second for 100 images, but 25 seconds for 500 images. The latter is unlikely to be acceptable to a user if the arrangement is created interactively.

In our first experiment, described in Chapter 4, the participants were asked to locate a given image within a set of thumbnails; in the second experiment they were asked to find a group of images matching a generic requirement. Both experiments had one condition where the set of thumbnails was arranged according to visual similarity, and one where it was arranged randomly. We expected that the participants would find the required images more quickly in the former condition.

The aim of the studies described in Chapter 5 was to evaluate similarity-based arrangements in a more realistic setting, using the simulated work task approach. We assumed that the image collection was annotated, allowing us to set a task that involved searching for images matching a specific-level requirement. We used designers as our participants, and asked them to select three images from a given set to accompany a passage of text. All of the images in the set were relevant to the requirement, and the designers therefore had to decide on their own selection criteria. We were interested in finding out whether organising such a topic-related set of images according to similarity (either visual, or based on the text of the annotations) would help the designers to narrow it down. One of the studies used the think-aloud method (a respondent strategy) to elicit more qualitative data.

The use of digital cameras for personal photography is becoming more widespread, and we consider this application area in Chapter 6. One of the main differences between a personal photograph collection and a commercial image library is that the user is also the photographer, and thus the images are very familiar to her. Markkula and Sormunen's study provided a foundation for the experiments in Chapter 5, but there was no equivalent study for personal photography. Therefore, before examining the potential usefulness of similarity-based arrangements in this area, we investigated the organisation and browsing of personal photograph collections, using respondent and field strategies, rather than controlled experiments.

In the first study in Chapter 6, we asked a group of keen photographers about how they organised their existing collections of prints and slides, and in the second study, we followed a group of volunteers through a six-month period of using a digital camera and a prototype photograph management system. The designers of this system did not include features for arranging thumbnails according to visual similarity, so we carried out that part of the investigation separately, using MDS arrangements created from sample sets of images provided by the participants.

Chapter 3

Comparing image similarity measures

As we mentioned in the previous chapter, multidimensional scaling can be used to arrange a set of images such that those which are most visually similar are placed next to each other. In the work described in this chapter, we had two main aims: to establish whether these arrangements should be useful in theory, and to compare the quality of arrangements created using different measures of visual similarity. An approach based on traditional information retrieval evaluation seemed ideal, because it allows a large number of possibilities to be tested quickly and conveniently. We therefore created two simple test collections of images from a library of stock photographs, and then used these to compare the relative performance of the visual similarity measures, both for retrieval (in multidimensional space) and for creating MDS arrangements (in two-dimensional space).

Other researchers (including Puzicha and his colleagues [93]) have compared visual similarity measures for the purpose of retrieval; we were concerned specifically with comparing them for constructing MDS arrangements, and our results for retrieval are given solely in order to show how the relative performance of the measures changes.

Rorvig [100] compared different layout algorithms and similarity measures for creating visualisations of text documents, but he made his comparisons purely by inspection, and did not attempt to quantify or test the significance of the differences.

Rogowitz and her colleagues [99] created MDS arrangements of a single set of 97 images, using two different visual similarity measures, and compared these (again by inspection only) to arrangements that were based on the similarity judgements of human assessors. They concluded that one of the similarity measures was closer to human perception than the other, but did not attempt a quantitative comparison. They implicitly assumed that there is a direct relationship between an image similarity measure's retrieval performance and its effectiveness for creating visualisations, an assumption which we explore in this chapter.

Name	*Features extracted*	*Method of measuring similarity*
a.1	Average HSV colour	Distance in HSV space
a.4 a.9 a.16	As above, but divide each image into 4, 9, or 16 equal sections	Calculate the distance in HSV space between each of the corresponding sections and then take the overall mean
h.l1	HSV histograms	L_1 distance
h.chi	HSV histograms	χ^2 statistic
h.jd	HSV histograms	Jeffrey divergence
emd	Colour signatures	Earth Mover's Distance
iris	Region summaries	IRIS measure

Table 3.1: The similarity measures that we compared, and the features that they require.

3.1 Measuring image similarity

The input to the MDS algorithm is a **similarity matrix**, which contains the pairwise similarities[1] of all of the images in the set. It is a triangular matrix with $\frac{n(n-1)}{2}$ entries, where n is the number of images in the set. In order to quantify the visual similarity of a pair of images, we first need to extract representative features from the images, and then define a suitable measure to compare them.

As we have already noted, primitive features such as colour and texture are easy to extract from images. The measures we chose to compare (with the features they require) are listed in Table 3.1, and described in more detail in the following sections. There are, of course, many possible versions of (and alternatives to) these; where possible, we chose simple measures from existing image retrieval literature, and implemented them ourselves. For the more complex measures, we obtained features and code directly from their inventors.

3.1.1 Average colour

This is an extremely simple method, using only a single colour to represent an entire image: the average hue, saturation, and value of all of its pixels. In the HSV colour space, hue is represented by the angle around a circle (for example, 0 degrees is red, 120 degrees is green, 240 degrees is blue), and saturation is represented by distance from the centre of this circle (the centre is least saturated, and the edges are most saturated). The third dimension (forming a cylinder) represents value, from black to white.

Measuring the similarity of two images is then just a question of measuring the Euclidean distance between their average colours in the HSV space, assuming that this is a reasonable approximation of their perceptual similarity. Because the space is cylindrical, the following formula must be used [111]:

$$d(i,j) = 1 - \frac{1}{\sqrt{5}}[(v_i - v_j)^2 + (s_i \cos h_i - s_j \cos h_j)^2 + (s_i \sin h_i - s_j \sin h_j)^2]^{\frac{1}{2}}$$

[1] In fact, *dissimilarities* are used: high values mean that the items are highly dissimilar. We treat the terms as interchangeable, as it is usually trivial to transform one to the other.

where i and j are colours in HSV space, and h_c, s_c, and v_c are, respectively, the hue, saturation, and value of a colour, c.

This is the a.1 measure in Table 3.1. It can be extended in order to take image composition into account, by dividing each image into a grid of equal-sized sections, and finding the average colour of each of those sections. Then, the distance between the average colours of corresponding sections can be calculated, as above. The final value of the measure is the mean of these distances. Measures a.4, a.9, and a.16 use 2×2, 3×3, and 4×4 grids, respectively.

3.1.2 Colour histograms

The main drawback of average colour is, of course, that it means losing all of the information about the large number of different colours which may be present in an image, and their relative distributions. If the colour space is quantised into bins, a colour histogram for an image can be produced by counting the number of its pixels that fall into each bin. The similarity between two histograms can then be measured, most commonly by comparing the contents of their corresponding bins.

There are many possible quantisation schemes; we chose the one outlined by Smith [111], which is based on the HSV colour space. Hue is assumed to be the most significant characteristic of a colour, and thus receives the finest level of quantisation, with the 360 degrees of the hue circle split into 18 sections of 20 degrees each. Saturation and value are assumed to be less important, and are each quantised to 3 levels. Pixels with no hue or saturation are treated as a special case, and split into 4 levels: black, dark grey, light grey, and white. Thus, there are 166 bins per histogram: 18 hues $\times$ 3 saturations $\times$ 3 values, plus 4 grey levels.

There are many different ways in which the resulting histograms can be compared in order to produce a single number representing the similarity of the corresponding images. We tried one measure from each of the first three categories described by Puzicha and his colleagues [93]: a heuristic histogram distance (L_1), a non-parametric test statistic (χ^2), and an information-theoretic divergence (the Jeffrey divergence).

As with average colour, it is possible to make these measures take some account of image composition, by dividing the images into sections, and aggregating the values of the measurement for each section. However, we did not pursue this, as it seemed to produce little or no improvement, especially considering the much higher number of features (and calculations) required.

L_1 distance

The L_1 distance (h.l1) is a simple and popular similarity measure. It involves adding up the differences between corresponding bins in the two histograms:

$$d_{L_1}(\boldsymbol{h}, \boldsymbol{k}) = \sum_i |h_i - k_i|$$

where $\boldsymbol{h}$ and $\boldsymbol{k}$ are histograms, and h_i represents the value in bin i of histogram $\boldsymbol{h}$.

χ^2 statistic

The χ^2 statistic (h.chi) measures how unlikely it is that one distribution was drawn from the population represented by the other:

$$d_{\chi^2}(\boldsymbol{h}, \boldsymbol{k}) = \sum_i \frac{(h_i - k_i)^2}{h_i + k_i}$$

Jeffrey divergence

This is an information-theoretic divergence (h.jd), measuring how compactly one distribution can be coded using the other one as the codebook [93]. It is calculated as follows:

$$d_J(\boldsymbol{h}, \boldsymbol{k}) = \sum_i \left(h_i \log \frac{h_i}{m_i} + k_i \log \frac{k_i}{m_i} \right)$$

where $m_i = \frac{h_i + k_i}{2}$.

3.1.3 Colour signatures and EMD

Histogram-based similarity measures are probably the most popular in current systems, but they are somewhat inflexible. Often, only directly corresponding bins are compared, disregarding neighbouring bins of similar colour. The measures are also sensitive to the chosen bin size: if it is too small, similar colours will be split across too many bins, but if it is too large, each bin will encompass too many colours, losing discrimination.

More adaptive methods have been suggested to get around these problems. Rubner, Tomasi, and Guibas [103] proposed using simple colour signatures, where each image is represented by a small number of colours (eight on average), with weights to indicate the importance of each colour in the image. They used a similarity measure called the Earth Mover's Distance (EMD), which reflects the minimal cost of transforming one signature to the other, and is a solution of a special case of the transportation problem from linear optimisation. They also illustrated the use of EMD to construct MDS arrangements of image sets, but did not compare it with other similarity measures for this purpose. EMD is another of the measures which were tested by Puzicha and his colleagues [93].

Rubner's C code for EMD was already in the public domain[2], and because we were using the same image library, he agreed to make his colour signatures available to us.

3.1.4 Image segmentation

All of the measures that we have discussed so far use global image features, although much current work in computer vision focuses on segmenting images into coherent regions. When constructing a visual query, this allows users to

[2]It is available from `http://robotics.stanford.edu/~rubner/emd/default.htm`

select individual regions from an image, rather than using all of it (such as in the Blobworld system [20]). It is also possible to incorporate region information into image features, which can then be compared with a pairwise image similarity measure.

Sinclair [110] has defined a method for unsupervised image segmentation; he provided us with an executable version that generates a region summary file for each image. An image is segmented into regions with broadly homogeneous colour properties, and regions have descriptors for colour, colour variance, area, shape, location and texture. Regions are then classified as either large (with an area of greater than 0.1% of the total image area) or small. Large regions are further classified as either textured or smooth, and small regions as regularly or irregularly shaped. Then, an image summary is constructed: the image is partitioned into nine sections of equal area, in a 3×3 grid, and a set of four global colour histograms (for large smooth regions, large textured regions, small regular regions, and small irregular regions) is created for each section; the histograms are normalised by image area. The largest (dominant) region in each of the nine sections is recorded.

He also provided us with C code to measure the similarity of two image summaries; the measure is known as IRIS (for Image Regions In Summary) and is designed to reflect both global image properties and the broad spatial layout of regions. The χ^2 statistic is used to determine the distance between colour histograms for each of the four region types in the two images being compared. The distance between dominant regions in the corresponding ninths of each image is given by the Mahalanobis distance in RGB colour space. The final value of the measure is a weighted sum of these distances.

3.2 Simple image test collections

As we discussed in Section 2.5.1, traditional information retrieval evaluations are based around test collections. This approach facilitates quick comparison of different methods of ranking results, and thus seems ideal here, as a convenient precursor to user experiments. However, as we have noted, there is no test collection for image retrieval. To create a test collection, we require a large set of images, with associated queries, and relevance judgements that indicate which images are relevant to each query. A common method of issuing a visual query in a content-based image retrieval system is query-by-example, where the user presents the system with an example of the type of image that would meet her requirement.

Real stock photograph collections are usually given some form of categorisation, providing a ready-made source of relevance judgements: if an image is used as a query, all of the other images from its category should be relevant to it. Obviously, it is also possible that images from other categories will be relevant, but we can overcome this by restricting the test collection to contain a set of categories with as little overlap as possible between their subject matter. Also, because we will be using measures based on visual similarity alone, it should be possible for anyone to identify by sight which category each image

belongs to, without any specialist knowledge. This can be achieved by choosing categories for their generic rather than their specific content; for example, anyone can identify that a photograph is of a mountain, but far fewer would be able to recognise a specific mountain (such as Mount Everest) without the aid of a caption.

We used the first two volumes of the *Corel Stock Photo Library* for all of the experiments described in this dissertation. Together, they contain 40,000 images, in 400 categories (listed in Appendix C), each of which has 100 images. From each library, we selected a subset of 20 categories that complied with the conditions above, to form two mini-collections, Corel 1 and Corel 2, whose contents are listed in Table 3.4. We used each of them separately, rather than combining them into a single collection, because there are only a limited number of generic categories in each library, and it was sometimes necessary to use the same type of category in both mini-collections (for example, both have a "flowers" category).

It is obvious that these collections are not at all realistic: for example, one would not generally expect 99 items (5% of the collection) to be relevant to every query. They should, however, provide a useful basis for comparison of similarity measures, as they will enable us to see how well each measure can separate the images in the relevant category from all of the others.

We introduced the cluster hypothesis in Section 2.1: a test collection satisfies the cluster hypothesis if the items that are relevant to a query tend to be more similar to each other, in general, than they are to non-relevant items. A way of testing this by observing the separation between two frequency distributions was devised by van Rijsbergen and Spärck Jones [119]. The accuracy of their method was questioned by Voorhees [122], who defined a test based on finding the most similar documents for each relevant document, and counting how many of them are also relevant. The analysis presented in this chapter is basically a more detailed version of these tests; in the particular case of query-by-example, every document (or image) in the test collection is also a query.

3.3 Comparing similarity measures for retrieval

Our basic procedure was to take a single image and use it as a query-by-example, ranking the remaining 1999 images in its test collection in order of their similarity to it. For each query, we produced a different ranking for each of the similarity measures under consideration; obviously, an ideal measure would rank the 99 relevant images (the rest of the query image's category) highest in the list. We repeated this process for all 2000 images in both the Corel 1 and Corel 2 test collections. Because these collections are fairly small, our first step was to precompute a similarity matrix for each measure and collection, from which the individual rankings could be produced without further calculation. We also created 100 2000-element similarity matrices filled with random numbers between 0 and 1, to give a baseline for comparison.

As we described in Section 2.5.1, the performance for a particular query at any given cut-off point in the ranking is conventionally quantified using

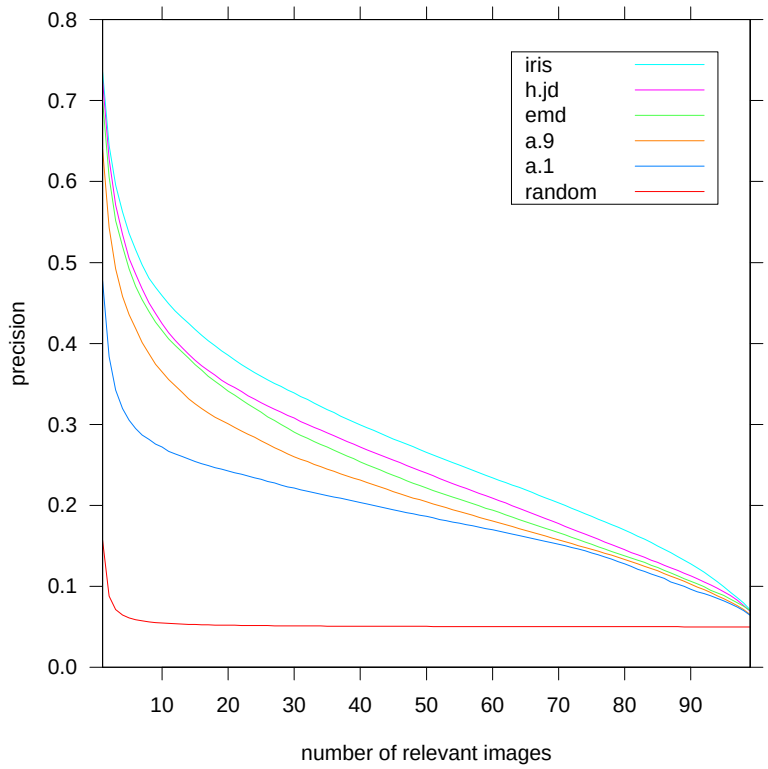


Figure 3.1: The recall–precision graph for the Corel 1 test collection. The line marked random is the mean of the 100 lines generated from the random similarity matrices, and is given as a baseline for comparison. The a.4, a.16, h.l1, and h.chi series have been omitted for clarity. The graph for Corel 2 is very similar, except that we do not have Corel 2 data for emd.

recall (the proportion of the relevant images in the collection that have been retrieved so far), and precision (the proportion of the images retrieved so far that are relevant to the query). Different retrieval methods can be compared by constructing a graph, plotting the value of precision at different levels of recall, averaged over every query in the collection. Here, each query has the same number of relevant images (99), and so the graph in Figure 3.1 simply plots precision at each relevant image in the ranking, instead of the usual practice of interpolating to 11 standard levels of recall.

It seems clear from this graph that even using a feature as simple as average colour (a.1) produces a large improvement over what would be expected by chance. Partitioning the images into a grid of equal areas results in a further improvement (only the series for a.9 is shown). The histogram-based measures perform even better: the Jeffrey divergence (h.jd) is shown on the graph, as it appears to be slightly superior to χ^2 (h.chi), which in turn is better than L_1 (h.l1). EMD (emd), for which Rubner was able to provide Corel 1 features only, performs approximately as well as the histogram-based measures, and the IRIS measure (iris) appears to have the best performance. Looking at the data another way: in order to find 10 relevant images for a Corel 1 query, it

would be necessary to retrieve 183 on average if the collection was ordered randomly, 37 if it was ranked using a.1, 27 using a.9, 24 using h.jd or emd, and 22 using iris.

However, we still have to test whether the observed differences between measures are statistically significant. To facilitate a more rigorous comparison, we require a single-number value which summarises a whole ranking, quantifying the performance of each measure for each individual query. We chose three different indicators, in order to examine different aspects of performance.

Firstly, we can simply calculate precision at a particular cut-off point in the ranking; because in this case there are 99 relevant images for each query, we chose **precision at 99**, so an ideal ranking would receive a score of 1.0.

However, simple precision does not take into account how high in the ranking the relevant images are. For example, using precision at 10, measure X might have relevant images at positions 1, 4, and 5, while measure Y has relevant images at 5, 8, and 10, but they would both get a score of 0.3 (3/10). We can therefore calculate precision at a number of different cut-off points, and take the mean of these values. In particular, if precision is calculated at each relevant item in the ranking, this is known as **average precision**, and is the equivalent of the area under a recall–precision curve like those in Figure 3.1. Using the same example, if there were only three relevant images in total, measure X would get an average precision score of $(1/1 + 2/4 + 3/5)/3 = 0.7$, and measure Y would score $(1/5 + 2/8 + 3/10)/3 = 0.25$.

Finally, we can use an indicator which takes only the highest part of the ranking into account, as this is arguably the most important. The most intuitive method would be to examine precision at a cut-off of 10 images, but this could have only 11 possible values, and would thus be unsuitable for use as a response variable in a parametric statistical test, and would be likely to produce many ties between similarity measures in a non-parametric test. Hull [55, 56] advocates using performance indicators which are averages, as these are likely to approximate continuous variables, making them appropriate for use in statistical tests. We therefore decided to use **average precision at 0.0 recall**, which is calculated in the same way as average precision, but using only the first tenth of the relevant items in the ranking (10 relevant images, in this case).

Then, to carry out a statistical test, it is necessary to construct a table with i rows and j columns, where each entry e_{ij} represents the performance of similarity measure j ($j = 1 \ldots m$, where m is the number of similarity measures being compared) for query i ($i = 1 \ldots 2000$). We created six of these tables, one for each combination of performance indicator (precision at 99, average precision, and average precision at 0.0 recall) and test collection (Corel 1 and Corel 2). They are summarised in Tables 3.2 and 3.3, which give the mean and standard deviation (across all 2000 queries in a test collection) of each performance indicator for each of the visual similarity measures being compared. To provide a baseline, we also calculated the values of the performance indicators for each of the 100 random similarity matrices. Precision at 99 ranged between 0.0480 and 0.0508, with a mean of 0.0495, which is the expected value of precision at any fixed cutoff in a random ordering of the images (99/1999). The

mean of average precision was 0.0529 (ranging from 0.0524 to 0.0535), and for average precision at 0.0 recall it was 0.0723 (0.0692–0.0752).

None of the performance indicators had a normal distribution, and their variances were unequal between conditions. This could not be remedied by transforming the data, and we therefore chose to use a non-parametric statistical test. Because we had more than two conditions to compare, we used the **Friedman test**, which is a non-parametric version of Analysis of Variance (ANOVA). Specifically, we used the version of the test that was proposed by Conover [29], and recommended for information retrieval evaluation by Hull [55]. With a non-parametric test, only the relative performance of the similarity measures for each query is taken into account, not the actual values of the performance indicators. The first step in carrying out a Friedman test is to replace each value in the table with a rank: *within each query*, the m similarity measures are ranked according to their relative performance, so that the worst measure receives a rank of 1, and the best receives a rank of m. Any tied measures are assigned a rank which is the mean of the ranks they would have received had they not been tied. For each of the performance indicators, Tables 3.2 and 3.3 show the mean rank of each similarity measure, across all 2000 queries in a test collection.

For each of the six combinations of performance indicator and test collection, the Friedman test was significant at a level of $p < 0.05$, which indicated that there were overall differences between the similarity measures. We then used a multiple comparisons test (Fisher's LSD, again as proposed by Conover [29] and recommended by Hull [55]) to establish which pairwise differences between similarity measures were significant at the $p < 0.05$ level. The results are represented by the horizontal lines in Tables 3.2 and 3.3: when there is *no* significant difference between similarity measures, they are connected with a line. Almost all of the pairwise comparisons are significant; only the a.9 and a.16 measures could not be separated in any case, and the emd and h.chi measures were not significantly different for the Corel 1 test collection (we do not have Corel 2 data for emd).

For a subset of the similarity measures, Table 3.4 expands the average precision figures in Tables 3.2 and 3.3 to show the variation between categories in more detail. It is interesting to see, for example, that the iris measure does not have the best performance for every category. It is also clear that some types of query are much easier to satisfy with a visual similarity measure than others: as one might expect, the easiest queries are those for categories where the images are most visually homogeneous (such as *Arabian Horses*, where all of the pictures appear to be of the same horses in the same field), and dissimilar to the images in other categories (such as *Divers & Diving*, where all of the images contain an uncommon shade of blue). In text retrieval evaluation, it is usual to find large differences between queries (for example, in TREC [50]).

We should perhaps note that the features for iris were created from 768×512 images (to allow the segmentation to work properly), while all of the average colour and histogram features were calculated from 96×64 thumbnails, which may be what gives iris some of its advantage. To test this, we repeated our calculations for the next best measure, h.jd, using the larger image size, and

		a.1	a.4	a.9	a.16	h.l1	h.chi	emd	h.jd	iris
Prec. at 99	*mean*	0.208	0.225	0.235	0.236	0.252	0.264	0.262	0.269	0.291
	s.d.	0.157	0.155	0.164	0.164	0.174	0.185	0.174	0.189	0.193
	rank	3.47	4.20	4.64	4.74	4.97	5.42	5.57	5.79	6.21

		a.1	a.4	a.9	a.16	h.l1	emd	h.chi	h.jd	iris
Avg. prec.	*mean*	0.189	0.210	0.223	0.225	0.239	0.246	0.253	0.258	0.283
	s.d.	0.150	0.155	0.168	0.169	0.188	0.186	0.202	0.206	0.213
	rank	3.16	4.04	4.75	4.79	5.00	5.45	5.52	6.00	6.27

		a.1	a.4	a.9	a.16	h.l1	emd	h.chi	h.jd	iris
Prec. at 0R	*mean*	0.324	0.413	0.452	0.455	0.484	0.508	0.513	0.523	0.550
	s.d.	0.250	0.296	0.314	0.318	0.312	0.312	0.319	0.321	0.324
	rank	2.79	3.97	4.72	4.77	5.14	5.61	5.72	6.09	6.19

Table 3.2: A summary of each measure's retrieval performance with the Corel 1 test collection. For each performance indicator, the similarity measures are sorted according to their mean rank, and the horizontal lines connect the measures which are not significantly different from each other at the $p < 0.05$ level.

		a.1	a.4	h.l1	a.9	a.16	h.chi	h.jd	iris
Prec. at 99	*mean*	0.190	0.227	0.245	0.236	0.239	0.256	0.259	0.295
	s.d.	0.123	0.142	0.149	0.146	0.148	0.151	0.151	0.155
	rank	2.88	3.98	4.34	4.37	4.46	4.87	5.12	5.99

		a.1	a.4	h.l1	a.9	a.16	h.chi	h.jd	iris
Avg. prec.	*mean*	0.172	0.210	0.229	0.222	0.225	0.239	0.244	0.282
	s.d.	0.109	0.134	0.152	0.142	0.145	0.157	0.158	0.164
	rank	2.70	3.91	4.26	4.40	4.48	4.87	5.33	6.05

		a.1	a.4	a.9	a.16	h.l1	h.chi	h.jd	iris
Prec. at 0R	*mean*	0.315	0.428	0.465	0.469	0.493	0.522	0.532	0.581
	s.d.	0.243	0.296	0.308	0.310	0.305	0.312	0.313	0.316
	rank	2.60	3.81	4.37	4.43	4.51	5.03	5.44	5.81

Table 3.3: As Table 3.2, but for the Corel 2 test collection.

ID	*Title*	a.1	a.9	h.jd	iris
1000	Sunrises and Sunsets	0.132	0.125	0.163	0.203
3000	World War II Planes	0.251	0.486	0.419	0.588
8000	Birds	0.071	0.080	0.094	0.084
13100	Flowers Volume II	0.093	0.086	0.163	0.142
22000	Bridges	0.115	0.134	0.148	0.145
29000	Exotic Cars	0.157	0.140	0.176	0.145
31000	Residential Interiors	0.172	0.226	0.281	0.397
52000	Butterflies	0.186	0.201	0.179	0.294
73000	Firework Photography	0.301	0.302	0.580	0.580
91000	Fruits & Vegetables	0.120	0.148	0.119	0.178
100000	Bears	0.114	0.128	0.120	0.112
104000	North American Deer	0.104	0.133	0.145	0.118
107000	Elephants	0.185	0.300	0.189	0.228
108000	Tigers	0.160	0.188	0.298	0.310
110000	Wolves	0.106	0.189	0.155	0.178
113000	Arabian Horses	0.466	0.595	0.559	0.561
156000	Divers & Diving	0.489	0.353	0.672	0.640
172000	Action Sailing	0.235	0.261	0.268	0.360
181000	Models	0.181	0.198	0.246	0.187
184000	Ice & Icebergs	0.144	0.193	0.183	0.221
	overall average precision	0.189	0.223	0.258	0.283
208000	Fungi	0.271	0.334	0.328	0.341
209000	Fish	0.124	0.137	0.197	0.198
221000	Flowers Close-up	0.241	0.241	0.204	0.283
225000	Freestyle Skiing	0.187	0.262	0.266	0.323
240000	Arthropods	0.121	0.123	0.139	0.217
268000	African Birds	0.090	0.105	0.093	0.108
273000	Performance Cars	0.158	0.234	0.211	0.279
300000	Surfing	0.216	0.228	0.404	0.368
314000	Dolphins and Whales	0.176	0.374	0.252	0.385
317000	Whitetail Deer	0.148	0.199	0.241	0.252
320000	Victorian Houses	0.142	0.244	0.376	0.390
326000	Wildcats	0.088	0.105	0.141	0.135
329000	Hot Air Balloons	0.073	0.098	0.151	0.194
332000	Fabulous Fruit	0.187	0.167	0.192	0.169
338000	Sailing	0.193	0.329	0.347	0.359
345000	Sunsets Around The World	0.204	0.177	0.255	0.311
351000	Trains	0.128	0.245	0.201	0.305
359000	Aviation Photography 2	0.193	0.270	0.151	0.322
364000	Kitchens and Bathrooms	0.348	0.411	0.533	0.512
388000	Women In Vogue	0.144	0.160	0.187	0.191
	overall average precision	0.172	0.222	0.244	0.282

Table 3.4: The categories used in the two test collections, with their average precision when their contents are used as queries, for each of four similarity measures. The top half of the table is Corel 1, and the bottom half is Corel 2. The shading of the cells shows the relative ranking of the measures for a particular category: the darker the cell, the worse the ranking.

	similarity measure								
	a.1	a.4	a.9	a.16	h.l1	emd	h.chi	h.jd	iris
relative time	1	3.8	8.6	15	8.8	230	10	24	16

Table 3.5: The approximate time needed to create a similarity matrix, given relative to a.1, for each of the nine measures. It is interesting to compare these timings with the results given in the previous section; for example, the a.9 measure is much faster than a.16, for about the same retrieval performance.

found that this did improve its performance (for example, the average precision for Corel 2 rose from 0.244 to 0.252), but it was still significantly worse than iris.

3.4 Calculating a similarity matrix

If there is sufficient storage available, pairwise similarities for the entire collection can be precomputed, so that whenever a similarity matrix for a given subset of the collection is required, the entries do not have to be calculated dynamically. However, as this is not always possible, Table 3.5 shows approximately how quickly each of the measures can be calculated, relative to the fastest, a.1. On a 296MHz UltraSparc-II processor, it takes 0.02 seconds of CPU to compute a 100-image similarity matrix using a.1, but approximately 4.6 seconds on average using emd. It seems clear that emd is unlikely to be practical where dynamic calculation of similarities is necessary. The table does not include the time taken to extract the image features, as this needs to be done only once; it takes longer for complex features (such as regions), which also tend to require a larger amount of storage.

3.5 Comparing similarity measures for visualisation

Having found significant differences between the similarity measures when used in the context of retrieval, we expected that these differences would remain in the context of visualisation, meaning that the best measures for retrieval would produce the best MDS arrangements. We therefore needed an objective method of evaluating the quality of these arrangements, in terms of how closely images from the same category are placed.

Our performance indicators can be adapted to apply to MDS arrangements, simply by producing rankings based on the two-dimensional Euclidean distance between images in the arrangement, rather than their (multidimensional) similarity according to the original measure. The two-dimensional equivalent of precision can be defined as the proportion of relevant images within a given radius of the current query image (using the image centres when measuring distances). The radius used is the distance from the query image to whichever relevant image represents the chosen level of recall, as illustrated in Figure 3.2.

Leuski and Allan [67] used the two-dimensional equivalent of average precision (which they called *average spatial precision*) to quantitatively compare

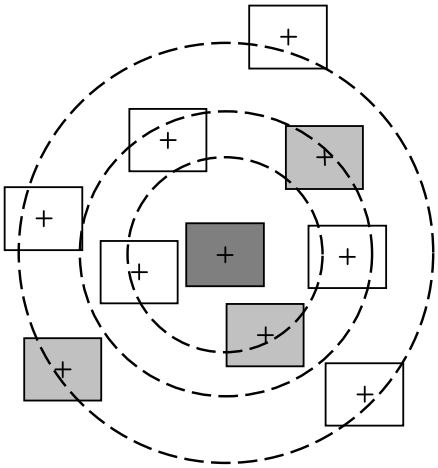


Figure 3.2: An illustration of how precision can be measured in two dimensions. The dark grey rectangle in the centre represents the query image, and the light grey rectangles represent images that are relevant to the query. The white rectangles are non-relevant images. Moving outwards from the centre, the dashed circles represent increasing levels of recall: at the closest relevant image, we see that 1 non-relevant image also falls within the circle, making the precision 0.5 at $r = 1$. At $r = 2$ there are 3 non-relevant images within the circle, giving a precision of $2/5 = 0.4$. Finally, at $r = 3$, precision is $3/8 = 0.375$. The average precision of the arrangement for this query image is therefore $(0.5 + 0.4 + 0.375)/3 = 0.425$.

different visualisations of sets of text documents. Each set consisted of the top 50 results of a query on the TREC test collection, and they were interested in exploring the potential of a visualisation as an alternative to the conventional ranked list, because others have found that the cluster hypothesis often applies to sets of query results [52]. TREC does not use query-by-example, so the query itself was not an object in their visualisations, and could not be used as the starting point for their calculations (unlike our example in Figure 3.2). Instead, they used the relevant document that was ranked highest in the list of results, assuming that the user had already found it and would start looking for other relevant documents from there.

Because a ranked list is one-dimensional, it is reasonable to assume that the user will traverse it sequentially from the top, and thus precision is a good measure of the perceived effectivess of the system, as it indicates how many results the user has to look at in order to find a certain number of relevant documents. This is not the case with two-dimensional precision, however, because it is unlikely that the user would move outwards in a circle from the starting point, as depicted in Figure 3.2. Note, therefore, that our use of two-dimensional precision is *not* an attempt to model the way in which a user would navigate through the arrangement; it is merely intended to provide a way of quantifying how well clustered the relevant images are.

In their later work [68], Leuski and Allan suggested that, if relevant documents tend to be grouped together in a visualisation, the optimal search strategy would be for the user to continually shift her attention to the centroid of the cluster of documents judged relevant so far, and then select the closest unexamined document to that point. In their experiment, discussed in Section 2.2, they found that participants came close to following the optimal strategy in practice. When the objects in the visualisation represent text documents, the user must open a document and read (at least some of) it in order to judge its relevance, and she cannot read more than one at a time. It is therefore in her interest to adopt a systematic strategy, so that she can make the best possible decision about which document to look at next, minimising the overall amount of effort. In contrast, the relevance of images can often be judged immediately from their thumbnails, and the user can shift her attention around the display very quickly, taking in a number of images at once. Thus, Leuski and Allan's model could be applied to arrangements of images, but it is probably less likely to be an accurate representation of the user's actions.

At this stage, we chose to cut down the number of similarity measures under consideration, retaining a.1, which is the simplest; a.9, which has equivalent performance to a.16 but requires a smaller number of features and calculations; h.jd, which is the best of the histogram-based measures; and both emd and iris, which are the only examples of their type.

Then, we applied MDS[3] to the 2000-image similarity matrices we created previously, producing two-dimensional configurations for each combination of similarity measure and test collection. We then repeated the process described in Section 3.3 to give tables for the 2D equivalents of precision at 99,

[3] using an algorithm based on simulated annealing; see Appendix A.

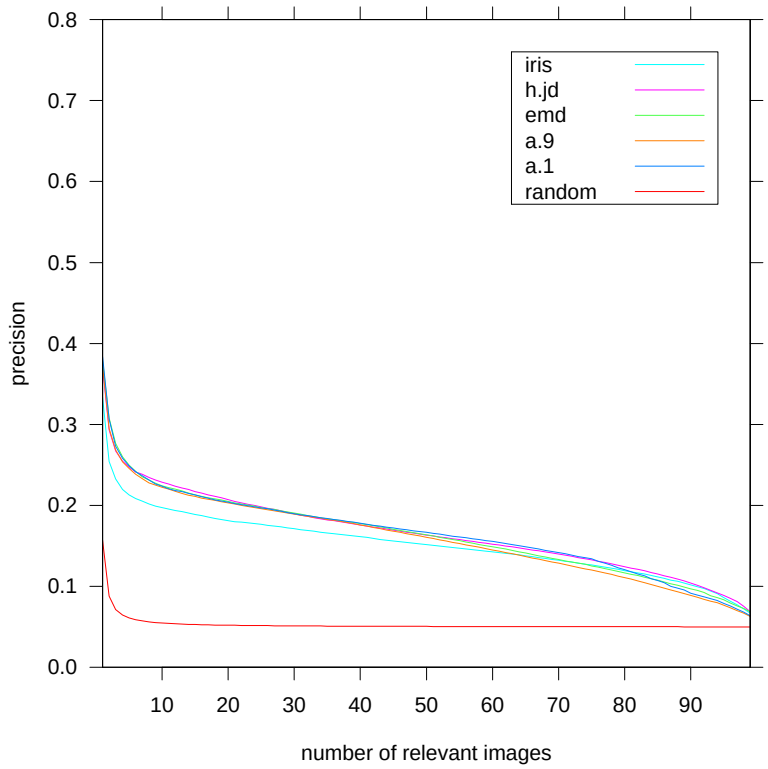


Figure 3.3: Two-dimensional precision against recall for the Corel 1 test collection, showing that the differences between the measures have almost disappeared. The line for random is the same as that in Figure 3.3. Again, the graph for Corel 2 is very similar, minus emd.

average precision, and average precision at 0.0 recall; summaries of these are shown in Tables 3.6 and 3.7, with a recall–precision graph for Corel 1 in Figure 3.3. We again used the Friedman test to look for significant differences (at the $p < 0.05$ level) in the performance of the similarity measures. This time, the iris measure generally performs worst (along with emd in two out of the three Corel 1 tables), and there are no other consistent differences between measures. At the radius of the 10th relevant image in the Corel 1 collection, on average there would be 183 images within the circle by chance; 51 in an arrangement created using iris; 45 using a.1, a.9, or emd; and 44 using h.jd.

Each MDS run produces a slightly different 2D configuration, as mentioned in Appendix A. We repeated our analysis for two other MDS runs (for each combination of similarity measure and test collection), and found that the results varied very little, with (for example) differences in mean average precision of no larger than 0.001.

Although there are large differences between the similarity measures in terms of conventional information retrieval, when used in conjunction with MDS the differences are very small. To look for an explanation of this, we have to consider MDS in more detail.

Appendix A describes the MDS algorithm we used, and how the energy

		iris	emd	a.1	a.9	h.jd
Prec. at 99	*mean*	0.165	0.179	0.182	0.178	0.182
	s.d.	0.117	0.138	0.146	0.127	0.130
	rank	2.85	2.93	3.04	3.08	3.10

		emd	iris	a.1	a.9	h.jd
Avg. prec.	*mean*	0.164	0.152	0.166	0.160	0.167
	s.d.	0.132	0.102	0.134	0.116	0.116
	rank	2.86	2.87	3.03	3.07	3.17

		iris	h.jd	a.9	emd	a.1
Prec. at 0R	*mean*	0.226	0.261	0.258	0.263	0.263
	s.d.	0.183	0.215	0.216	0.223	0.221
	rank	2.79	3.03	3.06	3.06	3.07

Table 3.6: A summary of each measure's performance in MDS arrangements, with the Corel 1 test collection. For each performance indicator, the similarity measures are sorted according to their mean rank, and the horizontal lines connect the measures which are not significantly different from each other at the $p < 0.05$ level.

		iris	h.jd	a.1	a.9
Prec. at 99	*mean*	0.156	0.173	0.173	0.175
	s.d.	0.104	0.110	0.115	0.115
	rank	2.27	2.54	2.56	2.63

		iris	h.jd	a.1	a.9
Avg. prec.	*mean*	0.142	0.153	0.158	0.158
	s.d.	0.083	0.086	0.095	0.094
	rank	2.23	2.53	2.59	2.66

		iris	h.jd	a.9	a.1
Prec. at 0R	*mean*	0.218	0.246	0.252	0.262
	s.d.	0.168	0.186	0.193	0.214
	rank	2.28	2.54	2.57	2.62

Table 3.7: As Table 3.6, but for Corel 2.

collection		*similarity measure*				
		a.1	a.9	emd	h.jd	iris
Corel 1	Avg. prec. ratio	0.876	0.718	0.669	0.647	0.537
	MDS energy	0.023	0.054	0.055	0.096	0.107
Corel 2	Avg. prec. ratio	0.919	0.711	–	0.627	0.504
	MDS energy	0.016	0.047	–	0.098	0.105

Table 3.8: This table shows how much of the retrieval performance of a measure is retained in its MDS arrangement. Average precision is used as the performance indicator, and the given ratio is the mean of average precision for MDS divided by the mean of average precision for retrieval. These values can be compared to the energy of each of the corresponding MDS arrangements; higher energy tends to result in a greater loss of performance.

value is calculated; this gives an indication of how closely the inter-point distances in the final two-dimensional arrangement approximate the corresponding dissimilarities (lower values are better). Table 3.8 shows the energy of the arrangements produced using the different similarity measures for the two test collections.

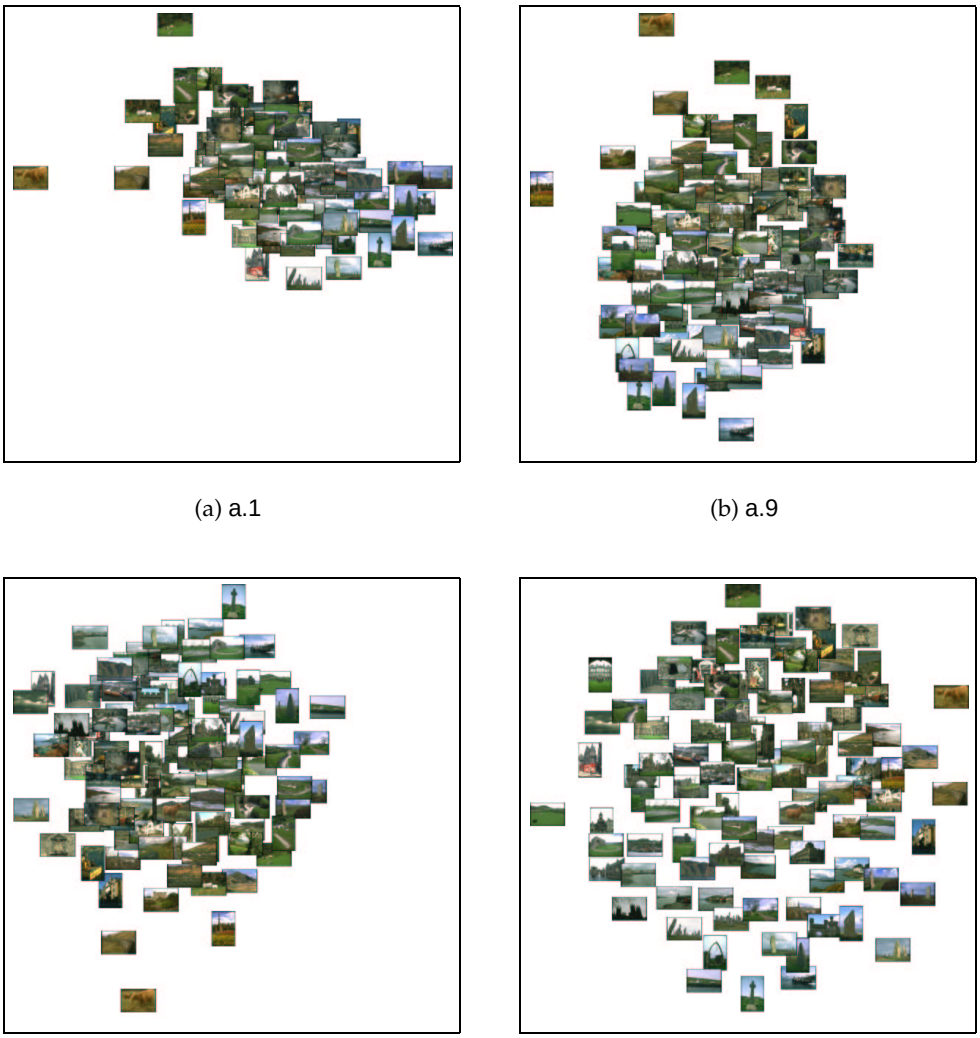
The simplest measure, a.1, is three-dimensional, and so a reduction to two dimensions does not result in much error, giving its configurations a low level of energy. However, more complex similarity measures tend to have a higher dimensionality, thus producing configurations with more energy (and error). It appears that the more complex measures tend to lose whatever advantage they have in high-dimensional space when forced into two dimensions by MDS. Thus, the best similarity measures for use with MDS will be those which make a good trade-off between effectiveness and dimensionality.

3.6 Discussion

Figure D.1 in Appendix D shows MDS arrangements of the 2000 images in Corel 1, for a.1, a.9, emd, h.jd and iris. The first three are tightly clustered, resulting in a lot of overlap between thumbnails, but in the arrangements created using h.jd and iris the thumbnails are spread more widely. Ironically, using a higher-dimensional similarity measure actually seems to be beneficial for presentational purposes: the images will be measured as less similar to each other in general, and therefore MDS will force them further apart, reducing overlap. Other than this, it is difficult to judge qualitatively which arrangement is best, reinforcing our quantitative results.

As we mentioned in Section 3.2, the collections we used were very artificial in nature, and it is therefore difficult to know whether these results would generalise to real image collections. In particular, the retrieval results show that in many cases the individual categories look different enough from each other to be separated even with a similarity measure based on features as simple as average colour.

The four arrangements in Figure 3.4 were created using what is probably a more realistic image set: all 100 images from a single category in the Corel



(a) a.1 (b) a.9

(c) h.jd (d) iris

Figure 3.4: A qualitative comparison of MDS arrangements based on four different similarity measures, using a set of 100 images of Scotland (category 92000).

collection. We can only compare these qualitatively, because the image set cannot be easily partitioned into related subsets. Again, there appears to be little difference between the arrangements, other than their shape and orientation, and repeating this process for other single-category arrangements produces the same outcome. It therefore seems that although the arrangements based on iris scored significantly worse than the others for most of the performance indicators, the difference is not large enough to be noticeable in practice.

As we have already discussed, the queries in our test collections were for a generic type of image, because it would be highly unreasonable to expect to satisfy a more specific requirement using a similarity measure based only on primitives (with no annotations available). We therefore tested how well clustered the images of the same generic type were, in MDS arrangements created using different visual similarity measures. Of course, there are many other possible aspects of the usefulness of such MDS arrangements, and we explore these further in the following chapters.

3.7 Conclusions

We can identify differences in performance among visual similarity measures when they are used to simulate the retrieval of results for a query-by-example from simple image test collections. The differences disappear, however, when the measures are used in conjunction with MDS to create two-dimensional arrangements of the same images. Measures that are more effective for retrieval tend to be more complex, and lose their discriminating power when their configurations are forced into two dimensions. Qualitatively, there is also little to choose between the arrangements produced by the different similarity measures, although the points tend to be more spread out in those created using higher-dimensional measures, thus allowing more images to be simultaneously visible.

Both for retrieval and for creating MDS arrangements, even the simplest measure easily outperformed a purely random ordering of the images, meaning that although similarity was measured only at the primitive level, this had a much higher correspondence with similarity at the generic level than would be expected by chance.

With no strong evidence to support any single similarity measure over the others, we chose to use iris for our subsequent user experiments, because of its effectiveness for retrieval, and the fact that it results in less image overlap when used with MDS.

Chapter 4

Initial user experiments

In the previous chapter, we described the results of a theoretical analysis that used methods borrowed from traditional information retrieval evaluation, and these showed that images with similar generic content are often grouped together in MDS arrangements based only on visual similarity.

For the work described in this chapter, we adopted an experimental strategy, as is commonly used in HCI evaluations. We carried out two user experiments, both of which involved directed browsing tasks. In the first experiment, the participants had to locate a given target photograph within a set of 80 thumbnails, and in the second experiment they were asked to find a group of photographs matching a given textual description, from a set of 100 thumbnails. Both experiments had one condition where the set of thumbnails was arranged according to visual similarity (using the IRIS measure, which we introduced in the previous chapter), and one where it was arranged randomly.

4.1 The first experiment

The aim of the first experiment was to establish whether users could find a given photograph more quickly in a set of images arranged according to their visual similarity than in a set arranged randomly. We chose this task because it involves a low-level visual search: if users are able to make sense of an arrangement based on visual similarity, they should be guided to the part of the display that contains the target image, and should therefore be able to find it more quickly than they would in a random arrangement.

We carried out a series of trials where we displayed a full-size target image to the participant, removed it from the screen, and then measured how long she took to find it within a set of thumbnails. In the visual condition, this set was arranged according to visual similarity. In the random condition, the set was arranged in a random order, but in a grid, rather than allowing free placement of the images anywhere on the screen. We chose to do this because current applications usually display thumbnails in a grid.

The participants had to remember what each target image looked like while they searched for it. Current theories of human memory [1, 4] suggest that there is a specialised type of working memory for storing visual informa-

tion, known as the **visuo-spatial sketch pad (VSSP)**. While searching, the participants would have maintained their mental image of the target on the VSSP, which is also where a familiar image recalled from long-term memory is placed [57, p213]. Superficially, then, it may seem that this experiment simulated the real-life task of searching for a familiar image. However, it is likely that the mental representation of a familiar image in long-term memory is different to that of an image recently seen for the first time. Although it is not clear in what form an image is actually stored, it is thought that initially (once it is perceived) it has a purely sensory representation. As this is processed and interpreted, a semantic representation is also created; this tends to persist for longer than the sensory version, which gradually becomes more vague [1, p216]. When a familiar image is recalled and placed on the VSSP, the mental image is not an exact reproduction of the original, but a reconstructed version based on what is remembered about the semantic interpretation (for example, which objects were present, and their approximate configuration) plus whatever remains of the sensory representation.

In general, human memory is subject to *levelling* (reducing the original to a simplified form) and *sharpening* (highlighting or embellishing unusual details). To use an example from visual memory [114, p194], an ellipse is usually remembered as being more circular than it was (levelling), and a gap in the outline of an otherwise typical shape is remembered as being larger than it was (sharpening). Colours tend to be levelled, so that they are remembered as more primary and more saturated than they actually were, and this may be common to both working and longer-term memory [114, p198].

4.1.1 Pilot studies

We carried out two small pilot studies prior to the experiment proper, each with two participants, and the first of these resulted in five major changes to the original design.

The target was originally displayed at thumbnail size, in an attempt to remove the variability that would be introduced in the process of mentally shrinking a full-size image to a thumbnail. However, this irritated the participants, who often moved their faces very close to the screen, and complained that they could not see what was in the target image. We therefore chose to display the target at full size.

It also became obvious that, when looking at the set of thumbnails, the participants could immediately filter them according to orientation (portrait or landscape), meaning that they only had to consider those which were of the same orientation as the target. To remove this source of variability, we decided to use only landscape-oriented images.

We originally allowed 30 seconds of search time, but if participants had not found the target within 20 seconds, they were unlikely to find it at all. We therefore set the limit at 20 to speed up the procedure.

Finding the target was surprisingly difficult: in 18.2% of the trials in the first pilot study, the participant either did not select an image, or selected the wrong one, resulting in a high level of missing data. In that study, the image

(a) (b)

Figure 4.1: Two examples of possible target images, from the set in Figure 4.2.

sets were of two different types: the non-target images were either drawn from the same category (in the Corel collection) as the target, or selected randomly from the entire library. The participants had much more difficulty with the first type, because in many cases they had not remembered enough about the target to tell it apart from any similar-looking images in its category. To cut down on missing data, we therefore decided to use only randomly selected groups of images. We also reduced the number of thumbnails in each set from 100 to 80.

All of these changes were implemented in the second pilot study, which had the same design as the main experiment, and the missing data rate fell to 7.3%.

4.1.2 Participants and apparatus

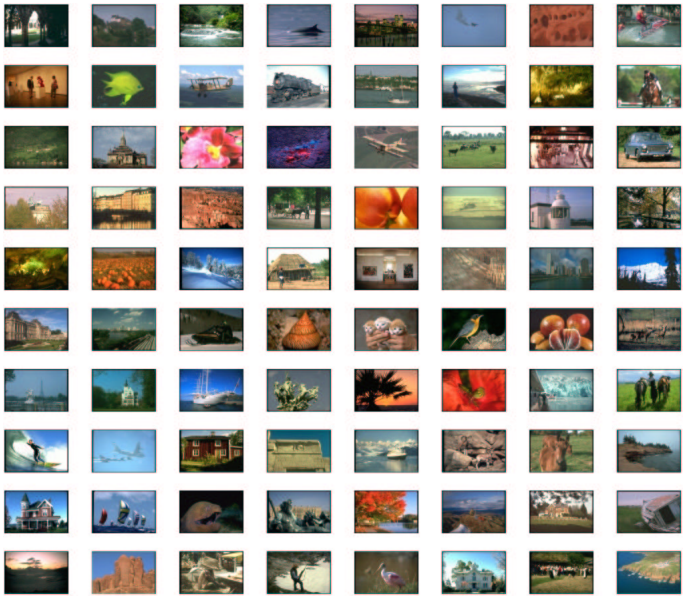
Sixteen participants were recruited from among the students and staff of the University of Cambridge. All had either normal or corrected-to-normal vision, with no colour blindness (self-reported). They were paid five pounds for their participation.

48 sets of 80 images were randomly selected from the *Corel Stock Photo Library 2*, which is described in Appendix C. Target images (such as Figure 4.1) were presented to the participants at 768×512 pixels, and thumbnails were displayed at 96×64 pixels. For each of the 48 sets of images, we created an MDS arrangement[1] based on visual similarity (using the IRIS measure, described in Section 3.1.4) and a random[2] grid arrangement (examples of both are shown in Figure 4.2), and saved them as large, maximum quality JPEG files. Participants were tested one at a time, all on a Windows NT 4 PC with 64MB of memory, and a 17-inch monitor set at 1280×1024 resolution. Each target image therefore covered 30% of the screen area, and each thumbnail covered 0.47%. A Java program was used to display the images and record timings.

[1] using the Newton-Raphson algorithm; see Appendix A.

[2] All of the random arrangements mentioned in this dissertation were created using Java's built-in pseudo-random number generation facility.

(a) MDS arrangement, based on visual similarity.



(b) Random grid.

Figure 4.2: Two arrangements of the same set of images, as used in the first experiment.

Participant	*Part one*					*Part two*				
1	PR	R	R	R	R	PV	V	V	V	V
2	PV	V	V	V	V	PR	R	R	R	R
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table 4.1: The experiment design. Each entry in the table represents a block of 12 trials, using either the **R**andom or **V**isual condition. The first block of each condition was a **P**ractice block.

We fixed the thumbnail size and number of images per set after the pilot studies, because there is a trade-off involved: the more images that are displayed simultaneously, the smaller they have to be to avoid obscuring each other. A grid has a regular structure which enables the most efficient use of screen space, but in an MDS arrangement a thumbnail can potentially be placed anywhere on the screen, making overlap almost inevitable (see Figure 4.2(a)).

It would have been possible to avoid overlap by implementing a mechanism to bring an image to the front when the user moved the mouse over it. However, rather than complicate the experiment, we simply decided to use only target images that had more than 70% of their area visible in the visual arrangement, so that the participants were never asked to find an image that was badly obscured (this would have produced outliers and missing data). As a result, approximately 10% of the images in each visual arrangement could not be used as targets.

4.1.3 Design

The experiment was within-subjects, so all of the participants used both arrangement types, and the design was balanced: half of the participants used the visual arrangement first, and half used the random arrangement first. Each participant saw the same arrangements, but these were presented in a random order within each condition. The 48 trials in each condition were presented in 4 consecutive blocks of 12, with breaks between blocks. The participants also received one practice block at the start of each condition. The design is illustrated in Table 4.1.

The target image for each set was selected at random for each participant (in the case of the visual arrangement, from those at least 70% visible, as described in the previous section), although a target was never used more than once in the whole experiment.

4.1.4 Procedure

Participants were asked to read a set of instructions (Figure D.2 in Appendix D), and were then given a practice block of whichever condition they were receiving first, followed by four further blocks of that condition. They then received the practice block and four further blocks of the other condition, as we have already described. In a single trial, participants were shown a target image for

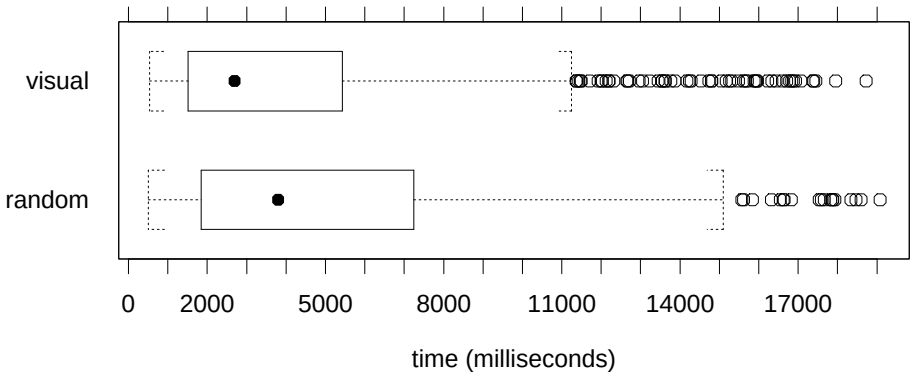


Figure 4.3: Distribution of times for the two arrangement types.

10 seconds. It was then removed from the screen, and they were asked to find it, as quickly as possible, in a set of 80 thumbnail images, and click on it using the mouse. The time limit was 20 seconds.

To give some consistency between trials, the participants had to move the mouse pointer to the centre of the screen (to click on a button) before the set of images appeared, and were asked not to move it again until they had found the target image with their eyes.

As soon as an image was selected, or when the time ran out, the correct image in the set was highlighted. This gave the participants some feedback about their performance, and was intended to help them remain interested in the task.

We told the participants that they were being timed, and that this would allow us to compare the two arrangement types to each other. We did not tell them how the arrangements were created, in case this biased them against the random arrangement. In particular, the instructions did not mention that visually similar images were clustered together in the visual arrangements, but we assumed that the participants would notice this during the practice block.

To provide qualitative data, the participants filled in a post-experiment questionnaire (Figure D.3 in Appendix D). This asked them to describe what sort of searching style they had used for each of the arrangement types, to explain what they thought were the advantages and disadvantages of each arrangement type, and to say which one they preferred. We also asked them if some kinds of target image were easier or more difficult to find than others. The whole procedure took about an hour per participant.

4.1.5 Results and discussion

The response (dependent) variable was time, the time taken between the set of images appearing and the participant clicking on the target image. The boxplot[3] in Figure 4.3 compares the distribution of this variable for the two

[3]In all of the boxplots in this dissertation, the solid black dot represents the median, and the box shows the limits of the middle half of the data (its length is the interquartile distance, IQD). The end of each whisker marks the nearest value that falls inside $1.5 \times IQD$ from the edge of

	Df	F	p
subject	15	1.93	< 0.050
block	7	1.00	0.429
trial	11	0.36	0.972
type	1	17.98	< 0.001
Residuals	1312		

Table 4.2: ANOVA for log(time). The multiple r^2 value is 4.20%.

arrangement types. Because of the positive skew, a log transform was applied prior to analysis to make the distribution more normal. We chose the statistical package S-Plus for data analysis, and used its linear regression features to construct a linear model for log(time) using the following predictor (independent) variables:

- subject, the ID of the participant (16 levels)
- block, the sequence number of the block, from 1 to 8 (an ordered factor)
- trial, the sequence number of the trial within the block, from 1 to 12 (an ordered factor)
- type, the way in which the image set was arranged (visual or random)

Then, the analysis of variance (ANOVA) function was used to extract information from the fitted model, as shown in Table 4.2[4]. Participants were significantly faster with the visual arrangement than with the random arrangement ($p < 0.001$), with a mean time of 4311 milliseconds for the former (s.d. 4011), versus 5107 milliseconds for the latter (s.d. 4148).

In the post-experiment questionnaires, most of the participants described making use of the grouping of visually similar images in the visual arrangement (quoted comments are preceded by a participant ID number):

01 *Noted the majority colour in the photo. Photos were generally grouped due to major colour, so searched around the right area.*

15 *I looked for other images with the same overall colour/brightness and then looked for the specific image.*

However, 7 of the 16 participants did not mention making use of the fact that visually similar images were clustered together in the visual arrangement, which suggests that they had not noticed it. If we split the participants into two groups (according to whether they noticed) and repeat the data analysis for each group, we see that type is significant in both: $F(1, 737) = 8.49$ ($p < 0.01$) for the 9 participants who noticed, and $F(1, 556) = 7.55$ ($p < 0.01$) for the 7 who

the box, and the circles indicate any values outside this range.

[4]All of the ANOVAs quoted in this dissertation were calculated using Type III (marginal) sums of squares. These correct for the effects of all of the other terms in the model, and thus do not depend upon the order in which terms were added.

did not. Also, the difference between mean times for the two arrangement types is consistent across the groups. It seems that the participants who were not consciously aware of the grouping by visual similarity had still used it to guide them in the direction of the target.

In 9.3% of the trials, the participants failed to select an image within the time limit, and in 3.0% they selected the wrong image. Therefore, a total of 12.3% of the trials had to be regarded as missing data; this was 14.5% of the random trials and 10.2% of the visual trials. In a few of the cases where no image was selected, it is likely that the participant would have eventually found the target, given a longer time limit, but in others participants simply forgot what the target image was. Also, some participants said that when they saw the target image in the set, it sometimes looked different to the way they had remembered it. In particular, reducing an image to the size of a thumbnail may result in the appearance of remembered objects changing dramatically.

08 *Often a certain feature noted in the large picture wouldn't be easy to discern in the thumbnails, throwing off the search method.*

13 *[It was difficult to find] pictures with recognisable small details which aren't so visible when reduced.*

Those participants who were consciously making use of the grouping of visually similar images in the visual arrangements said that sometimes the target thumbnail did not appear in the area of the screen that they expected. Consequently, they would scan the area repeatedly before realising that the target was just outside it. This may have been because the similarity measure and layout algorithm failed to arrange the images in the way that the participant expected, or because the participant's memory of the target was different to its actual appearance.

15 *[The visual arrangement had] an ordering that didn't always meet with my expectations.*

The effect of salience and position

Within a set of images, some always seem to stand out more than others, and we expected that the salience of a target would play an important part in how quickly it was located; research in visual perception and attention [58, 123] suggests that humans have fundamental mechanisms for detecting salience. The eye is structured such that the area with greatest acuity (the fovea) covers only 1–2 degrees of visual angle, and acuity then drops sharply towards the periphery. Movement of the eyes is necessary, to allow the area with highest priority to be seen at the greatest resolution. It is thought that a specialised process monitors the entire visual field, analysing it in parallel (in a bottom-up manner) to determine where overt attention should be directed. The area that is deemed to be the most salient is attended to next; in many cases it appears to **pop out** [123, p163].

In studies of visual attention, a task like the one used in this experiment would be called a **visual search** [127]: the participant must locate a given target within a set of **distractor** objects. Typically, simple objects such as letters and shapes are used in these studies, not photographs. In order to search, humans must consciously direct their attention via a top-down process, for example in order to look for a face in a crowd. There can be conflict between the top-down and bottom-up processes; for example, regardless of the current task, the attention is usually drawn towards movement in the peripheral vision. In this experiment, the participants were performing a top-down search for the target image, but certain images would always have popped out from the set, no matter what the target was. More salient targets should have been located more quickly.

The questionnaire asked if some target images had been easier or more difficult to find than others, and all of the participants agreed. In particular, they said that close-up images were easier to find, as well as those containing large areas of bright/saturated colour, and strong contrasts. The most difficult targets had the opposite attributes, that is, they showed distant objects or lots of small details, contained dull colours, or lacked contrast; landscapes and buildings were mentioned as examples.

Many of the participants noted that salience had affected their search styles, reporting that they had first quickly scanned the image set to see if the target "jumped out", and only if that failed did they search in a more methodical way:

05 *Seeing if it jumped out at me and then searching either outwards from the centre [for both types] or downwards from the top in a "reading" pattern [for random].*

08 *[For random] first a general scan, to see if anything stood out, then a methodical column-by-column scan, if I hadn't found it. [For visual] again, a general scan, then more focussed on the area I thought the image might be in (because of its colour).*

16 *[With random] for each picture I tried to memorise the most striking colour, and when presented with the grid, I would first try to match the colour, then the rest of the image. Did not scan in a particular order at first; but when I was not able to find a matching picture, I would scan the grid row by row. [With visual I] essentially used the same method as above.*

The multiple r^2 statistic gives the percentage of the variation in the original data that is explained by the linear model. Its value for the model used in Table 4.2 is only 4.20%. The target images were selected at random from each set, and their salience is likely to contribute to the unexplained variation. We therefore felt that it would be interesting to retrospectively introduce some measure of salience into the statistical analysis.

The participants' comments would suggest that salience could be measured using features such as overall brightness or colour saturation. However, studies of attention [58] and visual search [127] have found that the salience of an item is mainly determined by its context: a salient target is one which is

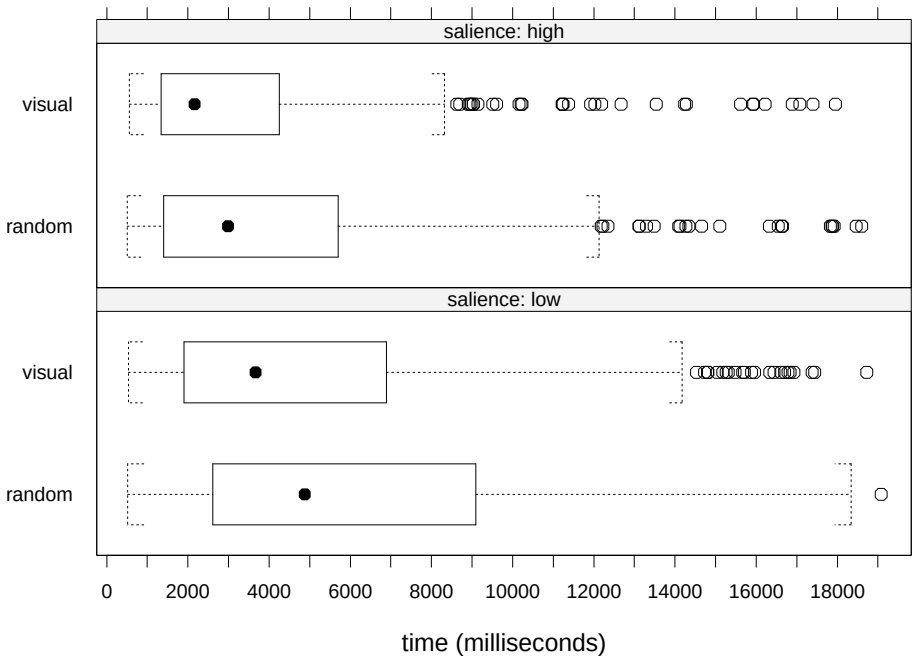


Figure 4.4: The effect of salience on response time, for both arrangement types. To create this figure, we simply split the data according to whether the salience of the target was higher or lower than the median. Overall, the mean time for the more salient half was 3931 milliseconds (s.d. 3718, 8.07% missing), and for the less salient half it was 5545 (s.d. 4323, 16.54% missing).

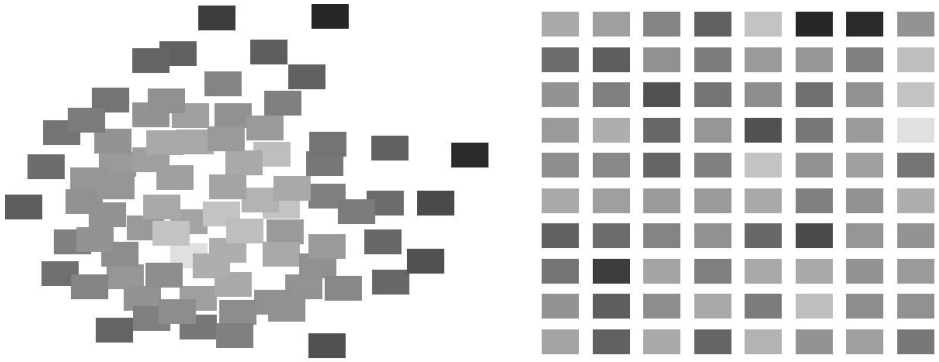
visually dissimilar to the distractors. Salience should therefore be quantified relative to the set of objects in which a target is presented, rather than based on absolute features. For example, in the randomly selected sets of images used in this experiment, a close-up picture of a flower was often highly distinctive, but it would be far less likely to stand out from a set that consisted entirely of flower photographs.

Thus, we estimated the salience of a target as the mean of its visual dissimilarity (according to the IRIS measure) to all of the other images in its set. For the targets used in the experiment, the measurement seems to be in broad agreement with the participants' opinions. In Figure 4.1, example (a) has the highest salience value of all of the possible targets (at least 70% visible) in Figure 4.2, while example (b) has the lowest. Figure 4.4 illustrates the effect of salience on response time, showing that more salient images seem to be found more quickly, regardless of arrangement type.

We also expected that the screen position of a target image in an arrangement would have some effect on the time taken to find it: participants always moved the mouse to the centre of the screen at the start of a trial, and so images placed at a greater distance from the centre should have taken longer to reach. Like salience, this was originally left as a source of random variation, dealt with by the fact that targets were selected at random. We therefore added another new predictor variable to the model: the distance in pixels of the target's

	Df	F	p
subject	15	2.02	< 0.050
block	7	0.76	0.622
trial	11	0.34	0.978
salience	1	127.97	< 0.001
distance	1	22.54	< 0.001
type	1	13.72	< 0.001
Residuals	1310		

Table 4.3: ANOVA for log(time), adding predictor variables for the salience of the target, and its distance from the centre of the screen. This causes the multiple r^2 value to rise to 12.76%.



(a) MDS arrangement.

(b) Random grid.

Figure 4.5: The salience (according to our measure) of the images in Figure 4.2, shown on a perceptually linear grey scale from most salient (dark grey) to least salient (light grey).

centre from the centre of the screen.

Table 4.3 shows the ANOVA for the linear model with the two new variables added. Despite the crudeness of the salience measurement, it is highly significant: more salient images were found more quickly. Images appearing closer to the centre of the screen were also found significantly more quickly. The significance of arrangement type is not affected by these additions, and the multiple r^2 value is 12.76%, which indicates that the new model explains more of the variation in the data.

However, matters are slightly more complicated than this. Figure 4.5 contains a representation of the salience of each image in the two arrangements from Figure 4.2. It is clear that in the visual arrangement, the most salient images appear around the outside, with the least salient in the middle; given our definition of salience, and the way in which MDS arrangements are created, this is what we would expect. The variables salience and distance are correlated for the visual arrangement ($r = 0.675$, $p < 0.001$), making it diffi-

		Visual only		*Random only*	
	Df	F	p	F	p
subject	15	0.76	0.725	2.27	< 0.010
block	3	0.16	0.922	2.17	0.091
trial	11	0.34	0.975	0.47	0.924
salience	1	40.54	< 0.001	60.85	< 0.001
distance	1	3.89	< 0.050	18.26	< 0.001
multiple r^2		9.91%		16.46%	

Table 4.4: ANOVAs for log(time), separated according to arrangement type, with visual on the left and random on the right. The residual degrees of freedom are 658 and 625 respectively.

cult to isolate the effects of salience and distance from each other in this half of the data. In the random arrangement, of course, there is no correlation ($r = 0.008, p = 0.814$).

We therefore divided the data in two, according to the arrangement type used, and repeated the data analysis for each half (Table 4.4). We can see that salience and distance are significant, both for the random arrangement (where they are independent of each other), and the visual arrangement (where they are correlated), although the F and p values for distance are much lower in the latter case. Thus, distance from the centre of the screen does have an effect on response time, but the size of the coefficients for distance in the two models suggests that it is very small, compared to salience. Figure 4.6 shows that for the random arrangement, images in the 20 grid positions closest to the centre were found slightly faster than those in the other 60 positions.

13 *[With random, I looked] near the middle and if that failed then systematic scan.*

When using the visual arrangement, participants' attention was guided to the area where the thumbnails were visually similar to the target, even if they were not consciously aware of this. Visual search studies suggest, however, that it is more difficult to find a target if it is surrounded by similar-looking distractors [127], because these lower the **local contrast** and make it less salient. So although it is easy to find the right area of the visual arrangement, the actual target is unlikely to pop out from within that area, which then has to be scanned serially. Because there is no such grouping in the random arrangement, individual images should have a greater local contrast with their neighbours, making the target more likely to pop out; if it does not, however, the whole arrangement has to be scanned sequentially, instead of just an area of it. This phenomenon perhaps explains why the difference in mean response time between arrangement types was fairly small.

12 *[With visual,] colour guided me to the right area, then detailed random scanning of images in that area. [...] Advantage was (1) some sorting had already been performed (2) could search on broad shapes/blocks of colour rather than specific objects in the picture. Disadvantages of [random]: no pre-sort, every image needed to be examined.*

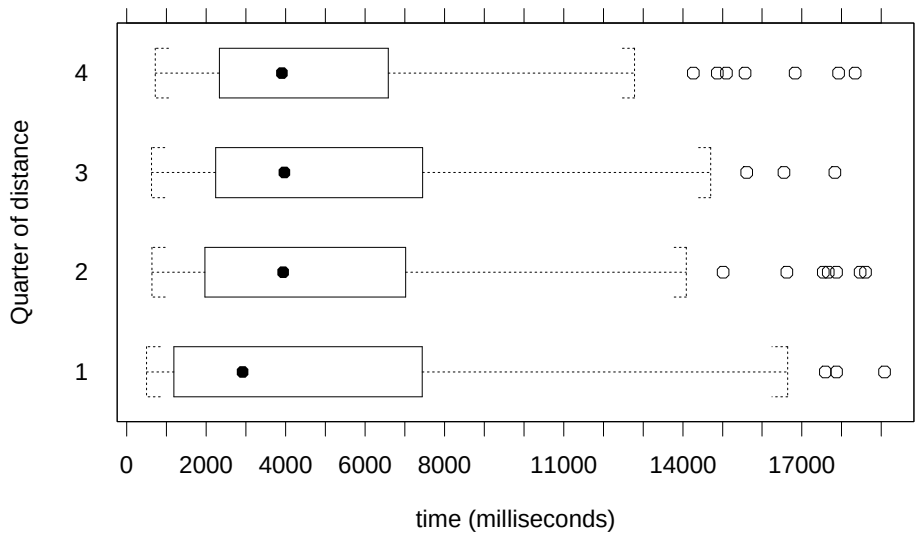


Figure 4.6: The effect of distance from the centre on response time, for the random arrangement only (where salience and distance are independent). The 80 possible grid positions are divided into four groups of 20, according to distance from the centre, and each box represents one group, so that, for example, the 20 closest to the centre are represented by box 1.

According to our current measure, the salience of a target is determined only by the set of images in which it is presented, and not their arrangement; the measure does not adjust for the fact that a target's actual salience depends greatly on its neighbouring images. A more accurate measure would be weighted such that images closer to the target contribute more to the value than those further away, or perhaps would include only the closest n images, or those within a certain radius. However, it would be difficult to determine the correct weighting (or cut-off point), because the size of the human visual system's useful field of view varies between tasks and displays [123, p157]. Itti and Koch [58] describe a way of modelling bottom-up attention within a single image, and Pirolli, Card, and Van Der Wege [89] propose an integration of visual attention theory and their own information foraging theory; it is possible that either of these approaches could be extended to modelling the user's perception of an arrangement of thumbnails.

The problem of image overlap

Objects in information visualisations are normally represented by points, but in the arrangements considered here, images are represented by thumbnails. Because these are bigger than a single pixel, there is a risk of overlap between them, resulting in some of the thumbnails being partially or wholly obscured. Before we conducted the experiment, we did not expect that overlap would be a problem, because overlapping images are likely to be very visually similar, and we had ensured that all targets would be at least 70% visible.

However, in the post-experiment questionnaire, ten of the participants said

	Df	F	p
subject	15	0.76	0.722
block	3	0.19	0.906
trial	11	0.38	0.965
salience	1	32.22	< 0.001
distance	1	3.57	0.059
visibility	1	4.35	< 0.050
Residuals	657		

Table 4.5: ANOVA for log(time), with the data from the visual arrangements only, incorporating visibility. The multiple r^2 value is 10.50%.

that the overlap in the visual arrangements had given them problems, making it difficult to see the edges of images and sometimes resulting in a remembered detail being obscured. Five of the participants expressed a preference for the random grid, citing lack of overlap and ease of systematic scanning (along rows or down columns) as their reasons. Three of these five had noticed the grouping according to similarity in the visual arrangement. Seven preferred the visual arrangement (five of whom had noticed the grouping by similarity; the other two said that the visual arrangement was more "fun" and the random grid was "boring"), and four had no preference (one of whom had noticed the grouping).

05 *[I preferred] the [random] method — advantage of being predictable and no overlapping images.*

16 *With the [visual] method, it was at times more difficult to match a picture when it was overlapped by another picture.*

We therefore defined a new variable, visibility, which is the percentage of the target that was visible in the arrangement (ranging from 70% to 100%), and added this to the model for the visual half of the trials; Table 4.5 shows the ANOVA results. The new variable is significant at a level of $p < 0.05$: less visible images take longer to find. As with distance, however, visibility is correlated with salience ($r = 0.299$, $p < 0.001$): overlapped images tend to be less salient. Figure 4.7 attempts to separate the effect of these two variables on response time. Similarly, visibility is correlated with distance ($r = 0.180$, $p < 0.001$): images placed closer to the centre of the arrangement are more likely to be overlapped.

4.1.6 Proximity grid

Given the participants' dislike of the image overlap in the MDS arrangement, and the fact that overlapped images seemed to take longer to find, we realised that it should be possible to combine the advantages of the two arrangement types by adapting the MDS arrangement to lay out the images in a more regular way. As a result, Wojciech Basalaj developed the **proximity grid** family of algorithms [7], which are described in Appendix A. Briefly, they guarantee

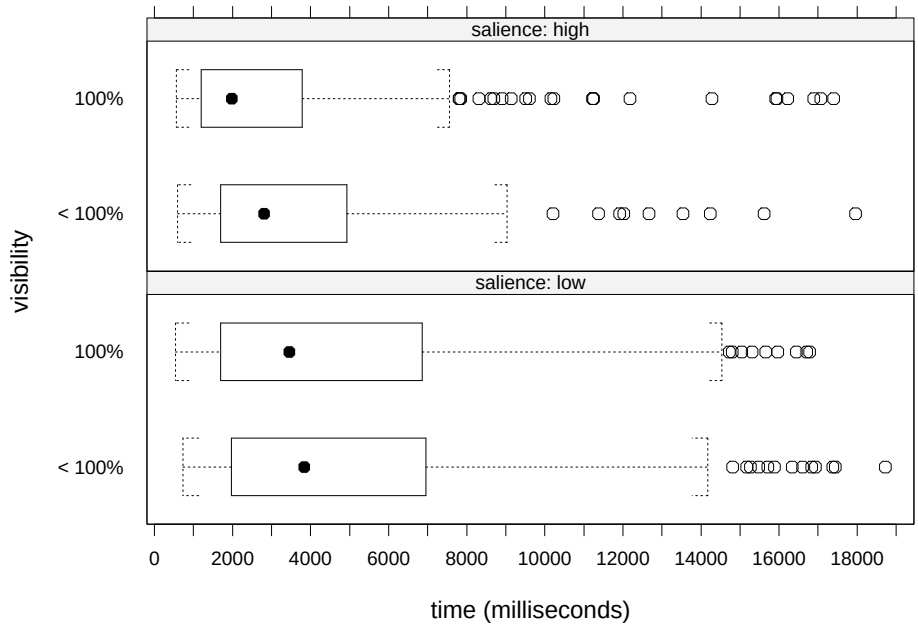


Figure 4.7: The effect of visibility on response time (for the visual arrangement only), again splitting the data according to whether the salience of the target was higher or lower than the overall median. The number of trials that should be represented by each box is (from the top) 254, 120, 190, and 204, but the missing data rates were 5.1%, 10.0%, 10.0%, and 16.7% respectively.

(a) Continuous. (b) 12 × 12 grid. (c) 10 × 10 grid.

Figure 4.8: Three arrangements of 100 images of Kenya, based on visual similarity. The continuous MDS arrangement is the most faithful to the original configuration, but has overlapping images. The 12 × 12 proximity grid removes the overlap while preserving much of the structure, and the 10 × 10 proximity grid is the least accurate but maximises the thumbnail size.

a minimum amount of separation between objects by forcing them to lie on a grid of a specified size, while still aiming to produce an accurate representation of their similarities. The resulting arrangements resemble those produced by systems that use self-organising maps, which are discussed in Chapter 2.

As an example of the desired effect, Figure 4.8 shows how the 100-image arrangement of Figure 1.1(b) (on page 15) would look as both a 12×12 and a 10×10 proximity grid. The former leaves 44 cells empty, and results in much of the original structure of the arrangement being preserved, while the latter is less accurate but has no empty cells, allowing the thumbnails to be larger. At present, only square proximity grids can be generated.

Although the MDS arrangements used in this experiment are continuous in theory, they are in fact discrete when displayed on the pixels of a computer screen: they can be thought of as very sparse grid arrangements, where each cell is one pixel in size. Thumbnail images are larger than one pixel, and therefore overflow their grid cells, causing overlap.

4.2 The second experiment

Having discovered in Chapter 3 that arranging images based on their visual similarity tends to cluster together those of the same generic type, we wanted to determine whether people would be able to take advantage of this in practice. The experiment carried out by Leuski and Allan [68] found that this was true for a visualisation of text documents, as we discussed in Section 2.2. In our second experiment, instead of having to locate a particular image within a set, the participants were given a textual description, and were asked to find images to match it. Again, the image set could be arranged randomly or according to visual similarity.

Research in visual memory suggests that when people are asked to imagine an object, they form a mental image of a stereotypical version of that object [57, p213] and its colour [114, p198]. Our participants may therefore be able to use an arrangement based on visual similarity to guide them in the direction of relevant images, in an analogous way to the first experiment. Of course, in this case the mental image will be much more vague than a memory of a recently viewed photograph.

4.2.1 Participants

We carried out a pilot study, with one participant, which resulted in some changes to the experiment instructions and questionnaire. Then, 20 participants were recruited from among the students and staff of a variety of departments within the University of Cambridge. All had either normal or corrected-to-normal vision, with no colour blindness (self-reported). They were paid five pounds for their participation.

ID	*Category name*	*Description given to the participants*
208000	Fungi	fungi (e.g. toadstools)
209000	Fish	fish
221000	Flowers Close-up	close-up photos of flowers
225000	Freestyle Skiing	skiers or snowboarders
240000	Arthropods	small invertebrates (e.g. insects, spiders, moths, wasps)
268000	African Birds	birds
273000	Performance Cars	racing cars
300000	Surfing	surfing
314000	Dolphins and Whales	dolphins or whales
317000	Whitetail Deer	deer
320000	Victorian Houses	houses, from the outside
326000	Wildcats	big cats (e.g. wildcats)
329000	Hot Air Balloons	hot air balloons
332000	Fabulous Fruit	fruit
338000	Sailing	sailing
345000	Sunsets Around The World	sunsets
351000	Trains	trains
359000	Aviation Photography 2	planes or helicopters
364000	Kitchens and Bathrooms	the interiors of houses (e.g. kitchens, bathrooms)
388000	Women In Vogue	female fashion models

Table 4.6: The textual descriptions that were given to participants in the second experiment, as well as the original category name and ID number from the Corel 2 mini-collection.

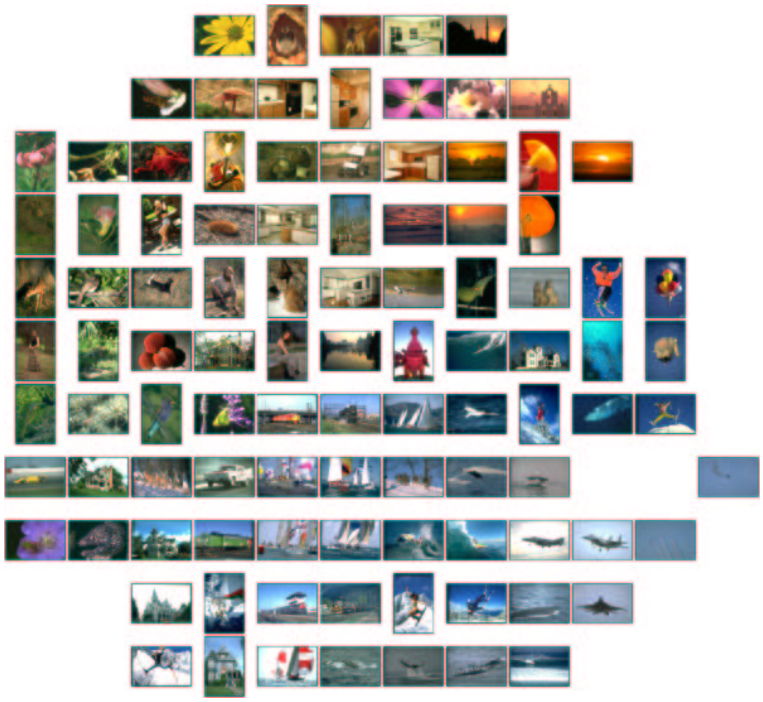
4.2.2 Apparatus

The Corel 2 miniature test collection, as described in the previous chapter, was used for this experiment. Its 2000 images were split into 20 sets of 100, so that each set contained between 3 and 7 images from each of the 20 categories, with no duplication of images within or between sets. In each trial, the participant was given a textual description of the images she should search for; this was usually an edited version of the original category name, as shown in Table 4.6. We chose to vary the number of images taken from each category, because in real search tasks the user does not know in advance how many relevant items will be present. Also, if a participant found four relevant images quite quickly, but knew that there were exactly five present, she might then spend a long time hunting for the last one, instead of simply assuming there were no more and moving on, thus skewing her search time.

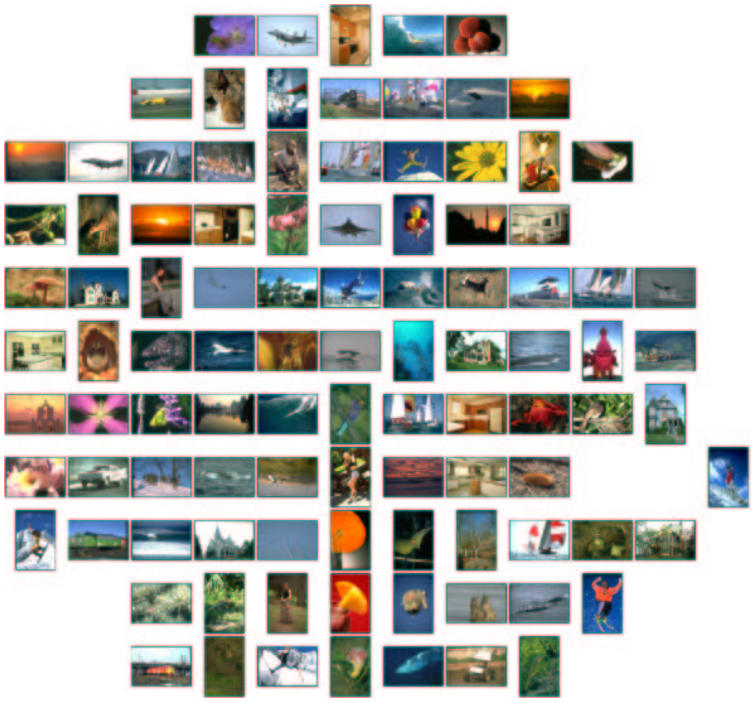
In the visual condition, the set of 100 thumbnails was arranged according to visual similarity (using the IRIS measure), and then placed in a 12×12 proximity grid[5]. For the random condition the same images were rearranged into a random order, within the same grid positions. Examples of each type are given in Figure 4.9.

We used the same PC as in the first experiment, this time with 192MB of memory, and its monitor set at 1600×1200 resolution. The set of thumbnail images was shown in an overview display of 1000×1000 pixels, with a magnification facility to allow the user to see a small area of the display in more detail (as shown in Figure 4.10).

[5] using the Newton-Raphson method, followed by a greedy algorithm; see Appendix A.



(a) MDS proximity grid, based on visual similarity.



(b) Random grid.

Figure 4.9: Two arrangements of the same image set, as used in the second experiment.

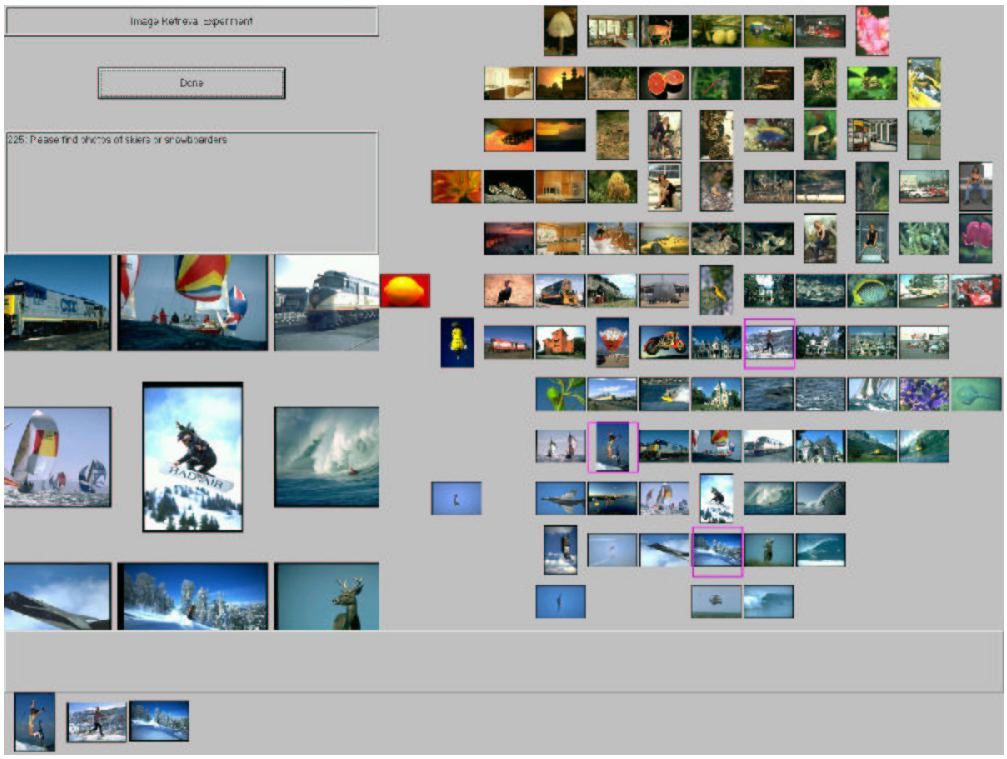


Figure 4.10: The application used to run the second experiment. The description of the required photographs appears in the top left, just below the "Done" button. The participant must then search among the 100 thumbnail images shown on the right, which in this case are arranged according to their visual similarity. On the left there is a 3× linear magnification of the area currently under the mouse pointer. Selected images are highlighted, and copied to the bottom of the screen. An image can be deselected by clicking on the copy, or on the highlighted thumbnail.

As before, each arrangement was created in advance as a single large image, this time in Windows bitmap (BMP) format. The magnification was produced by saving this image at 3000×3000 pixels, and displaying only the portion currently around the mouse pointer at full size. The full size images were 240×160 pixels (2% of the screen area), meaning that the thumbnails were displayed at 80×53 (0.22% of the screen area).

The experiment software was a simple Visual C++ application, written by Wojciech Basalaj. The arrangement bitmaps were very large (27 MB each), and we preferred Visual C++ primarily because it offered more explicit control over memory allocation than Java.

A magnification facility would also have been of assistance to the participants in the first experiment, but was not necessary as they had already seen the target image at full size. In this experiment, however, the participants needed to be able to see each image at a reasonable level of detail, in order to decide whether it matched the description.

4.2.3 Design

We adopted a within-subjects design, so that each participant did two blocks of 10 searches: one using visual arrangements and one using random arrangements. We balanced the design so that half of the participants did the visual block first and half did the random block first. Within a block, the searches were given in a random order. Every combination of the 20 image sets and 20 descriptions was used, so that each participant carried out a different search with a particular set. Before using each arrangement type, the participant did 4 practice searches with it; the image sets used for this were constructed from different categories to those used in the main part of the experiment.

Again, the participants were not told how the arrangements were created, to avoid biasing them against the random condition. Also, given that some participants had not noticed the clustering by visual similarity in the first experiment, we felt that it would again be interesting to study whether this affected performance, and could not have controlled for it in advance.

4.2.4 Procedure

At the beginning, each participant read the experiment instructions (Figure D.4 in Appendix D). In a single trial, the description of the required photographs was displayed, and then the participant had to select as many matching images as possible from the set of thumbnails, clicking on the "Done" button (Figure 4.10) once she was happy with her selections.

We decided that the participants should have a longer time limit (two minutes) than in the first experiment, because the task was more complex than simply locating a given target image. To encourage the participants to complete the task as quickly as possible, without sacrificing accuracy, we offered a financial incentive: they were told that whoever had the best combined ranking on speed and accuracy would be awarded an extra ten pounds. A

similar incentive was used in the experiment described by Veerasamy and Heikes [120].

The participants also filled in a post-experiment questionnaire (Figure D.5 in Appendix D), which asked about their search styles for the different arrangement types, and about what made a search easier or more difficult. The whole procedure took approximately 45 minutes.

4.2.5 Results and discussion

The experiment software logged each participant's actions, giving us a number of possible timings per trial to consider as dependent variables:

- first, the time of the first image selection
- last, the time of the last image selection or deselection (before pressing "done")
- average, the average time taken to make a selection (last divided by the number of selections)
- done, the time at which the participant pressed the "done" button, to indicate that the search was over

We also recorded the sequence of grid cells that each participant covered with the mouse. These trails give some indication of how much active searching the participant did, since the mouse had to be moved over a thumbnail in order to see a magnified version. Therefore, we could also consider

- length, the length of the mouse trail, in grid cells

None of these variables were normally distributed, so we had to choose appropriate transformations to apply to them, before constructing linear models in S-Plus: we used the square root of last and length, and applied a log transform to first, done, and average. There were three observations missing for first, last, and average, where the participant pressed "Done" without selecting any images. The two minute time limit was never exceeded.

Table 4.7 gives the ANOVA results for each of the five response variables. The following predictor variables were added to each model:

- subject, the participant ID (20 levels)
- category, the category being searched for (20 levels)
- set, the ID of the image set (20 levels)
- present, the number of relevant images in the set, from 3 to 7
- type, arrangement type, either visual or random

Initially, we included the sequence number of the trial as a predictor, but it was never significant, and removing it produced better linear models (with higher values of multiple r^2 and lower residual standard error) for all of the response variables.

		first		average		last		done		length	
	Df	F	p	F	p	F	p	F	p	F	p
subject	19	3.03	< 0.001	5.79	< 0.001	9.52	< 0.001	28.99	< 0.001	32.98	< 0.001
category	19	3.10	< 0.001	8.06	< 0.001	5.22	< 0.001	13.27	< 0.001	4.67	< 0.001
set	19	0.65	0.867	1.36	0.145	1.59	0.055	1.60	0.054	1.07	0.386
present	1	11.13	< 0.001	60.59	< 0.001	16.45	< 0.001	0.84	0.361	0.67	0.414
type	1	0.05	0.822	37.79	< 0.001	33.21	< 0.001	7.15	< 0.010	87.35	< 0.001
multiple r^2		29.37%		53.32%		51.89%		71.23%		70.77%	

Table 4.7: ANOVA results for the second experiment. A log transform was applied to first, average and done, and we used the square root of last and length. The residual degrees of freedom are 337 for the first three columns, and 340 for the last two.

Variable	*Visual*		*Random*	
	mean	(s.d.)	mean	(s.d.)
first (sec)	7.79	(7.50)	8.47	(8.03)
average (sec)	6.51	(5.08)	8.32	(5.32)
last (sec)	25.46	(15.16)	30.84	(13.38)
done (sec)	41.49	(17.35)	43.74	(17.27)
pause (sec)	15.90	(10.73)	12.61	(11.97)
length (cells)	54.73	(31.15)	70.80	(30.32)
recall (%)	80.40	(23.12)	80.88	(21.22)
precision (%)	94.63	(17.48)	95.29	(13.03)

Table 4.8: Comparing the arrangement types, for both speed and accuracy.

Differences between arrangement types

The predictor we are most interested in is type. Figures 4.11 and 4.12 compare the distributions of first, last, done, and length for the two arrangement types, and Table 4.8 enumerates the differences in means for each of the response variables. Participants were significantly faster with the visual arrangement for all of average, last, done, and length, but not for first, although its mean value was slightly lower for the visual arrangement.

To measure accuracy, we used adapted versions of the traditional information retrieval evaluation measures: recall is defined as the proportion of the relevant[6] images present in the set that were actually selected by the participant in a given trial, and precision is the proportion of the images selected that were relevant. Their means and standard deviations for each arrangement type are given in Table 4.8.

The precision measure is not particularly interesting in this case, because it simply shows how often the participants confused the categories. As these were deliberately chosen to have little overlap, it is not surprising that precision was usually very high (with a median of 100%). Participants did have difficulties with certain searches, leading to scores of below 100%. For example, some participants thought that dolphins and whales counted as fish, especially

[6] As in the previous chapter, a relevant image is defined as being one that is part of the described category in the original collection.

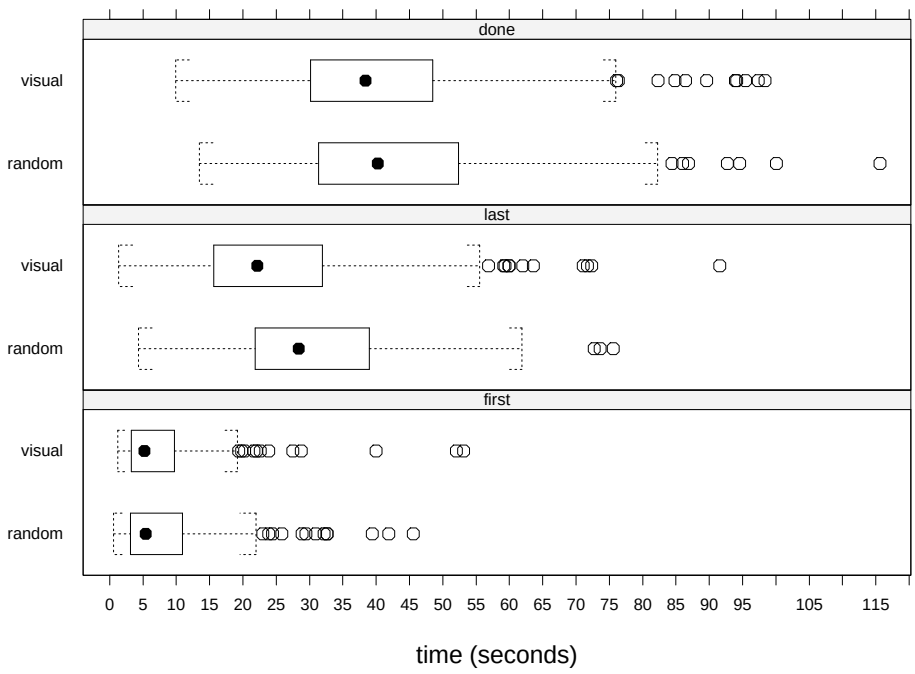


Figure 4.11: Distribution of first, last, and done for the two arrangement types.

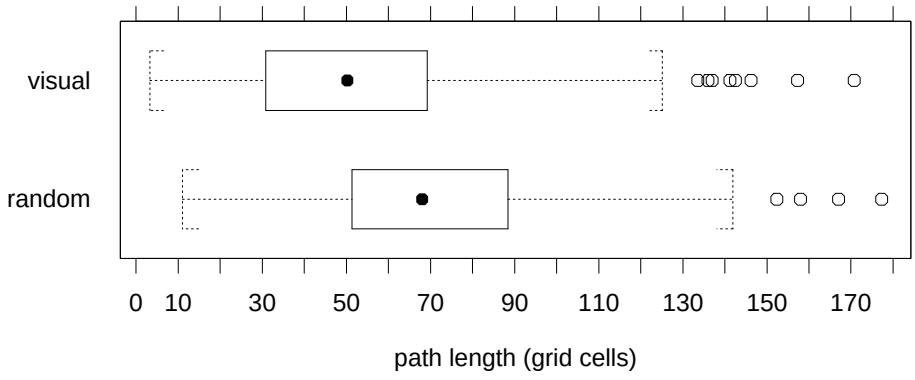


Figure 4.12: Distribution of length for the two arrangement types.

if they were asked to search for "fish" before finding out that "dolphins and whales" was a separate category. The other main source of confusion was the fact that the small invertebrates were sometimes pictured on flowers, meaning that some participants chose these for the "close-ups of flowers" search.

When considering accuracy, therefore, we are primarily interested in recall. Like precision, it has a very negatively skewed distribution, and thus parametric statistical tests are not appropriate. Non-parametric testing is also problematic because the variable can only take on a limited range of values, which would produce many tied ranks in a test. We can get around this problem by aggregating the data, calculating the mean of recall for each arrangement type, per participant, producing twenty pairs of values which can then be compared with a Wilcoxon signed-rank test, giving a p value of 0.784. It therefore seems that there was no significant difference in the level of recall between the arrangement types.

There is likely to be some trade-off between last and recall for this task, because the longer the participant searches, the more relevant images she is likely to find. However, if she has already found all of them, there is no point in continuing until the time limit expires. Because the participants in this experiment did not know how many relevant images were present, they repeatedly had to decide whether they should continue searching, or press "Done". It is interesting to note, then, that participants paused for longer with the visual arrangement between making their last selection and pressing "Done": we created a new response variable, pause, by subtracting last from done (Table 4.8), and the difference between arrangement types is significant ($F(1, 337) = 26.13$, $p < 0.001$). This suggests that with the visual arrangement, participants were less likely to feel that they had found all of the relevant images, even though, as we have already seen, there was actually no significant difference in recall between arrangement types.

Four participants commented on this in the questionnaire; it seems that although they were often able to find a cluster of relevant images with the visual arrangement, they still felt that it was necessary to check the rest of the set, in case some images from the category were not visually similar to those in the original cluster. This is probably an artefact of the experimental conditions.

02 *With [visual], I expected some photo hidden on the edge which I would miss.*

08 *[With random] I felt like the targets were more dispersed, so I needed to look everywhere. [...] I felt like I'd had a brief glance at every single option [...], whereas with [visual], I felt less sure about having found all the pictures.*

18 *With [visual] I was always worried that I'd missed a huge group whereas with [random] I thought if I'd missed any it would only be 1 or 2.*

Table 4.7 shows that present is significant for three of the response variables: when there were more images present, the first image was found more quickly, and it took less time on average to find each one, but it took longer in total to find all of them. It is also significant for the new variable, pause

Variable/noticed		*Visual*		*Random*		*Significance*	
		mean	(s.d.)	mean	(s.d.)		
first (sec)	Y	5.95	(4.33)	9.16	(8.04)	$F(1,110) = 3.25$	$p = 0.074$
	N	9.01	(8.82)	8.02	(8.03)	$F(1,187) = 1.71$	$p = 0.193$
average (sec)	Y	4.89	(2.60)	8.88	(6.50)	$F(1,110) = 34.81$	$p < 0.001$
	N	7.59	(5.98)	7.96	(4.37)	$F(1,187) = 1.92$	$p = 0.168$
last (sec)	Y	19.20	(10.56)	28.92	(12.00)	$F(1,110) = 28.27$	$p < 0.001$
	N	29.62	(16.32)	32.11	(14.12)	$F(1,187) = 2.59$	$p = 0.109$
done (sec)	Y	33.10	(11.27)	39.77	(16.47)	$F(1,112) = 11.69$	$p < 0.001$
	N	47.08	(18.45)	46.39	(17.35)	$F(1,188) = 0.18$	$p = 0.675$
pause (sec)	Y	13.57	(7.72)	10.09	(10.85)	$F(1,110) = 9.80$	$p < 0.010$
	N	17.44	(12.11)	14.28	(12.42)	$F(1,187) = 8.17$	$p < 0.010$
length (cells)	Y	37.01	(17.67)	60.94	(25.55)	$F(1,112) = 60.69$	$p < 0.001$
	N	66.54	(32.62)	77.38	(31.54)	$F(1,188) = 17.61$	$p < 0.001$
recall (%)	Y	74.85	(25.29)	73.79	(24.26)	not tested	
	N	84.11	(20.85)	85.61	(17.49)	not tested	
precision (%)	Y	93.66	(20.38)	95.32	(14.79)	not tested	
	N	95.28	(15.30)	95.26	(11.77)	not tested	

Table 4.9: Comparing the arrangement types, dividing the data according to whether participants noticed or did not notice the difference between them.

($F(1,337) = 29.02$, $p < 0.001$): when there were more images present, participants paused for a shorter time, probably because they were likely to have found a larger number of images and thus felt more confident that they had found all of them.

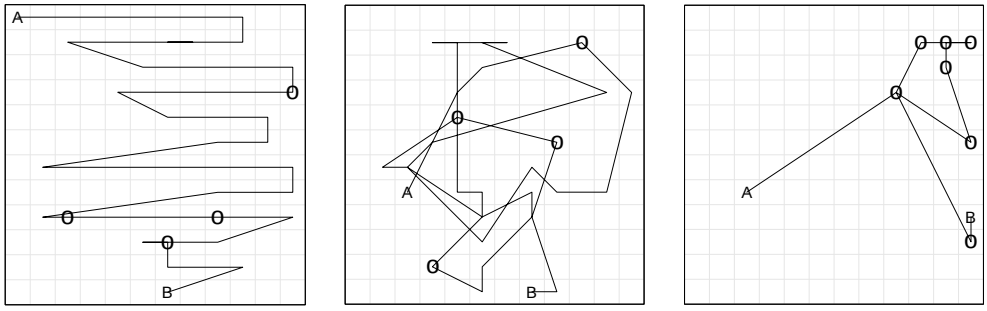
06 *If there were only 2 pictures found, I kept on looking. If 10, I stopped.*

Table 4.7 also shows that the predictors subject and category are significant for all of the response variables: some participants did the task more quickly than others, and some searches were easier than others. We consider both of these in the following sections.

Differences between participants

When asked about their search styles, 12 of the 20 participants did not mention making use of the clustering of visually similar images in the visual arrangement, which suggests that they had not noticed the difference between the two arrangement types. Those who did notice were relatively much faster with the visual arrangement than those who did not, for whom no significant improvement over the random arrangement is indicated, except with regard to length of mouse trail (Table 4.9). In contrast to the results of the first experiment, for this task participants needed to be consciously aware of the arrangement according to visual similarity in order to take advantage of it when browsing.

However, the mean recall of those who noticed was lower, *regardless of arrangement type*, which perhaps indicates that those participants who were inclined to prioritise speed over accuracy were also more likely to notice the



(a) Systematic. (b) Haphazard. (c) Efficient.

Figure 4.13: Three examples of mouse trails from individual trials, illustrating the main searching styles. The start of each trail is marked with an **A**, and the end with a **B**. The circles show the positions of the relevant images; the systematic and haphazard examples are both taken from searches involving random arrangements, but the efficient style can only be used with a visual arrangement.

difference between arrangement types. We have already mentioned that a participant's decision to press "Done" will depend on how confident she is that all of the relevant images have been found, but of course it will also depend on how important accuracy is to her, in relation to speed. This is reflected in the fact that pause was lower for the participants who noticed, for both arrangement types; a t-test comparing log(pause) for the two groups of participants is significant, $p < 0.01$.

To investigate this, we looked more closely at participants' search styles: as well as considering their own general descriptions from the questionnaire, we studied plots of the mouse trails for each trial (for example, those in Figure 4.13). It is important to note that these only show which image the mouse pointer was over, which is not necessarily the image that the participant was actually looking at. Some participants did tend to use the mouse pointer to keep track of where they were looking on the overview display, but others only moved it to a thumbnail if they wanted to see the magnified version.

We noticed that there were three main search styles. The **systematic** approach involved covering the screen from top to bottom, usually in a boustrophedonic[7] manner (as in Figure 4.13(a)).

10 *I started from the top left hand corner image running the mouse over all of the pictures in sequence and looking at the enlarged versions of the images which weren't so clear.*

16 *Mainly left-right top-bottom scan, occasionally jumping if something relevant caught my eye.*

[7]defined in the Oxford English Dictionary as "alternately from right to left and from left to right, like the course of the plough in successive furrows".

In contrast, participants who chose the **haphazard** style attended to different areas of the screen in no apparent order, as they noticed relevant images (as in Figure 4.13(b)).

18 *Scanned all the pictures looking for common/stereotypical things associated with each category, then after spotting the obvious ones looked slightly more closely for ones that were less obvious.*

Finally, the participants who noticed the difference between arrangement types were able to take advantage of the fact that relevant images tended to be clustered together in the visual arrangement by adopting an **efficient** style, usually quickly scanning the whole screen in order to find an area containing a cluster, then moving to that area to look for relevant images within it (as in Figure 4.13(c)). This made them much faster with the visual arrangement than the participants who did not notice, who simply tended to use the same style (either systematic or haphazard) for both arrangement types.

02 *Started in a certain area of the screen (photos were usually close together), then moved on from there.*

16 *Quick scan of the whole layout, then jump to an area of interest.*

18 *Images that fitted each category seemed to be grouped more closely together so looked for clusters.*

We wondered if those participants who were inclined to prioritise speed over accuracy were more likely to adopt a haphazard approach, and if this in turn made them more likely to notice the difference between arrangement types, because they did not attempt to follow a fixed route through the image set. Similarly, those participants for whom accuracy was more important than speed might have decided to search systematically, which could have made them less aware of a difference between arrangement types.

To compare the haphazard and systematic styles directly, we decided to look only at the data from the random arrangement, for which all of the participants had to adopt one of these two styles. Although there is too little data in each cell to draw any definite conclusions, Table 4.10 shows that, as expected, the participants who searched haphazardly seemed to be faster than those who searched systematically, and obtained a lower level of recall, on average.

However, it also shows that there seems to be *no relationship* between the search style adopted by a participant, and the likelihood of her noticing the difference between arrangement types. Plus, of the participants who adopted a systematic style, those who noticed the difference between arrangement types were faster, and had a lower level of recall. We are unable to explain why the participants who noticed the difference between arrangement types seemed to prioritise speed over accuracy, and it would be interesting to replicate this experiment to see if the same phenomenon occurred.

Fewer participants noticed the similarity-based organisation in this experiment than in the previous one, probably because this time it was less obvious

Style	*Noticed*	
	yes	no
systematic	5	8
	30.74 sec	35.34 sec
	76.73%	88.99%
haphazard	3	4
	25.95 sec	25.64 sec
	68.90%	78.84%

Table 4.10: Here, the participants are split into groups according to whether they adopted a systematic or a haphazard searching style for the random arrangement, and whether or not they noticed the difference between arrangement types. Each cell of the table contains (from top) the number of participants represented by the cell, and the means of last and recall for those participants, using the data from the random arrangement only.

that there was any difference at all between the arrangement types, as they were both presented as grids. Ironically, our improvement of forcing the images into a grid to remove overlap may have made participants less likely to notice the organisation by visual similarity, thus preventing them from taking advantage of it. In a real system, however, users would be aware in advance of any organisation according to visual similarity, meaning that these issues regarding noticing or not noticing are unlikely to be important.

Interestingly, when asked to express a preference, the eight participants who had noticed the difference were split between the two methods. Five said that they preferred the random arrangement, four of these because they felt more confident that they had found all of the relevant images, as discussed above. Only three preferred the visual arrangement; for these participants it would seem that the increased efficiency of the search outweighed any concerns they may have had about missing out relevant images.

04 *Easier to use expectation & knowledge to cut down the search.*

16 *There was usually (but not always) a correlation enabling you to find clusters of useful images.*

20 *Narrowed search down straight away.*

Differences between searches

As in the first experiment, the participants were asked about the factors which made a search easier or more difficult, and obviously these had a large impact on both search time and recall. Again, the salience of the individual images was important, with participants commenting that searches were easier if the relevant images contained strong colours, high contrast (especially between subject and background), or large objects, recognisable at thumbnail size. More than half of the participants said that it was difficult to find images where the subject had little contrast with the background, and dull colours and small objects were also mentioned. Considering the group of images being

searched for, most of the participants said that consistent appearance within the group made them easier to find, especially if they did not resemble images from any other group. Two participants mentioned that both the "interiors of houses" and "female fashion models" searches had these qualities. On the other hand, it was difficult at first glance to separate images of surfing from those of sailing, or dolphins.

Table 4.9 shows that the participants who had noticed the difference between arrangement types tended to find the first relevant image more quickly with the visual arrangement than with the random arrangement, although the difference is not quite significant. As we expected, it seems that they were often able to form a mental image of a relevant photograph and then were assisted in searching for it by the organisation according to visual similarity. In the questionnaire, 16 of the participants agreed that they formed mental images from the descriptions, and the other 4 were neutral. Five participants specifically mentioned that a search was easier if the relevant photographs had a predictable appearance; for example, sunsets were easy to find because it was likely that they would be orange. However, birds could be pictured against a number of different types of background, and several of the hot air balloons were of unexpected shapes.

16 *For some descriptions, it was easy to translate a "mental image" into an idea of colour e.g. beach – blue/sandy, skiers – white, etc. For other descriptions e.g. house interior it was hard to predict what colour they might be.*

Of course, as soon as the participants did their first search, they saw examples of images that would be relevant to later descriptions, and this perhaps gave them a basis from which to form their mental images.

Two-dimensional precision revisited

In the previous chapter, we described how performance indicators from conventional information retrieval can be applied to an MDS arrangement, allowing us to measure how well it has clustered images of the same generic type. We were interested in finding out whether this would be related to the participants' speed of locating relevant images in this experiment, and therefore calculated an average two-dimensional precision value for each trial. The participant's first selection could have been any of the 3–7 relevant images present, and so we calculated the average precision at each of these in turn, and then took their mean.

Overall, the mean value of average precision was 0.168 for the visual arrangements, and 0.071 for the random arrangements. This confirms that images from the same category tended to be more closely grouped together in the former. Note that in Tables 4.8 and 4.9, the mean of average is lower than the mean of first for the visual arrangement, but they are about the same for the random arrangement. This seems to suggest that once the participant had found the first relevant image in the visual arrangement, she was able to find the others more quickly because they were usually close to it.

08 *I mainly used first looking around, finding one target and assuming that the other targets would be in the same area. Once I'd found one, it was easier to find the others than [with random].*

To investigate this further, we tried adding average precision to our linear models as a predictor, and then performed ANOVAs. It was significant for all of average, last, done and length; a higher average precision value meant a faster search. It may therefore be possible to use this measure of the degree of clustering of the relevant images to predict how quickly a user will find them. However, we have already seen that average precision is related to arrangement type, and it is also correlated with the number of relevant images present in the set ($r = 0.277$, $p < 0.001$): obviously, it tends to be higher when $7/100$ are relevant than when $3/100$ are relevant. These relationships make it difficult to separate the effect of average precision from that of type and present, and therefore these results may be unreliable.

We also found in the previous chapter that there was very little to choose between arrangements created using different measures of visual similarity. The iris measure was used for the visual arrangements in this experiment, but out of interest we subsequently produced versions which were based on the a.1 measure. We found that their average precision values were comparable to the originals, and therefore believe that if we replicated this experiment using a.1-based visual arrangements, the results would be very similar.

4.3 Conclusions

The first experiment found that the participants could locate a given image more quickly in a visual MDS arrangement than in a randomly arranged grid of thumbnails, although the salience of the target image within the set seemed to be the dominant factor. In post-experiment interviews, many of the participants said that the overlap in the continuous MDS arrangement had given them problems, and that they preferred the regularity of the random grid; this result inspired the subsequent development of the proximity grid algorithms, with which the best aspects of the two arrangement types can be combined.

In the second experiment, the participants had to find a group of images matching a requirement based on generic content. They were faster when the set of thumbnails was arranged according to visual similarity (in a proximity grid) than when it was arranged randomly, especially when they were consciously aware of the difference between the arrangement types. However, with the visual arrangements, the participants generally paused for longer after making their last selection, because they felt less sure that they had found all of the relevant images. The primary advantage of the visual arrangements was that they generally grouped the relevant images together, so that once the first image had been located, it was easier to find the others. A visual arrangement could also be helpful for finding the first image, if the participant was able to form an accurate mental image of the general appearance of relevant photographs before starting to search.

As we noted in Section 2.6, a number of studies of real image collections have found that requests for generic content are less common than those for specific content, such as named people or places. Specific names must be recorded using annotations, so that the user knows what a photograph actually depicts. The owners of image collections have already made a lot of investment in annotation, and this enables users to browse the contents of a category, or the results of a text-based query. A semantically related set of images like this could be arranged according to visual similarity, which may help the user to get an overview of the different types of photographs that match the requirement, and then choose those which are most visually suitable. This is what we consider in the next chapter.

Chapter 5

Simulated work tasks

The experiments in the previous chapter involved tasks where the participants were asked to locate a known image, and a group of images matching a generic requirement. In Section 2.6 we noted that studies of the requests submitted to general-purpose photograph collections have found that users most commonly have requirements at the specific level, such as for named people or places. To satisfy a requirement of this kind, the photographs must be suitably annotated, so that the user can enter a text query (or navigate to a named category), and then browse the resulting set of thumbnails.

We were interested in finding out whether arranging this set according to visual similarity would help the user to make her final selection(s) from within it. We again used the IRIS visual similarity measure, which was introduced in Section 3.1.4. As we noted in Section 2.6.1, Markkula and Sormunen [77, 78] found that journalists usually preferred to issue a broad textual query and then browse through the results, applying secondary selection criteria (which included some based on visual attributes and aesthetics) while browsing, rather than specifying them in the initial query.

When the collection is annotated, a text-based similarity measure can be used in conjunction with MDS to arrange a set of images. In theory, this should result in the most meaningful automatic organisation possible, grouping together the photographs that are similar at whichever levels of content are covered by the captions. Again, this may be helpful to the user in making selections from the set, and therefore we also consider caption-based arrangements in this chapter. The caption similarity measure we used is described in Appendix B, and Figure 5.1(a) gives an example of an arrangement created using this measure, identifying some of the resulting clusters.

We carried out two experiments, plus a follow-up, and all of these were based on a simulated work task [13] that involved selecting photographs to be used as illustrations, using designers as our participants. As we mentioned in Section 2.5.3, Jose, Furner, and Harper [60] adopted a similar approach in their evaluation of a spatial query interface. Its findings were based only on the participants' questionnaire responses, without considering the efficiency of their actions.

5.1 The infodesign 99 study

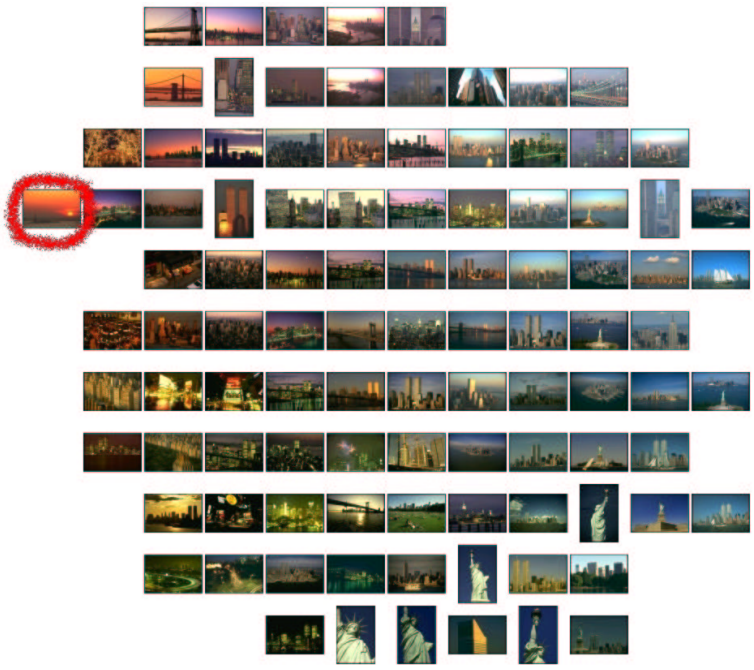
Initially, we were simply interested in investigating whether designers would find either type of similarity-based arrangement useful for selecting images, and also if it was helpful to have both types of arrangement available at once. For example, imagine that a designer is looking for some photographs of New York, and has both of the arrangements in Figure 5.1 available. In the caption-based arrangement, the circled image of the Statue of Liberty at sunset appears next to all of the other Statue of Liberty photographs. In the visual arrangement, however, the same photograph is grouped with the other sunsets, giving the designer a different perspective, and perhaps assisting her in finding a group of images that work well together visually.

In keeping with the simulated work task methodology, we wanted to set a reasonably realistic task, giving the participants a role in a situation, with a vague requirement that they would be able to develop further on their own. We settled on the domain of travel guides, because approximately 150 of the 400 categories in the Corel collection are focused on a single place (a continent, country, region, or city); the category names are listed in Appendix C. The participants were therefore told that they had been asked to choose photographs to illustrate a set of destination guide articles for a new travel World Wide Web site. Each article was to be an overview of a different location, to appear on a separate page. However, as it would probably have taken a long time for them to read a real article, we gave them summarised versions, and told them that these conveyed the general impression of the final article.

The participants had a set of 100 photographs available for each location, and their task was to choose three of them, to be used together, for each article. We asked them to select three, rather than just one, because we were interested in exploring whether the similarity-based arrangements would be helpful for finding images to work well together as a set. Markkula and Sormunen [77, 78] found that the journalists in their study would try to choose images that complemented those already selected for the page, looking for a balance of different styles and colours. They also mentioned that the journalists were willing to browse through large numbers of thumbnails, only reformulating their queries if they had more than about 100 results; in this experiment, the 100 images in a set were intended to represent either the contents of a category, or the top 100 results of a query for the place to be illustrated. All of the images were already relevant, in that they were related to the topic of the requirement, so the participants had to apply further criteria in order to decide on their final selections.

Unlike our previous experiments, we chose to tell the participants how the arrangements were created, because this was not immediately obvious in the case of the caption-based arrangement. We decided not to make a random arrangement available, because (as noted below) we only had a small amount of time with each participant, but also because it was difficult to see how we could explain it without making it seem like an obvious straw man.

(a) Caption similarity. Some of the resulting clusters are highlighted, with outlines and labels, but the experiment participants did not see these.



(b) Visual similarity.

Figure 5.1: Two 12×12 MDS proximity grid arrangements of 100 images of New York, as used in the infodesign 99 study.

5.1.1 Participants

To obtain access to a community of designers, we set up the experiment at the infodesign 99 information design conference[1]. Our 18 participants were all attendees, who volunteered their time during the half-hour conference breaks, so this gave us a tight time constraint. All had normal or corrected-to-normal vision, with no colour blindness (self-reported).

Beforehand, we carried out a pilot study, with one participant: a system administrator, who is also an amateur artist. The original design had one practice search, and then four further searches in the main part of the experiment, but this was cut to three because the pilot participant took longer than expected to complete the task. In addition, some changes were made to the experiment instructions, to emphasise that there were no correct answers, and it was therefore entirely up to the participants to decide on the criteria they would use to make their selections; in particular, they did not have to adhere slavishly to the given text.

5.1.2 Apparatus

We selected four places from the categories in the Corel collection: these were Paris (223000), Kenya (253000), Alaska (309000), and New York (244000), which was used as a practice. For each of these, we created two 12×12 proximity grid arrangements[2] of the 100 images, one arranged according to caption similarity, as in Figure 5.1(a), and the other according to visual similarity, as in Figure 5.1(b). We used two PCs, both running Windows NT 4 (one had 192MB of memory and the other had 128MB), with 17-inch monitors set at 1600×1200 resolution. Each was used for half of the participants, and usually there were two people working at the same time. The thumbnail images were 96×64 pixels, or 0.32% of the screen area, and the magnified images were 288×192, 2.9% of the screen area.

We selected places where the available Corel images were of good technical quality, and corresponded reasonably well to the pieces of text. These were abridged from the online *Rough Guide*[3], to a length of approximately 200 words each, and are given in Appendix D, starting on page 208.

To display the images to the participants, and allow them to make their selections, we used a modified version of the Visual C++ program described in the previous chapter, illustrated in Figure 5.2. The two different arrangements of the image set were named "Caption" for caption-based similarity, and "Image" for visual similarity; the participants could switch between them using radio buttons. The highlight on selected images was retained after switching, to allow the participants to see how the surrounding context of their selections changed. Up to four images could be selected at a time.

The overview display was allocated a larger area of the screen than before (at 1600×1200 resolution, it was 1200×1200 pixels instead of 1000×1000)

[1] held in Cambridge in July 1999. See `http://www.idu.co.uk/id99/`

[2] using the Newton-Raphson method followed by a greedy algorithm; see Appendix A.

[3] `http://travel.roughguides.com/`

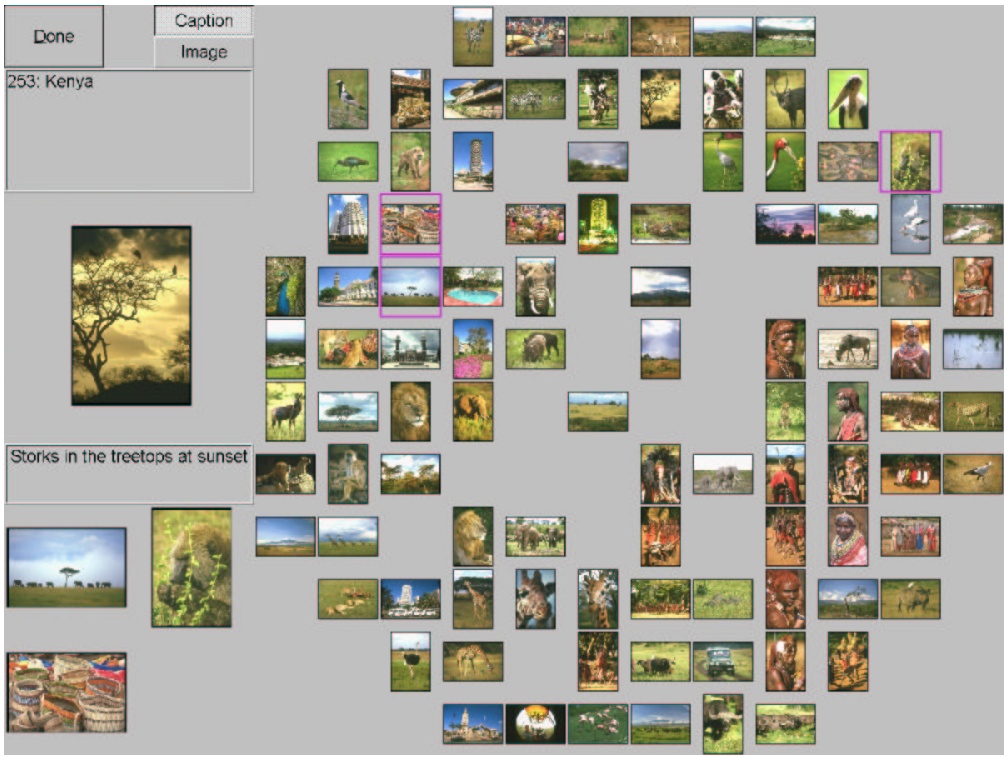


Figure 5.2: The experiment software used for the infodesign 99 study, showing a caption-based arrangement of 100 images of Kenya. Compared to the previous interface (Figure 4.10), the overview display is larger, and the magnified area (centre left) is smaller. The label just below this area shows the caption of the image that is currently under the mouse pointer. The radio buttons at the top left were added to allow the user to switch between the two different arrangements of the image set. This time, up to four images can be selected at once; these are copied to the area at the bottom left of the screen.

to allow the thumbnails to be slightly bigger. This meant that the magnified area had to be smaller (400×400 pixels instead of 600×600); the images were still shown at $3\times$ zoom, but with fewer simultaneously visible. The caption of the image currently under the mouse pointer was displayed, to give the participants more detail about its content. As before, user actions (selections, switches, and mouse movements) were logged and timed.

5.1.3 Design and procedure

Firstly, participants read the experiment instructions, shown in Figure D.6 in Appendix D. Then, they had one practice search (New York), during which the software was demonstrated to them, and they were allowed to practise using it until they felt comfortable. We explained the two arrangement types, and showed the participants how to switch between them. In the main part of the study, they did three further searches (Paris, Kenya, and Alaska). All six possible orderings of these three places were used.

Statement	*Agreement*					*Mean*	*Med*
	0	1	2	3	4		
"The arrangement of photos by **caption** similarity was useful"	0	3	3	9	3	2.7	3
"The arrangement of photos by **image** similarity was useful"	3	4	3	6	2	2.0	2
"The task would have been just as easy with a random arrangement"	6	3	3	2	4	1.7	1.5
"It was useful to have two different views of the same set of photos"	1	4	5	4	4	2.3	2
"The two views complemented each other well"	1	6	6	3	2	1.9	2

Table 5.1: Responses to five of the questionnaire items from the infodesign 99 study. 0 represents "strongly disagree" and 4 represents "strongly agree".

A single search proceeded as follows. First, the name of the place was displayed on-screen, and participants read its associated text, which was provided on paper. They had to choose which arrangement type to use first, by pressing its radio button. This caused the set of images to appear, and also started the timer for the search. Participants were then free to select or deselect images until they were satisfied with the chosen three, at which point they pressed the button marked "Done" (see Figure 5.2). There was no time limit, but they were encouraged to work quickly. While searching, they could switch between arrangement types as often as they wanted.

Once they had completed the experiment, participants filled in a questionnaire (Figure D.7 in Appendix D), where they indicated their agreement or disagreement with a series of statements, and could write down further comments if they wished. The original questionnaire used a scale of 1 (highest rating) to 5 (lowest rating), but to be consistent with the rest of this dissertation, in the following tables the scale has been changed to 0 (lowest) to 4 (highest).

5.1.4 Results and discussion

We expected that the participants would find both the caption-based arrangement and the visual arrangement to be useful in their own right, and also in tandem. This section gives summaries of the participants' questionnaire responses, as well as data about their actual usage of the experiment software, gathered from the log files. Again, quotes from the questionnaires are marked with a participant ID number; these include some comments from the pilot participant (**P**).

Usefulness of different arrangement types

Table 5.1 shows that two-thirds of the participants agreed that the caption-based arrangement was useful. Participant 03 commented that "it gave me a breakdown of the subject", and participant 04 said that it helped her to link the images to the text, as did the pilot participant:

P *If the text mentioned somewhere then I could find all the pictures of that particular place more quickly.*

Opinion was more divided on the usefulness of the visual arrangement. Participant 12 preferred it because it seemed to help him with finding generic

images:

12 *I preferred the image layout as it was easier if for example I wanted a picture of a lion it was easier and quicker to find the animal images and then search from there.*

Eight people rated caption as more useful than visual, three gave visual a higher rating, and seven gave them the same score. The ties make statistical analysis difficult, but a two-tailed Wilcoxon signed-rank test gives a p value of 0.099, so the difference in ratings between arrangement types is significant only at the $p < 0.1$ level.

Nine people disagreed with the statement "the task would have been just as easy with a random arrangement", meaning that half of the participants felt they got something useful from at least one of the arrangements. Some participants mentioned that they would always want to look at every available image before making their choice, and were therefore inclined to think that it made no difference to them how the images were arranged.

P *Because I was trying to find the most aesthetically pleasing pictures I found I had to look at all of them, so the arrangement didn't particularly matter.*

In 40 of the 54 searches, participants chose to look at the caption-based arrangement first, and on average spent 63% of their total search time using it, and 37% using the visual arrangement (the mean time taken per search was 140 seconds). These averages hide a large amount of variability between participants in the time spent using each arrangement type. Seven participants heavily favoured the caption-based arrangement (using it for 85% of the time, or more); four of these used it exclusively, and rated the visual arrangement's usefulness as 0, 0, 1, and 1. Three participants heavily favoured the visual arrangement (using caption-based for 22% of the time, or less), and the remaining eight showed no obvious preference (spending between 39% and 69% of their time using the caption-based arrangement).

Six of the latter group also agreed that it was useful to have two different views of the same set of photographs (eight agreed in total; see Table 5.1). Participant 16 said that it was "useful to check both". Four of these (five in total) also felt that the two arrangements complemented each other well:

01 *The caption mode I preferred to start with, that made sure there was enough diversity "contentwise" (subject related) then I'd use the image mode to make sure they complemented each other as a set.*

08 *I used "image" first then checked out my selections by using "caption".*

The difference in usage time is also reflected in the number of selections made with each arrangement. Six participants made all of their selections (including those images subsequently deselected) using the caption-based arrangement, and one (05) made all of them using the visual arrangement. The others were somewhere in between, and on average, the participants made 61% of their selections using the caption-based arrangement.

In two-thirds of all searches there was at least one switch between arrangements; the most switches in any one search was four. Seven participants selected an image in one arrangement and then deselected it in the other; one did this seven times, and two others did it five times each, suggesting that they were indeed taking advantage of having two different views of the same image set. However, one participant stated explicitly that he disliked this facility:

05 *Moving from one style of selection to another was distracting and broke concentration — I did it less and less as I went through.*

Of course, the caption-based arrangement can only be as good as its captions, and two participants commented on its limitations for considering different levels of meaning. For example, in the "Kenya" set (visible in Figure 5.2), the wildlife photographs are scattered around the screen, because the captions only contain the names of the individual species, such as "lion" or "giraffe", and no generic word like "animal" to connect them.

07 *Meaningful captions — especially with Alaskan exercise — animal related shots were spread out all over screen in both views, but would prefer them all together for speed of selection. Captions including terms such as "wildlife", names of regions (Lower East Side, or Left Bank) might have been useful.*

08 *Caption could be extended to include broader subjects, e.g. Modern Architecture of Paris.*

Rose and his colleagues [102] have experimented with using the WordNet lexical database to automatically expand image captions (adding more specific and more general words), and we believe that an approach like theirs could lead to better-structured caption-based arrangements. For example, the most generic terms could be used to create top-level clusters such as "animals" or "buildings", with more specific terms being used to determine the substructure of each cluster. Of course, for the particular case where the images are of places, an especially useful arrangement would be one based on geography; the appropriate metadata would be immediately available if the photographs were originally taken with a camera equipped with a Global Positioning System (GPS) receiver.

Interestingly, two participants felt that the caption-based arrangement was useful because it did *not* group together visually similar images:

05 *[I preferred] caption because it randomises the arrangements.*

15 *[I preferred] caption. It made me choose visually the pictures that caught my attention and decide due to the picture itself and not the theme or similarity.*

In Section 4.1.5, we discussed how a random arrangement can produce more local contrast than a visual arrangement, causing individual images to be more salient. Because semantically related images are not necessarily visually similar, it seems that a caption-based arrangement also has this property. We will return to this issue later in the chapter.

Criterion	*Importance*					*Mean*	*Med*
	0	1	2	3	4		
Technically good, striking photographs	0	0	5	5	8	3.2	3
Photographs that were relevant to the given article	0	0	0	6	12	3.7	4
Photographs that worked well as a set of three	1	2	4	4	7	2.8	3
Photographs that fitted your own impressions of the place	0	3	5	8	2	2.5	3

Table 5.2: Questionnaire responses regarding participants' selection criteria. 0 represents "not at all important" and 4 represents "very important".

Selection criteria

As we mentioned earlier, we assumed in this study that the user has already carried out some restriction of the image collection, by issuing a query, or opening a category, and that all of the images in the displayed set are relevant to the given requirement. Table 5.2 shows how important the participants felt that certain selection criteria had been to them while choosing from this set. The only statement in the whole questionnaire that every participant agreed with was the importance of finding images that were relevant to the given text, even though the experiment instructions had encouraged them not to rely too much on it.

07 *[I was particularly satisfied when I] was able to find photographs which I felt matched the written "overview" of the proposed article.*

This was regarded as more important, on average, than the technical quality of the photographs, and is perhaps a factor in the favour shown to the caption-based arrangement. A few participants mentioned that they would have used the visual arrangement more in a different type of task, where meaning was less important.

Eleven of the participants agreed that they had wanted their selections to work well together as a group.

05 *In all cases it was a combination of editorial sufficiency and visual completeness — a set.*

Figure D.8 in Appendix D shows the images that were most commonly selected. Other selection criteria offered by the participants were:

02 *Reinforce stereotypes/clichés of the places or avoid stereotypes/clichés.*

03 *Colour or "mood" of the photo, e.g. NYC — I didn't want to use a calm sunset because the article described it as energetic.*

06 *Photographs that induce the intended image of the place in the viewer, set a mood.*

08 *Ones that I thought could be successfully cropped to give me what I was after.*

A number of participants expressed dissatisfaction with the images available, often because of their technical quality, or because it was difficult to find a photograph to suit a particular aspect of the text. One said that if you do not have the right images, it makes no difference how they are arranged.

02 *I didn't think that the images particularly captured the excitement of the text.*

14 *My impression was that the sets of photos didn't suit the textual descriptions. I couldn't find images that matched the descriptions satisfactorily.*

16 *No beach pictures for Kenya. No Anchorage pics for Alaska. No choice of river pics for Paris.*

The user interface of the experiment software

There were no major problems with the experiment software; participant 02 commented "the program is very easy to use and makes the task straightforward". However, a few suggestions for improvements were made.

Two participants (05 and 08) said they would have liked the space to make more than four interim selections, suggesting nine or twelve. Participant 18 said it would have been useful to be able to make "heaps" of related images, so that one could be chosen from each group. He would also have liked to delete images from the arrangement that were definitely unwanted. Participant 06 said that the user should be able to rearrange the chosen images, perhaps seeing them at a larger size with some sample text on a mocked-up page.

Participant 11 also wanted "a chance to see the 3 chosen ones on a separate screen, bigger, together". Participant 01 said that he would have liked to see close-ups of the images, so that he could decide if a picture could be cropped to use a detail from it. Opinion was divided about whether the magnified images were too small (Table 5.3), but ten of the eighteen participants agreed that they were. Participant 13 said that although she would have liked the magnified photographs to be bigger, she understood the trade-off that had to be made in terms of screen space, and felt that "having all the thumbnails on one screen is preferable to scrolling through several screens of larger thumbnails".

Participant 03 said that "the hierarchy of small image with larger on the side worked well for me", although two participants (06 and 14) said that the thumbnails were too small. Participant 14 said he would have preferred to use more screen space for the thumbnails, removing the dedicated area for the magnified view and replacing it with a pop-up magnification.

Finally, one participant suggested that it would be useful to have a facility to issue queries based on the captions, highlighting the matching photographs.

Other details

Table 5.3 shows that only two of the participants were unclear about the task, and three did not think that the task was realistic. They stated that a more realistic task would involve a more detailed specification of the intended page layout of the article.

Information designers tend to work primarily with text, and only eleven of our eighteen participants had prior experience of carrying out picture selection. These eleven spent 49% of their time (on average) using the visual arrangement, compared to 19% for the other seven, and gave it a mean usefulness rating of 2.2 (median 3), compared to 1.0 (median 1).

Statement	*Agreement*					*Mean*	*Med*
	0	1	2	3	4		
"I had a clear idea of what I was supposed to do"	0	2	3	5	8	3.1	3
"I thought the task was realistic"	1	2	4	7	4	2.6	3
"I have prior experience of carrying out picture selection"	3	3	1	4	7	2.5	3
"The articles gave me 'mental images' of possibly suitable photos"	0	0	2	8	8	3.3	3
"The magnified photos were too small"	3	2	3	7	3	2.3	3

Table 5.3: Further questionnaire responses. 0 represents "strongly disagree" and 4 represents "strongly agree".

As in the previous experiment, almost all of the participants agreed that the texts gave them mental images of possibly suitable photographs (Table 5.3), corroborating one of the findings of the experiment conducted by Jose, Furner, and Harper [60].

5.2 The Anglia experiment

Because of time constraints, the participants in the infodesign 99 study had only a short period to form an opinion of each arrangement type. We felt that the findings were sufficiently interesting to warrant a further experiment, where the participants were given more time to become familiar with each one. This experiment was more formal, returning to the model of directly comparing an arrangement based on visual similarity to a random arrangement. As well as finding out if designers would prefer the similarity-based arrangement to a random one, we were interested to know if it would help them to carry out the given task more quickly. We did not use a caption-based arrangement, primarily because we felt that it would be too complex and time-consuming to compare three arrangement types in this way, but also because of the inconsistencies introduced by the lack of detail in the Corel captions, which we discussed earlier.

We had chosen not to use a random arrangement in the infodesign 99 study, because of our concern that the participants would be biased against it from the beginning by being told that it was simply random. For this experiment, our solution was to name it "Library"; we told the participants that the images were arranged in the order that they had been provided by the creators of the stock photograph library[4]. The experiment instructions were also carefully worded, to try to ensure that both arrangements were presented as being equally valid.

Some of the infodesign 99 participants felt that the task would be more realistic if a sample page layout was provided. For example, Jose, Furner, and Harper [60] gave their participants a template for the layout of a leaflet, rather than a piece of text. We consulted a lecturer in graphic design from Anglia Polytechnic University in Cambridge, who was also our contact for recruiting participants. His opinion was that it was not necessary to specify a page lay-

[4]We did consider actually using this ordering, but were dissuaded by the fact that in some categories it was obviously meaningful, whereas in others it was not.

out, and that in fact it might be too constraining: a designer would probably expect to have some control over page layout, and so a normal picture selection process would be more iterative and creative than simply slotting photographs into pre-defined spaces. As we noted in Section 2.5.3, Garber and Grunes [47] found that art directors at advertising agencies tend to be responsible for a project's overall artistic concept, as well as selecting or commissioning suitable images.

The participants in the infodesign 99 study expressed a wide range of preferences: some liked both types of arrangement, some liked only one, and others liked neither. It was difficult to identify the source of these individual differences, beyond the advice of another conference attendee, that "you can't expect designers to agree on anything…". We wanted to consider individual differences in more depth in this experiment. A number of previous studies (such as [23, 116]) have found that participants' spatial ability can be correlated with their performance in tasks involving navigation through a user interface. We were interested to know if differences in spatial ability would affect people's usage of similarity-based visualisations of image sets, and therefore asked our participants to take Set I (12 questions) of Raven's Advanced Progressive Matrices (APM) [96], a culture-free test of spatial reasoning that specifically evaluates the ability to think about abstract categories.

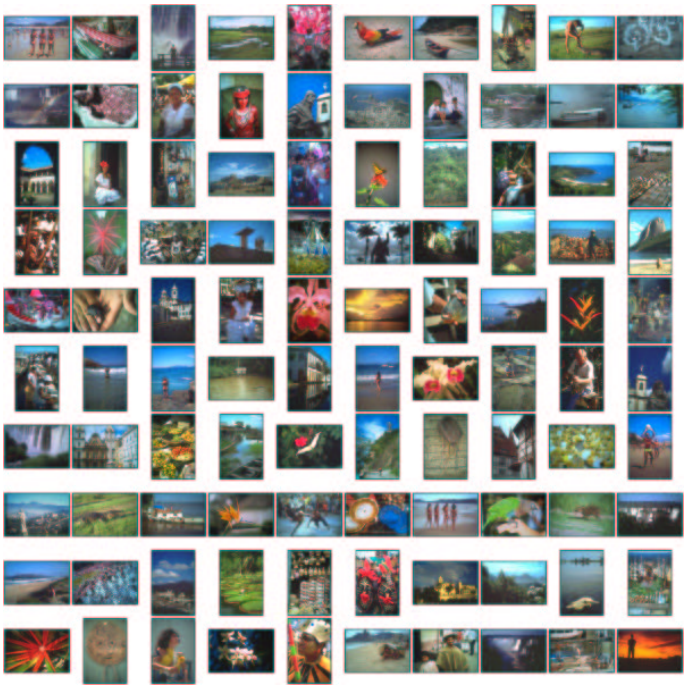
5.2.1 Participants

Our participants were ten students of graphic design from Anglia Polytechnic University, Cambridge. Eight were at the end of the second year (of three) and two were at the end of the first year. Again, all had normal or corrected-to-normal vision, with no colour blindness (self-reported). They volunteered to participate via their lecturer, and we did not know in advance how many would arrive. The experiment was conducted in two sessions, each with five students at once. They were paid five pounds for their participation.

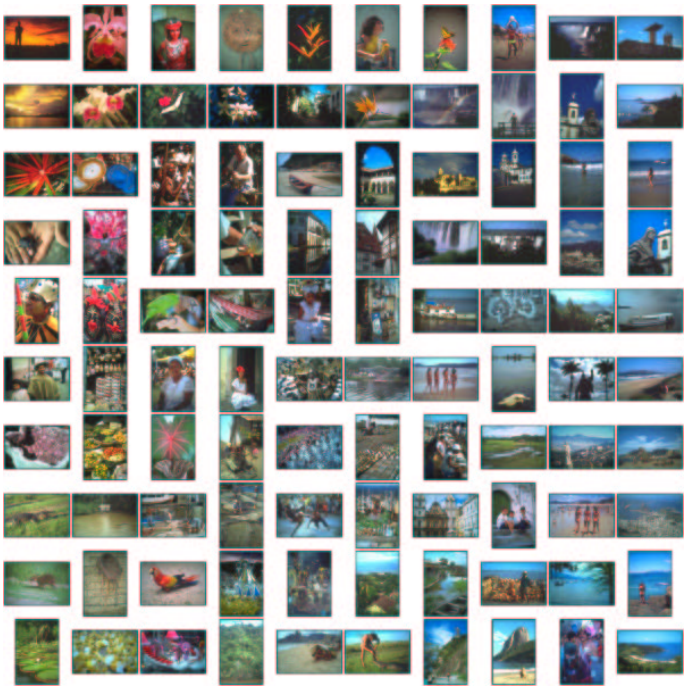
We carried out a pilot study, identical to the main experiment, with one participant: a philosophy student at the University of Cambridge, who had previously worked temporarily at a graphic design agency. Her results are not included in the quantitative analysis, although as she took more time over her questionnaire than the participants in the main part of the experiment, some of her comments have been quoted below (marked with a **P**). No changes were made to the design as a result of the pilot study, but the post-experiment questionnaire was altered slightly.

5.2.2 Apparatus

We selected nine places for the main part of the experiment: Brazil (93000), Canada (230000), Death Valley (39000), Denmark (121000), Ireland (385000), Jamaica (328000), Kenya (253000), Nepal (187000), and Yellowstone National Park (94000). A tenth, Devon and Cornwall (291000), was chosen for use as a practice. For each of these, we created a random arrangement and an arrangement based on visual similarity (for example, Figure 5.3). In both cases the



(a) Random.



(b) MDS proximity grid, based on visual similarity.

Figure 5.3: Two 10×10 grid arrangements of 100 images of Brazil, as used in the Anglia experiment.

100 images were placed into 10×10 proximity grids[5], rather than the 12×12 of the previous two experiments. We did this because we had to use a lower screen resolution (1024×768) and wanted to maximise thumbnail size; these were 74×49 pixels, or 0.46% of the screen area, and the magnified images were 222×148, 4.2% of the screen area. We used five Windows 98 PCs, all of which had 17-inch monitors and 64MB of memory.

Again, we chose places where the Corel images were of high technical quality, and we abridged the pieces of text from the online *Rough Guide* to each place, to a length of approximately 150–200 words. In the infodesign 99 study, relevance to the text was important to all of the participants when making their selections, and some complained that the available photographs were not always a good match. For example, the text might mention a particular landmark, causing the participant to look for a photograph of it, which could be frustrating if no such photograph was present in the given set. We were keen to avoid this in the Anglia experiment, especially as it might have affected the participants' opinions of the different arrangements. One possibility was to get rid of the texts altogether, and simply ask the participants to choose photographs based on their own impressions of the place. However, people tend to know a lot more about some places than others, and the pieces of text provided the participants with some useful background information. The lecturer we contacted also felt that removing the texts would make the task less realistic. Our eventual solution was to edit the texts very carefully, so that where possible, we chose passages that focused on creating an impression of the place, rather than describing particular landmarks, in order to encourage participants to decide for themselves what they should look for. The texts are given in Appendix D, starting on page 214.

Some changes were made to the experiment instructions used in the infodesign 99 study. Firstly, we explicitly mentioned that the texts were meant to provide background information on each place, as well as giving the participants an impression of what the finished Web page article would convey. Instead of stating that the photographs were to "illustrate" the text, we said that they would "appear on" a Web page with the text. We also mentioned that creativity was encouraged, and that the images should work well together as a group. The final version of the instructions is shown in Figure D.9 in Appendix D.

The experiment software (Figure 5.4) was the same as in the infodesign 99 study, but with two modifications. Firstly, the radio buttons for switching between arrangements were labelled "Library" and "Visual". Secondly, following comments from the infodesign 99 participants, there was space to select up to nine images at a time, instead of only four. The experiment software and data were recorded onto CD-ROMs, and the experiment was run directly from these, with log file data being saved to floppy disc.

[5] using the method based on a genetic algorithm; see Appendix A.

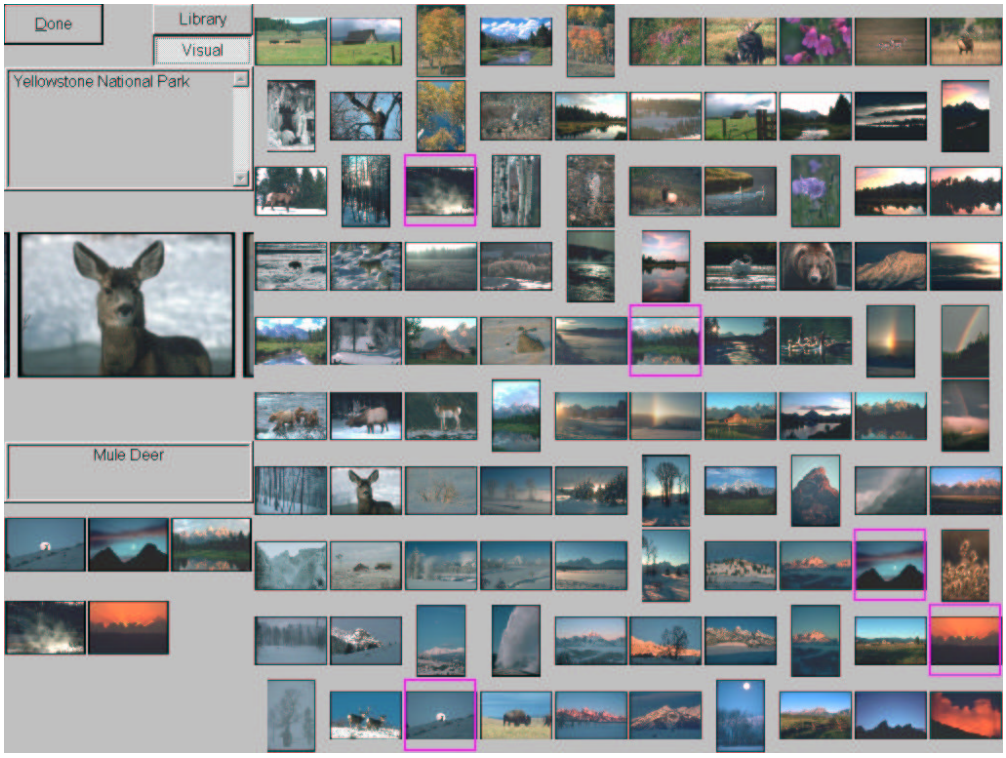


Figure 5.4: The software for the Anglia experiment. Nine images can be selected at a time, compared to four in the version shown in Figure 5.2.

Participant	*Part one*		*Part two*
1	V,X	R,Y	B,Z
2	R,X	V,Y	B,Z
3	V,Y	R,X	B,Z
4	R,Y	V,X	B,Z
⋮	⋮	⋮	⋮

Table 5.4: Anglia experiment design. The first letter in each column is the type of arrangement: either **V**isual, **R**andom, or **B**oth. The second letter is the group of places used: X, Y, or Z.

5.2.3 Design

In the infodesign 99 experiment, participants always had both types of arrangement available. This time, we wanted them to use each one on its own, so that they could appreciate its strengths and weaknesses fully, before going on to having a choice between the two. We also wanted to compare the arrangements directly, in terms of how quickly selections were made. The experiment was therefore divided into two parts. In part one, the students chose photographs for six places, using one type of arrangement for the first three, and the other for the second three (a within-subjects design). Half of the participants used the visual arrangement first and half used the random arrangement first. Then, in part two, both types of arrangement were available; the students had to choose one of them to use initially, and could then switch between the two views as they wished.

The experiment design is illustrated in Table 5.4. For part one, we chose six of the places and divided them into two groups. Group X was Denmark, Jamaica, and Nepal, and group Y was Death Valley, Ireland, and Kenya. We then counterbalanced both the type of arrangement and the group of places, so that the first type of arrangement used (**V**isual or **R**andom) alternated with every participant, and first group of places (X or Y) alternated with every second participant. We also systematically varied the order of places within a group. The remaining three places (Brazil, Canada, and Yellowstone National Park) formed group Z, and this was used for every participant in part two, with **B**oth arrangements available.

This design allowed us to compare the visual and random arrangements to each other, using the data from part one, but we could not directly compare using a single arrangement to using both of them together, because different groups of places were used in parts one and two.

The APM test was administered between part one and part two. This gave the students a break from the task, and was also intended to reduce the chance (in part two) of them simply favouring the arrangement type they used most recently.

5.2.4 Procedure

On arrival, each participant sat down at a PC, and was given a number of sheets of paper: the experiment instructions (including a description of the task and a step-by-step list of what they should do; see Figure D.9 in Appendix D), a personal information questionnaire (name, contact details, and so on; see Figure D.10 in Appendix D), and the accompanying texts for each place (one place per page, in alphabetical order). They read the instructions and then filled in the questionnaire.

Then, they started up the experiment software for the practice search, "Devon and Cornwall". In this case both of the arrangement types were available, and we explained them to all five participants in the group, along with the software and its features. They were then asked to practise using it until they were comfortable, and had made three selections. At this point they could press the "Done" button to close the practice and start up part one of the experiment, which they did by double-clicking on a script file corresponding to their participant ID.

Ideally, we would have introduced each arrangement type just before the participant was about to use it, rather than explaining both at the same time. However, as there were multiple participants, working simultaneously but in their own time, this would have proved awkward, unless all of them were using the same arrangement type at once. This was not possible, because we did not know the number of participants in advance, and thus had designed the experiment to balance the order of arrangement types with every other participant.

To reiterate, in part one each participant did six searches: three with one arrangement type (on its own) and then three with the other. They did this in their own time, with no limit, but were encouraged to be quick. Then they were given the APM test, and were asked to complete that in their own time. In part two they had both arrangement types available and could switch between them.

The procedure for a single search was largely the same as in the infodesign 99 experiment, except that in part one, the participant did not have to choose which arrangement to use, because only one type was available. The name of the place was displayed, and the participant read its associated text. Then, the radio button(s) for the available arrangement(s) became enabled once the image data was loaded, and clicking on one of these buttons caused the corresponding arrangement to be displayed, and started the timer for the search. Again, once the participant had a set of three images she was happy with, she pressed the "Done" button, and moved on to the next place name. Participants' actions were logged and timed by the experiment software.

At the end they were given a three-page post-experiment questionnaire to fill in (Figure D.11 in Appendix D). They rated their agreement or disagreement with a series of statements on a seven-point scale, where 0 was "strongly disagree" and 6 was "strongly agree". There was also space for them to make comments.

5.2.5 Results and discussion

In this section, we analyse the data logged by the experiment software, and summarise the participants' questionnaire responses, quoting their comments where appropriate. As with the infodesign 99 study, our results are split into sections considering the usefulness of the different arrangement types, the criteria participants were using when selecting photographs, and the user interface of the experiment software.

Usefulness of different arrangement types

In part one of the experiment, the visual and random arrangements were directly compared to each other. Our response variables in these searches were:

- done, the time taken before the "Done" button was pressed
- length, the length (in grid cells) of the participant's mouse trail across the arrangement

Figures 5.5 and 5.6 show the distributions of these variables for the two arrangement types. Although length was approximately normally distributed, done was not, so a log transform was applied to the latter before analysis. The two variables were significantly correlated ($r = 0.784$, $p < 0.001$).

We again used S-Plus's linear regression features, and constructed linear models for both log(done) and length. The following predictor variables were used in each model:

- subject, the ID of the participant (10 levels)
- trial, the sequence number of the search, from 1 to 6 (an ordered factor)
- place, the ID of the place (6 levels)
- type, the type of arrangement used (visual or random)

Then, the ANOVA function was applied to the fitted models, as shown in Table 5.5. For both done and length, type is significant at a level of $p < 0.05$. Participants were *slower* with the visual arrangement, as can be seen in the boxplots and in Table 5.6. Also, trial is significant, at a level of $p < 0.01$ for length and $p < 0.001$ for done: in general, participants took less time over their selections as they went along. This could be because they became more practised, or started taking the task less seriously, or some combination of these. But, because of the balanced design, this does not affect the significance of type.

We also constructed linear models for the (transformed) variables first, last, and pause, as defined in the previous chapter: the time of the first selection, the time of the last selection (or deselection) before "Done" was pressed, and the time between last and done. There was a significant correlation between last and done ($r = 0.996, p < 0.001$), and so the ANOVA results for last were very similar to those on the left of Table 5.5. None of the predictor variables were significant (at a level of $p < 0.05$) for either first or pause.

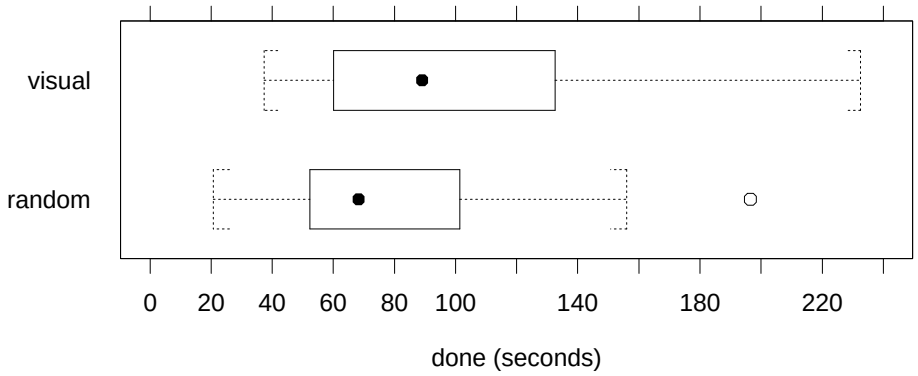


Figure 5.5: Distribution of done for the two arrangement types.

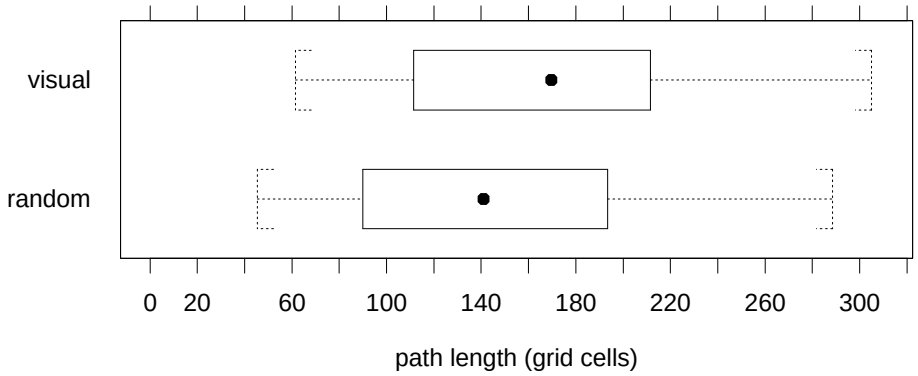


Figure 5.6: Distribution of length for the two arrangement types.

		done		length	
	Df	F	p	F	p
subject	9	7.19	< 0.001	7.42	< 0.001
trial	5	5.31	< 0.001	5.10	< 0.010
place	5	1.31	0.278	2.43	0.052
type	1	6.34	< 0.050	5.18	< 0.050
Residuals	39				
multiple r^2		72.88%		72.54%	

Table 5.5: ANOVAs for log(done) and length.

Variable	*Visual*		*Random*	
	mean	(s.d.)	mean	(s.d.)
done (seconds)	103.78	(56.02)	81.31	(41.39)
length (cells)	169.55	(67.42)	144.32	(64.43)

Table 5.6: The means and standard deviations of done and length for the two arrangement types.

Feature	*Type*	*Agreement* 0 1 2 3 4 5 6	*Mean*	*Med*	*p*
"It was enjoyable to use"	Visual	0 0 0 2 4 4 0	4.2	4	
	Random	0 1 4 1 2 2 0	3.0	2.5	0.019
"I thought it was useful"	Visual	0 0 0 3 2 4 1	4.3	4.5	
	Random	0 2 2 1 4 1 0	3.0	3.5	0.020
"It made it easy for me to find the photos I wanted"	Visual	0 0 0 0 4 2 4	5.0	5	
	Random	0 4 0 2 3 1 0	2.7	3	0.007
"It made it easy to find photos that complemented each other"	Visual	0 0 1 1 3 2 3	4.5	4.5	
	Random	0 3 1 3 0 3 0	2.9	3	0.133

Table 5.7: Comparing participants' ratings of the two arrangements in the Anglia experiment. 0 represents "strongly disagree" and 6 represents "strongly agree".

Table 5.7 summarises the participants' agreement scores for four statements regarding the random and visual arrangements, and shows the results of comparing the pairs of scores using two-tailed Wilcoxon signed-rank tests. For the first three statements, the scores for visual were significantly higher, $p < 0.05$. This is not the case for the fourth statement, despite the difference in mean scores, because three of the participants gave the random arrangement a better rating than the visual arrangement.

As we discussed in Section 2.5.3, other research has established that users' requirements develop during the picture selection process, so for example an initial search for images of New York may lead to sub-requirements for the Statue of Liberty, a taxi cab, a night scene, and so on. As we have seen, images of the same generic type can often be grouped together in an arrangement based on visual similarity. Participants' comments suggest that this is mostly why they found it useful:

02 *If you had an idea in mind, say a sunset, you have them all together, easier to make the choice. [...] Already knowing the sort of pictures I wanted, the visual arrangement suited that.*

03 *Visual puts photos in groups, i.e. it is easier to find one type of photo. You have direct comparisons to make a choice of photo.*

05 *Good once you know what you want.*

09 *Good for finding genres, e.g. landscape, night-time, daytime etc.*

However, as we noted in Section 4.1.5, the grouping together of visually similar images lowers the local contrast, making neighbouring images less distinct from each other than in a random arrangement. The following were given as disadvantages of the visual arrangement:

05 *Makes choosing out of certain group more hard.*

09 *Easier to miss a photo completely.*

Arrangement type	*Satisfaction*							*Mean*	*Med*
	0	1	2	3	4	5	6		
Visual	0	0	3	7	5	11	4	4.2	4.5
Random	0	0	4	8	13	5	0	3.6	4

Table 5.8: Comparing the visual and random arrangements in terms of the satisfaction ratings participants gave for each search.

10 *My eye was often drawn to one set of colourfully interesting images while ignoring the rest.*

This may help to explain why the participants were faster when using the random arrangement in part one: often the participants did not really know what they were looking for, especially at the beginning of a search, and simply wanted to browse through the set, seeing what was there. In this case the random arrangement could be useful, as the local contrast enabled individually strong images to stand out, rather than be lost among their visually similar neighbours (Figures 5.3(a) and 5.3(b) may be helpful in understanding this effect). Participants listed the following as advantages of the random arrangement:

02 *You saw the whole selection, but mixed up so the ones you liked sprang out at you.*

05 *Mixed up — so didn't structure my choosing.*

06 *Gave good overall sense of what the choice was.*

But of course, the random arrangement also has disadvantages: it "meant visually trawling through images that were in no order" (10), making it "hard to quickly find a genre" (09).

In summary, when the participants were using the random arrangement in part one, they could simply grab three eye-catching photographs, but because single images stood out less in the visual arrangement, they were perhaps forced into thinking harder about what they were looking for, thus taking longer to complete their search. In other words, a visual arrangement seems to support directed browsing, while a random arrangement can be helpful for undirected browsing.

We could not use measures like recall and precision to assess the effectiveness of the participants' selections, because all of the images in each set were relevant to the given topic, and the participants were applying their own selection criteria. In the post-experiment questionnaire, they were simply asked to rate their satisfaction with their selections for each search (from 0 to 6, where 0 is "not at all satisfied" and 6 is "very satisfied"). The overall satisfaction scores for the two arrangements are shown in Table 5.8: the mean and median scores were higher for the visual arrangement. A simple comparison of the individual scores for the visual arrangement and the random arrangement, using a two-tailed Wilcoxon rank-sum test, gives a p value of 0.054, just missing significance at the $p < 0.05$ level. Looking at the totals for each participant (not

shown in the table), six of them were more satisfied with the searches they did using the visual arrangement, one was more satisfied with those using the random arrangement, and the remaining three had the same total rating for both types. It is possible that because the participants generally took longer to make their choices when using the visual arrangement, they felt that they had looked harder, and were therefore more satisfied, but we could find no relationship between satisfaction score and either done or length.

In part two of the experiment, the participants had both arrangement types available, and could switch between them. We had expected that they would overwhelmingly favour the visual arrangement. It was chosen first in 21 of the 30 searches, and on average was used for 66% of the search time (so random was used for 34%). Five participants heavily favoured the visual arrangement, using it for 73% or more of the time; one used it exclusively. Two heavily favoured random, using visual for 22% of the time or less, and the other three favoured neither, using visual between 56% and 66% of the time. Three participants made all of their selections (and deselections) using the visual arrangement, and two made all of them using the random arrangement, with the others using a combination of both. On average, the participants made 63% of their selections using the visual arrangement.

Despite generally rating the visual arrangement more highly, many of the participants seemed to feel that it was worth looking at both arrangements: in total, they switched between the arrangements in 18 of the 30 searches, and the most switches in any one search was 3. Only one participant (05) seemed to be actively using the two arrangements together, deselecting images previously selected in the other arrangement. Five of the ten participants (numbers 01, 02, 05, 09, and 10) said that they preferred having both arrangements available, especially to see if an image stood out in one arrangement that they hadn't spotted in the other (note that they refer to the random arrangement as "library"):

05 *The more variety the better. I preferred using each arrangement as a layout to 'scan' to see if the same images stuck out to me.*

09 *Can track back to make sure I have seen all the photos. More flexible. I preferred visual, though it is sometimes a bit awry. Library was too disjointed to quickly browse.*

10 *Both had merit. Useful with visual if I wanted to make sure [my selected] images were not too similar. Good to start off with library for general "first impression" of images, and then switch to visual.*

This also has its disadvantages, however: it may be "too complicated overall" (06), and may cause the user to "lose track of where things are" (P). Four participants (03, 04, 07, and 08) and the pilot study participant said they preferred using visual on its own. One participant (06) preferred the random arrangement, but said that "visual was good for a quick browse". This range of preferences is also reflected in participants' responses to the statement "I liked having two different arrangements of the same set of images" (Table 5.9).

Statement	*Agreement*							*Mean*	*Med*
	0	1	2	3	4	5	6		
"I liked having two different arrangements of the same set of photos"	0	1	3	2	1	3	0	3.2	3
"It made no difference to me how the photos were arranged"	2	1	0	3	1	2	1	3.0	3

Table 5.9: Further questionnaire responses relating to the usefulness of the different arrangement types.

Statement	*Importance*							*Mean*	*Med*
	0	1	2	3	4	5	6		
Technically good, striking photographs	0	0	0	1	3	2	4	4.9	5
Photographs that worked well as a set of three	1	0	0	0	0	6	3	4.8	5
Photographs that were relevant to the given text	0	0	0	2	1	4	3	4.8	5
Photographs that fitted your own impressions of the place	1	0	1	3	1	2	2	3.7	3.5

Table 5.10: The importance of different selection criteria to the participants in the Anglia experiment.

Four participants agreed with the statement "it made no difference to me how the photos were arranged" (Table 5.9); for three of them, this seemed to conflict with their other questionnaire responses, suggesting that although they rated the visual arrangement more highly, they were happy to use either.

Finally, in part two, the mean value of done was 85.06 seconds, and the mean of length was 149.46 cells; the mean satisfaction score was 4.1, with a median of 4. These results are not directly comparable with those from part one, because a different set of places was used.

Selection criteria

Table 5.10 shows participants' ratings of the importance of different criteria when they were making their selections. In the infodesign 99 study, the most important factor was the relevance of the selected photographs to the given text, then their technical quality, and then their coherence as a group. This time, these three criteria were roughly equivalent in importance, which may be because of the changes we made to the experiment instructions, or differences between information designers and graphic designers.

When participants gave their reasons for being particularly satisfied or dissatisfied with their selections, they often mentioned that they were satisfied when the images complemented each other, for example that they "fitted well together compositionally" (03) or made "a good trio — mainly because of colour" (05). As their choice was limited to the initial set of 100 images, many of the comments with regard to satisfaction referred to this set, rather than the three they had selected. "Diversity" (04) was important, with participants preferring to choose from a "wide ranging series of images" (10) to end up with "a varied selection, e.g. nightlife, wildlife, day life" (02).

When asked if they had been using any other criteria, some participants mentioned that they wanted to have a set which was representative of the place, giving a "good sense of nativeness" (06) and showing "a variety of attractions as well as the country's real culture" (09) while containing "not too

many cliché pictures" (06). Figure D.12 in Appendix D shows the most commonly selected images of each place.

02 *My ideas came from what was mentioned in the text, picking out the major points and landmarks.*

03 *Photos had to be clear, crisp, and compositionally correct, and all had to be different images but must fit together.*

04 *After reading the article, I knew what I wanted — if it said high peaked mountains I'd have them in mind before I searched.*

10 *After building a mental image of the place and then looking at the photos my ideas were governed by the images. [...] I tried to look for photos that shared the diversity of the region.*

P *I looked for images that complemented one another as a set (colour/composition) and showed a range of the subjects available. I tried to vary the subject matter so the finished set had a good range of interest.*

Participant 05 and the pilot participant were able to describe their selection processes in more detail. They adopted different strategies: 05 would cut down the initial set of 100 into a candidate subset, and choose the final three from among those, but P usually chose one image and concentrated on finding two other photographs to go with it.

05 *1. General scanning choosing lots I liked in both library and visual. 2. Cancelling out some on close inspection. 3. Choosing 3 different ones to broaden view and colours etc.*

P *I tended to select one initially striking image (working on colour and composition — something dynamic but usually simple) and then work around it. I often tried the set without it but ended up keeping it.*

In the previous chapter, we considered the effect of salience on how quickly a single target image was found within a set. For this experiment, we decided to test whether salient images were more likely to be selected by the participants, again by defining an image's salience as its average dissimilarity to all of the other images in its set. In total, 197 unique images were selected during the experiment, and 703 were never selected. A comparison of these two sets with a Wilcoxon rank-sum test shows that the selected images had a significantly higher level of salience ($p < 0.05$), although the difference in means is small. The difference is present regardless of the arrangement type used to make the selections.

The only questionnaire comment relating to salience came from the pilot participant, who suggested that perhaps the use of thumbnails encouraged the selection of distinctive images, especially close-ups.

P *The thumbnails were quite small — I think this made me more likely to choose clearer, more graphic pictures — the detail shots were a bit lost.*

Statement	*Agreement*							*Mean*	*Med*
	0	1	2	3	4	5	6		
"I have prior experience of carrying out picture/photo selection"	1	0	2	0	3	1	3	3.9	4
"I thought the task was realistic"	0	0	0	0	3	5	2	4.9	5
"It was easy to find suitable photos from those available"	0	0	1	0	7	2	0	4.0	4
"I browsed the photos with 'mental images' of what I wanted"	0	0	1	2	4	3	0	3.9	4

Table 5.11: Other questionnaire responses from the Anglia experiment.

All but one of the participants said that they had found it easy to locate suitable photographs (Table 5.11), and seven agreed that they had browsed the photographs with mental images of what they were looking for, echoing our previous results.

The user interface of the experiment software

Participants were asked if there were any other features they would have liked the experiment software to have. As in the infodesign 99 study, most of them seemed generally happy with it.

09 *I found it pretty complete, easy and natural to use. I like being able to view photos without clicking on them.*

Two participants (plus the pilot participant) said that it would be useful to them to be able to see selected images at a larger size.

05 *Perhaps to view each single picture on its own with black background (and maybe choice of three too?).*

P *I would have liked to see the selected thumbnails bigger — and been able to move them around to compose the set.*

Two participants mentioned that they would have liked the set of photographs to be labelled according to a "theme", with one imagining this could be done with the existing visual arrangement, and the other suggesting that he would have liked a different, more semantic type of arrangement.

04 *Perhaps in the case of visual, a little bit of text stating the theme i.e. "plants" or "people".*

10 *Perhaps a way of sorting the images by general theme rather than just colour similarities, e.g. landscapes, people, towns etc.*

Other details

The raw scores on the APM test were between 7 and 11 (out of 12), which is from "low average" to "high" for the 18–39 age range. We thought that perhaps participants with higher test scores would have a greater preference for the visual arrangement, because they should have found it easier to understand how the two-dimensional placings of the images reflected their relative

similarities. However, we found no relationship between test score and any of the experimental variables or questionnaire responses. Given our other results, it appears that a participant's spatial ability was less important than the amount of directed browsing she was prepared to do: perhaps those participants who found most benefit in a similarity-based arrangement were those who were more likely to attempt to narrow their search in a directed manner, rather than being content to take an undirected approach and select whichever images popped out.

Table 5.11 shows that, as in the infodesign 99 experiment, not all of the participants had prior experience of carrying out picture selection. This time, however, there was no relationship between experience and performance in the experiment. All of the participants agreed that they thought the task was realistic, confirming our belief that it was not necessary to provide them with a sample page layout.

5.3 The Anglia follow-up

After analysing the results of the Anglia experiment, we contacted the students again (about a month later), and asked if they were willing to be involved in a follow-up study. This was at the beginning of a summer break, and so only two of the original participants were available. They were asked to carry out a number of searches while thinking aloud, and we recorded them on video. They used visual and random arrangements, as before, and were also exposed to a number of variations, including caption-based arrangements, as used in the infodesign 99 study.

Our primary aim was to gain more of a qualitative insight into the differences between arrangement types than is possible with post-experiment written comments. We also hoped to find out in more detail about how an initial, general requirement becomes specialised into different sub-requirements as searching proceeds. Finally, we wanted to quickly test the potential usefulness of variations on the basic arrangement types, without having to carry out further formal experiments.

5.3.1 Participants and apparatus

Our participants were numbers 04 and 05 from the original Anglia experiment. We used a PC running Windows NT 4, with 192MB of memory, and a 17-inch monitor set at a resolution of 1024×768. The same experiment software was used, with some slight modifications to accommodate variations in arrangement type, as described in the next section. Table 5.12 lists the image sets and arrangement types that were used. The proximity grid arrangements were created using the genetic algorithm method, and the continuous arrangements were produced with the Newton-Raphson method (see Appendix A).

Corel category ID	*Place*	#	*Random*	*Visual*	*Caption*	*Text?*	*Select?*
253000	Kenya	100	10×10	10×10	–	yes	yes
121000	Denmark	100	10×10	10×10	–	yes	yes
128000	Russia, Georgia & Armenia	100	10×10	10×10	–	no	yes
65100	Japan	100	10×10	10×10	–	no	yes
118000	Greek Islands	100	–	–	10×10	yes	yes
82000	San Francisco	100	–	–	10×10	yes	yes
244000	New York	100	–	10×10	10×10	yes	yes
223000	Paris	100	–	10×10	10×10	yes	yes
244000	New York (plus labels)	100	–	–	10×10	yes	no
93000	Brazil	100	–	10×10	–	yes	no
93000	Brazil	100	–	12×12	–	yes	no
93000	Brazil	100	–	cont	–	yes	no
230000	Canada	100	–	10×10	–	yes	no
230000	Canada	100	–	12×12	–	yes	no
230000	Canada	100	–	cont	–	yes	no
88000	Czech Republic	100	–	12×12	12×12	yes	yes
166000	Turkey	100	–	12×12	12×12	yes	yes
102000 & 215000	France	200	–	15×15	15×15	yes	yes
53000 & 64000	Mexico	200	–	15×15	15×15	yes	yes

Table 5.12: The image sets and arrangement types used in the follow-up study, in the order that they were presented to the participants. The number of images shown (#) was 100 for all but the last two places, France and Mexico, where two categories were combined to create a set of 200 images. The *Random*, *Visual*, and *Caption* columns show which arrangement types were available for each place; normally proximity grid arrangements were used (and the grid size is given here), but "cont" indicates that a continuous arrangement was used. The *Text?* column shows whether background information about the place was provided to the participants, and the *Select?* column indicates whether participants were required to select images from the set, or simply comment on the arrangement.

Figure 5.7: A mock-up showing a caption-based arrangement of 100 images of New York, with superimposed labels, manually extracted from the captions.

5.3.2 Design and procedure

The first two searches (Kenya and Denmark) were repeated from the original experiment, to allow the participants to practice thinking aloud, and get used to the experiment software again. Both visual and random arrangements were available, in 10×10 proximity grids. Again, the thumbnails were 74×49 pixels each (0.46% of the screen area), and could be viewed at $3\times$ zoom. The next two searches were of two new places (Russia and Japan); these were the only searches which had no accompanying text, as we were interested to see if this would make a difference.

Then, we introduced the caption-based arrangement, and the participants did two searches (Greek Islands and San Francisco) using it, on its own. An extra radio button, marked "Caption", was placed alongside the "Visual" and "Library" buttons, to allow the caption-based arrangement to be selected. As before, if an arrangement type was unavailable, its button was greyed out. In the next two searches (New York and Paris) both a caption-based arrangement and a visual arrangement were available, and the participants could switch between the two, as in the infodesign 99 study.

Following that, there was a break from searching, and the participants were shown a mock-up of another variation: the caption-based arrangement of New York that they had already seen, but this time with short textual labels superimposed on it (see Figure 5.7), with the aim of making the structure of the arrangement more immediately obvious. Lin's self-organising map displays of document collections [71], as discussed in Section 2.2, have similar labels, and his techniques could be applied here. Chalmers, Ingram, and Pfranger [21] have suggested a simple method of identifying clusters within a visualisation,

and labels could then be created automatically by picking out commonly occurring words from each cluster. For the mock-up, however, we simply did this manually[6].

Next, we introduced more sparse arrangements, using two places from the original Anglia experiment (Brazil and Canada). We showed the participants both of these image sets (arranged by visual similarity) in a 10×10 proximity grid, a 12×12 proximity grid, and in a continuous MDS arrangement. This is like the comparison made in Figure 4.8 on page 75, except that the participants did not see all three versions simultaneously. We explained to them that, compared to the 10×10 grids they were used to, the 12×12 and continuous versions showed progressively more accurate representations of the original similarity of the images, but that as a result the image thumbnails had to be smaller, or that some overlap was introduced. The participants were not required to make selections from these arrangements. In both the 12×12 and continuous versions, the thumbnails were 60×40 pixels, 0.31% of the screen area.

Then the participants returned to the task of selecting images to accompany text. For the next two places (Czech Republic and Turkey), they again had a caption-based arrangement and a visual arrangement available, this time in 12×12 proximity grids.

For the final two places (France and Mexico), the participants had to select from 200 rather than 100 images. Both caption-based and visual arrangements were available, in 15×15 proximity grids. The thumbnails were 48×32 pixels (0.20% of the screen area), and could be viewed at $4\times$ zoom, instead of 3. We added a simple text-based query facility to the interface: pressing a "Search" button launched a dialogue box where the participants could enter some text, causing any photographs whose captions contained that text to be highlighted in yellow (the magenta highlights of any selected images remained visible). We explained and demonstrated this to the participants. A "Clear" button could be used to remove the highlights from the most recent query.

5.3.3 Results and discussion

We recorded the sessions on video, and subsequently transcribed the participants' comments. After reading the transcripts, we defined an initial **framework** [98] containing a number of categories, and used this to classify the comments; each comment could be placed in more than one category. The process took several iterations, which involved refining the framework. The top-level headings are shown in Table 5.13, and these are broken down further in the following sections. It is easy to see, for example, that more than half of the participants' comments were about their selection criteria.

Heading	*04*	*05*
Usefulness of different arrangement types	83	81
Browsing strategies	18	36
Selection or rejection criteria	156	155
The user interface of the experiment software	10	36
Total	267	308

Table 5.13: The top-level headings of the framework, showing the total number of comments that the participants made about each topic.

Category	*04*	*05*
Visual arrangements being useful for directed browsing	16	3
Visual arrangements having obvious structure	14	7
Visual arrangements causing similar images to appear to merge	1	9
Random arrangements being useful for undirected browsing	6	5
Caption-based arrangements being useful for undirected browsing	4	6
Caption-based arrangements being useful for directed browsing	3	4
Caption-based arrangements *having* obvious structure	5	3
Caption-based arrangements *lacking* obvious structure	2	8
Comments about the usefulness of superimposed labels	3	5
Caption-based arrangements would be more useful if you knew the place	2	1
Comments showing initial disorientation at more sparse arrangements	6	1
Comments mentioning 12×12 proximity grids of 100 images	7	14
Comments mentioning continuous MDS arrangements of 100 images	6	11
Comments mentioning 15×15 proximity grids of 200 images	8	4
Total	83	81

Table 5.14: The part of the framework relating to the usefulness of the different arrangement types, showing the number of comments classified under each heading.

Usefulness of different arrangements

Table 5.14 shows that the participants found both visual and random arrangements to be useful, depending on their current requirement (echoing the results of the original experiment). When they were looking for something in particular (directed browsing), the structure provided by a visual arrangement could help them to narrow down the set, especially when the requirement was for image content at the primitive or generic levels.

04 *That's now got me thinking of a temple shot. Which I'll click on to the visual for, see if there's any temples. Which there is, there we are, found that straight away.*

05 *I prefer a darker one or a different one, which I'm going to look for now on this side of the screen, which is easier, so I much prefer the visual for this.*

05 *I'm definitely seeing a diagonal line here between lots of skyscraper, sky shots, and city shots, at night.*

The grouping of visually similar images also resulted in lower local contrast, however, giving the images less chance of standing out individually.

05 *They're all merging together, my eyes aren't concentrating enough on each one, I think they're too similar.*

05 *I've found loads of weird ones that I didn't think I'd... I would never have picked them out in a million years with visual. I'll just see where that one is [in visual]. It's there... I didn't even cover that area particularly. The whole thing... too similar, I suppose.*

When looking without a particular requirement in mind (undirected browsing), a random arrangement (referred to as "library" by the participants) was helpful, because of its higher local contrast.

04 *I'm going to click back on to the library again, because I just feel more comfortable with that in the first instance, just for skimming.*

04 *I'm staying off the visual just yet [and using library] because I don't really know what I'm looking for. Just looking to see what attracts me.*

05 *[I switched to library] so I could look all over it and just see if something actually caught my eye.*

We used caption-based arrangements in the follow-up study, but not in the original experiment. Like random arrangements, these tend to have good local contrast, because images that have similar captions are usually not visually similar. This was also noted by two of the participants in the infodesign 99 study (Section 5.1.4).

[6]We did consider doing it automatically, but the lack of detail in the Corel captions again caused problems here.

05 *I like the setout… I like library anyway, so it's similar […] because it tends to make different ones stick out a bit more.*

05 *I can't decide, so I'll have a quick look at the caption. Yeah, that's helped, […] makes me want to look round randomly like that, rather than structured. But I quite like that.*

In the infodesign 99 study we found that the usefulness of these arrangements is affected by the level of detail present in the captions. Because the Corel captions tend to focus on names (the specific level), the resulting arrangements usually grouped together photographs of a particular named area or landmark, for example in the Czech Republic set:

04 *Now that I know this is the Prague area I'm sort of staying here, because I think it's got quite a lot more to offer, visually.*

The results for the generic level were more mixed, because this is not always covered by the captions. The pictures of "women in dresses" described below were in fact clustered together because their captions all contained the word "Moravia" (a region of the Czech Republic), and the photographs all happened to show people in traditional costume.

05 *I've just immediately looked at these ones and I've decided I really don't want women in dresses. So I can immediately just bar that whole section, which is really helping me.*

However, as suggested by the earlier quote about the Prague images, the main problem that the participants had with the caption-based arrangement was finding the area that contained images matching their requirement. This is because the basis of its structure is usually not immediately obvious, unlike a visual arrangement. For example, if the arrangement in Figure 5.1(a) on page 95 did not have the extra labels, the images of Times Square and Manhattan Bridge would be difficult to locate quickly, especially for someone who is unfamiliar with New York.

05 *I'm trying to work out how the captions go together. […] Is it in alphabetical order?*

This was the motivation behind creating the mock-up with superimposed keywords (Figure 5.7), which elicited a number of positive comments. We believe that if such labels could be reliably created automatically, they would be of great help to users in locating images matching a requirement, and also in getting an overview of the range of photographs present.

04 *Yep, that helps, […] all the areas, […] I know now […]. Without those it would just be a matter of guesswork, so, they are helpful, especially for me who likes to take perhaps three photos from different areas. So yeah, […] that's far more helpful than just simply the caption at the side, because […] from the images alone I couldn't really tell where [they were taken] unless I have a good knowledge of the city. Or country.*

Both participants stated that the caption-based arrangements would be more useful if they were more familiar with the place.

04 *The thing I think I'd find beneficial with the caption, at the moment, is if I had more of a knowledge of the city, or country. As it is, I don't know the areas, I'm only going on the images that I recognise.*

05 *If I knew more about these places I'd probably have more particular places that I wanted, in which case caption would definitely… [be useful].*

The participants were shown two variations on the 10×10 proximity grid they were used to: a 12×12 proximity grid arrangement, and a continuous MDS arrangement. The 12×12 grid was perceived as being more useful than the 10×10, as the extra space caused clusters to be more obvious.

04 *It was a little bit disconcerting when you first did it because I'm just so used to the* 10×10*. It looked more messed up, actually. But now that I'm looking at it, yeah, I can see the connections going on, and probably if I'd started with that it would make more sense, actually, yeah.*

The added structure seemed to result in a change in navigation style, jumping between clusters instead of scanning the arrangement row-by-row.

04 *I think you'd probably bounce from group to group, […] because they're in shapes, […] the mind doesn't read it like a page, any more.*

05 *I want to look at the ones on the outside [of the visual arrangement], because they're different.*

However, with a more sparse grid, the image thumbnails have to be smaller. Initially, neither participant minded this, saying that there wasn't much difference, and that they spent a lot of time looking at the magnified version anyway. However, after actually using a 12×12 grid, participant 05 said that the smaller thumbnail size was "more of a problem than I thought it would be."

Neither of the participants liked the continuous arrangement; they both felt that it would make systematic scanning of the image set more difficult.

04 *That's horrible, I mean, you'd just miss everything, […] your eye would probably pick out the strongest colours and you'd end up missing loads of little things like that up there, and, nah, I don't like that.*

05 *It's cluttered, it's kind-of… messy, so everywhere I go I keep thinking I've left a few, I haven't looked at a few.*

They also disliked the overlapping images; both of these disadvantages of continuous MDS arrangements were identified by the participants in our initial experiment (Section 4.1.5).

04 *[The overlap] just kind-of defeats the purpose really, yeah, I mean you want to be looking at the photos, you know. […] It makes you feel like you want to move them, you want to see what's going on underneath, and you can't.*

05 *I still like to be able to see the whole photo with a border around it. Borders make such a difference. I want to see the entire thing.*

It is possible, of course, that some of the negative reaction can be explained by the fact that the participants had, until this point, always used a grid-based method of viewing image sets.

04 *Perhaps if I'd started off with this, maybe I'd think the* 10×10 *was horrible, but I doubt it! [laughs]*

05 *Maybe it's because I'm used to software [...] as very kind-of mathematical and gridded, the way I operate with a mouse and on a computer, I much prefer it to be structured. If I was working in real life, I'd have all my photos like that on the table, and I'd work better like that, but because it's on software, [...] and the photos are small anyway, you definitely need [...] them to be structured, in a grid.*

With 200 instead of 100 images, presented in a 15×15 proximity grid, the thumbnails were even smaller, making them harder to see, and the participants found it "overwhelming" (04) to try to choose only three from so many images, twice the number they had become used to. This also made it more difficult for them to locate an image that they had come across earlier but hadn't selected.

04 *My God, oh, so many. Just to see what's going on. [...] In one way it's nice, you think, oh, great, I must have plenty to choose from here, it's quite nice in that sense. But, yeah, it's a little bit too much.*

Browsing strategies

Participants' browsing strategies (Table 5.15) varied with each search, but generally, they wanted to get an overview of the available images first, before using the zoom facility to look in more detail at those which immediately stood out. With the visual or caption-based arrangements, they could concentrate on a particular area of the screen if they had an idea in mind after reading the text, or developed one while browsing the images. If they were having trouble finding images they liked, or wanted to make sure they hadn't missed any out, they might then try to cover the screen more methodically, usually from top to bottom, left to right.

04 *At the moment I'm pretty happy with my three choices, but I don't feel I've looked through the photographs enough yet to be totally happy with them.*

05 *Now I'm just moving around in order, so that I can cover the whole screen without missing out any.*

05 *I keep going to the top left-hand corner, probably from habit. [...] I guess if I was Hebrew I'd be a bit different.*

Category	*04*	*05*
Discussing the order in which the images are looked at	5	11
Wanting to make sure none are missed out	7	6
Switching between arrangements	5	10
Choosing a large number of candidate images	1	9
Total	18	36

Table 5.15: The part of the framework relating to participants' browsing strategies, showing the number of comments classified under each heading.

It was often helpful to the participants to have two different arrangements of the same set of photographs. As well as using each one for a different type of requirement, they could switch between them to see if different images stood out in the other arrangement, and to make sure they were happy with their selections. It also provided variety and helped to maintain their interest in the task. After switching arrangements, participants sometimes used an existing selection as an anchor point from which to start looking.

04 *I'm just going to go back to the library again, just to have a quick, see if there's anything I've missed.*

05 *All the different methods of grouping them, work well together [...] because if I get stuck or a bit stagnant with one, I know I've got that for backup, and it kind-of refreshes [...] my interest.*

Sometimes their initial choice of arrangement type was purely arbitrary, with participant 05 saying at one point that she'd chosen the caption-based arrangement "probably just because it's the first option".

As she did in the original experiment, participant 05 tended to choose a large number of candidate images (the experiment software allowed up to nine to be selected at once), and then cut that down to three: "I can make a proper decision after I've chosen a few."

Selection criteria

The transcripts also allowed us to study in more depth how the initial requirement developed from simply finding three photographs of the given place. As we noted in Section 2.5.3, Garber and Grunes [47] found that art directors often alter their picture selection criteria while a search is in progress. Table 5.16 shows the different types of selection criteria that our participants used, and how frequently each one was expressed. We chose to distinguish between those occasions when they stated a requirement and then searched for images matching it ("S"), and those when they were already looking at a particular photograph and justifying their decision to select or reject it ("J"). The former indicates directed browsing, and the latter undirected browsing.

There were notable differences between the two participants in this regard. Participant 05 tended to pick images out from the set and then comment on why she'd chosen them, building a shortlist which she would then cut down

Criterion	*04*		*05*	
	S	J	S	J
being of a generic object or type	21	18	5	4
conveying a particular mood or impression	7	8	5	11
having certain visual properties (e.g. colour, composition)	3	5	4	20
being of good technical quality	0	2	0	10
relating to something mentioned in the text	8	1	3	0
working well together as a group	7	16	6	17
standing out, catching the eye	15	2	11	13
being obviously of that place and not elsewhere	4	7	2	14
covering different/enough aspects of the place	3	9	2	6
being of a particular landmark, or part of a country	8	9	3	5
personal (subjective) reasons, or context	1	2	2	12
Total	*77*	*79*	43	112

Table 5.16: The part of the framework relating to criteria for selection or rejection of images, showing the number of comments classified under each heading. They are further categorised according to whether the criterion was expressed as part of a **S**earch requirement, or to **J**ustify a decision to select or reject a particular photograph.

to form her final three, making further comments about why she was keeping or rejecting certain photographs. She was less likely than participant 04 to express a requirement and then look for photographs to match it, admitting half way through the experiment "that's the first time I've actually had an agenda to cover".

At first, participant 04 tended to stick closely to the given text, but after she did the two searches where no text was provided, her reliance on it faded, and she became more likely to develop requirements in the course of looking through the photographs. Participant 05, on the other hand, always treated the text purely as background information, especially if she already had some knowledge of the place.

05 *Because I've been there, I've interpreted [the text] in my own way, and I agree with what this said basically, so [...] because I agree with that ...] I kind-of discarded that and just kept to my own views because I thought it'd be similar anyway.*

The participants did not usually say that they were looking for a photograph of high technical quality, as this was implicit, but sometimes a photograph would be rejected for being of low quality, for example "cheesy", or "too dark" (both 05). Also, they would sometimes comment on the overall quality of the whole set of available photographs and their suitability for the text.

05 *The text is [...] basically saying [New York is] a wild and crazy place, and these photos aren't exactly wild and crazy, in my view, [...] I was expecting punk rockers and mad transvestites, and so it's kind of limited my want to look at them all, because they're not what I wanted, really.*

The participants would occasionally select images for their visual properties, usually colour or composition. Participant 04 mentioned that "I'm really sticking to the shots that have got some big focus going on", probably because close-up images tend to be distinctive at thumbnail size. Visual properties were an important factor in determining whether a photograph was eye-catching, and many images were chosen for this reason. Often they were simply happened upon, but sometimes participants would explicitly state this as a requirement:

05 *I'm just generally looking around for anything that catches my eye.*

The most commonly expressed requirements were for generic objects or types of photographs, for example a "city shot" (04), or "animals" (05). Conveying a particular mood or impression could also be important.

04 *Now that's lovely, I think I'll keep that, that's a really nice romantic shot.*

05 *[The text says] "travelling independently", see I'd choose different photos if I was going for a kind-of backpacking holiday rather than a touristy one.*

As before, it was very important to the participants that their selected photographs should work well together as a group, complementing each other while containing a diverse range of colours and subjects. When the participants had selected a number of candidate images, choosing the final three was usually a matter of deciding which ones worked best together. If they had only selected two, the participants would often say that they wanted a third image to work well with the others.

04 *Find something [...] in contrast to the first photograph I took, when I was thinking of the exotic. [...] I've got two sort-of beigey, goldy subjects already, so I want something to contrast to that, have a little bit of variation going on there.*

04 *I'm going to look for three photographs that are of different matter. So maybe the inside of a building, the outside of something, and maybe some landscape shot.*

05 *I'm quite happy with those three. Red, green and blue, they're like complementary colours.*

It was important that photographs should be obviously of the place in question and not anywhere else, and also that they should cover enough different aspects of the place.

04 *I'm looking for sort-of white buildings at the moment. [...] They're mentioned in the text, and also when I think of [...] Greece, I think of little white square buildings.*

05 *You want it to be recognisable, you don't want it to look like any old city, which some of these photos definitely could make it.*

Selecting photographs of a particular landmark may help to make the set more obviously representative of the place.

05 *I've just suddenly remembered, I was looking at these and I thought, oh, this could be any city, like Hong Kong or somewhere like that, so what do I need for New York, well, Statue of Liberty.*

At the same time, avoiding clichés is desirable.

04 *I'm going to go for Times Square actually, because I'm trying to stay clear of the Statue of Liberty, it's too much of a... cliché, I think.*

04 *I think it's going to be hard to avoid putting in the Eiffel Tower.*

Not having much personal knowledge of the place seemed to make the participants more likely to concentrate on how the photographs looked.

04 *One thing that I've noticed with the Greece one in comparison to this one: because I've been to some of the Greek Islands, the captions were more useful. [...] Never been to San Francisco, captions could say anything really. So I'm really just [going to go] purely on the photos.*

05 *I'm tending to go to more arty shots though, [...] because I don't know that much about the country.*

As well as knowledge of the place, participants sometimes had other subjective reasons for making selections; a photograph might relate to one of their interests, for example. Context was also a factor, with participants' choices occasionally being affected by the images they selected in previous searches.

The user interface of the experiment software

Participant 05 was more likely than 04 to explicitly make comments about the experiment software (Table 5.17). After the original experiment, she said that she wanted to see her selected photographs at full size, and this time she commented that she would like to be able to rearrange them, to decide which ones work well together.

She felt that the magenta highlights around the selected thumbnails were "slightly distracting", and that perhaps they should be less bright. However, she did like having the selections identified within the overview display, especially with a visual arrangement, because "you know what areas you've gone for more".

There were no complaints about the size of the thumbnail images in the 10×10 proximity grids, but as already noted, participant 05 felt that they were too small in the 12×12 grids, and both participants said this about the 15×15 grids.

We introduced the text query facility for the last two places, where the participants had 200 images to choose from. This was popular in theory: the participants liked the idea of being able to describe what they wanted, with 05 saying it made her feel she was being more "properly selective". Of course, they first had to decide which words to type in.

Category	*04*	*05*
Wanting to move the selected images around	0	6
The usefulness of the selection highlights	1	8
The size of the thumbnails and magnified images	5	10
The search facility	4	12
Total	10	36

Table 5.17: The part of the framework relating to the user interface of the experiment software, showing the number of comments classified under each heading.

04 *Ah, now this... This, I like, but, the same old thing again, if you didn't know anything about France, you'd be stuck! [...] But then you've got the text.*

05 *The search, to me, is going to be a lot quicker, so I immediately went for search and just put in "beach". I'm quite happy, that I've chosen a beach one, because that's what I wanted.*

They were somewhat discouraged upon realising that the image captions may not be detailed enough to contain descriptions of the content at each level.

05 *Because I don't know the software, [...] I'd need more time to, experiment with different words, like, I mean, [if I search for] "people", how, I don't know how the captions really work. [...] Need to know how somebody's captioned it, if they've done it really thoroughly, or really logically, or really descriptive, or...*

Participant 04 spelled one of her query terms wrongly, entering "chateux" instead of "chateau" when selecting from the photographs of France, meaning that there were no results, even though some of the captions did contain "chateau". She subsequently returned to using a browsing strategy, perhaps having lost confidence in the query facility.

04 *I think I'm going to put "chateux", because I wouldn't mind a little castle. [No results]. Whoops. Ah, oh well, that's a shame. [...] So, yet again, when all else fails, [...], see what's going on, quick skim round.*

5.4 Conclusions

The experiments in this chapter all used the simulated work task approach, to test MDS (proximity grid) arrangements in a realistic setting, using designers as the participants. We assumed that a broad text-based search had already been performed (in this case via the selection of a category), and that the participant was browsing through the results, which were all relevant to the topic of the requirement.

The infodesign 99 study compared a visual arrangement to a caption-based arrangement, and found that the participants tended to prefer the latter, because it grouped together semantically similar images. The Anglia experiment

was more formal, and compared a visual arrangement to a random arrangement. Although the participants were faster with the random arrangement (efficiency), they preferred the visual arrangement (satisfaction), and felt that they had made better selections using it (effectiveness). Frøkjær, Hertzum, and Hornbæk [43] noted that these three variables are not necessarily related to each other, and that efficiency may be less important than the other two when the task is complex. The Anglia follow-up was exploratory, aiming to gain more qualitative insights into the usefulness of the different arrangement types, and test a number of different variations.

In general, arranging a set of thumbnail images according to their similarity seemed to be useful to the participants, especially when they wished to narrow down their search to a particular subset of the presented images. Using visual similarity, the set can be divided into simple genres, although the local contrast of the arrangement is reduced, so individual images stand out less. A caption-based arrangement does not have this problem, and helps to break down the set according to meaning, although its usefulness depends on the level of detail in the available captions, and labels may be necessary to help the user understand the basis for its structure. Even if annotations are not available, a purely random arrangement can also be useful, especially when the user does not have a particular requirement in mind, because individual images usually contrast to their neighbours and thus are more likely to stand out.

There is also evidence that, for some people, having access to different arrangements of the same set of images is useful. As the user's requirement develops, different views of the same set of images may support different selection criteria, or aspects of utility, as discussed in Section 2.5.3. Switching between arrangements also gives the user a fresh view of the image set.

This chapter has dealt with image browsing in stock photograph collections, whose contents are mostly unfamiliar to their users. In the next chapter we consider the organisation and browsing of personal photographs, an application domain that is increasing in importance as the use of digital cameras becomes more common.

Chapter 6

Personal photography

Photography has been a popular pursuit since the advent of Kodak's colour Instamatic cameras in the 1960s. More recently, digital cameras have become widely available, and because they do not need film, they are very cheap to use, and photographs can be viewed immediately after they are taken. The cameras themselves are still expensive, but as prices fall, consumer surveys predict that their use will become more widespread, resulting in large personal collections of digital photographs. Computer-based systems to store these photographs, facilitating future browsing and retrieval, will therefore be required.

Although others have studied personal photography from a sociological and anthropological point of view (such as Holland [54]), there has been very little research attention given to how people organise and browse their photograph collections. The two studies in this chapter used respondent and field research strategies (as introduced in Section 2.5). In the first study, we interviewed keen photographers about their current practices with regard to their physical photograph collections, and asked them for their opinions of a number of possible features of a computer-based system for managing personal photographs. Subsequently, AT&T Laboratories Cambridge developed such a system, known as *Shoebox*, and this was used in the second study. To investigate how digital photograph collections will be organised and browsed, and to evaluate the usefulness of Shoebox's features, a group of volunteers were each given a digital camera and a copy of the software, and were interviewed at the beginning and end of a six month trial period. Their actual usage of the system was recorded in log files.

Of course, we were also interested in finding out whether using multidimensional scaling to arrange a set of photographs according to visual similarity, as described in previous chapters, would be useful for personal photographs. In the initial study, interviewees were shown examples of MDS arrangements of stock images, and were asked for their opinions of its potential usefulness. Although the Shoebox developers originally intended to incorporate MDS features in some form, the constraints on the project made this impossible. Instead, we asked the participants in the Shoebox trial to make a small set of their photographs available for the creation of an MDS arrange-

ment; most of them agreed, and their comments about it were recorded.

6.1 The initial study

As we discussed in Chapter 2, the users of a general-purpose photograph collection are generally unfamiliar with its contents, and want to search the collection to find images matching a requirement. In contrast, the principal users of a personal photograph collection are the photographer and members of her family; they are very familiar with its contents, which relate to their memories of life events. We therefore expected that the typical requirements associated with a personal photograph collection would be somewhat different to those already identified for unfamiliar collections.

This distinction between familiarity and unfamiliarity is also present with textual document collections. The bulk of information retrieval research makes the assumption that the documents in a collection will be unfamiliar to the user, but there has also been some interest in studying personal information management: how people organise familiar documents such as electronic mail messages or word processed files. In an early study, Malone [75] interviewed office workers about how they organised the paper documents on their desks, and observed that there were large individual differences in degree of organisation. Some of his participants had systematically organised almost all of their documents into what he called *files*: explicitly titled units with the elements intentionally arranged in a particular order. Those at the other extreme tended to let documents build up into *piles*, with elements which were not intentionally arranged in any order. Others had some mixture of the two methods. Often a document was left in a pile so that it would be a visible reminder of something to be done.

Later, Barreau and Nardi [6] considered management of electronic documents at work. They identified three types: **ephemeral** (only needed for a short time, such as a "to do" list), **working** (frequently used documents that are part of the user's current work), and **archived** (old documents that are infrequently accessed, such as reports on previous projects). In the context of a computer system, the word "file" typically has a different usage to that employed by Malone: it refers to a single document, a set of which may be contained in a named directory. Directories may also contain sub-directories, and users can set up their own hierarchy of these in order to classify their files. When Barreau and Nardi set their participants the task of finding a particular file, the preferred method was choosing a directory and browsing through it, rather than trying to remember what the file was called and using a query facility; this was only used as a last resort. It was easier to remember where they put something, and recognise the file name from those in the directory, instead of having to recall it. Because each participant had created her own directory structure, naming the files herself, finding one again was unlikely to be difficult unless she had forgotten which directory the file was in, or the directory contained hundreds of files. Like paper documents, electronic documents also have an important function as reminders of things to be done, for example

being left on the desktop at the end of a day, to be noticed the next morning.

Whittaker and Sidner [124] studied workers' usage of electronic mail messages, and, like Malone, they identified individual differences with regard to organisation. Many of their participants had given up on filing their e-mail, or only did it occasionally, resulting in large numbers of messages remaining in the "inbox". To these participants, categorising e-mail did not seem worth the effort, because the rough chronological ordering of the messages in the inbox was enough to find them again, if needed. Some researchers, such as Lansdale and Edmonds [65], have argued that perhaps the default method of organisation for all electronic documents should be a chronological ordering, instead of using directories. Users would not be forced to categorise or even name a document, and they could retrieve it by remembering its context, rather than its location in a hierarchy.

Because they are familiar to their users, personal photographs are part of the wider class of personal documents, and many of these earlier findings are likely to be applicable. For example, it appears that familiarity with a collection makes browsing more common than querying as a means of locating a particular item, and that the context provided by chronological ordering can be helpful when searching a disorganised collection. However, the existing studies have only considered documents used as part of people's work practices, and personal photographs are not normally taken for this purpose. More specialised research is therefore required, and for the initial study we decided to adopt a similar approach to the researchers mentioned above, concentrating on interviewing a small number of participants in depth, to gain a qualitative insight into how people currently organise their personal photograph collections.

6.1.1 Participants and method

We recruited twelve photography enthusiasts, and asked each of them the same set of questions, in a structured interview, while taking notes on their answers. They were asked about their current practices, and then gave their opinions of possible features of a computer system for managing personal photographs. The list of questions is given in Appendix D, starting on page 229. The interviewees were seven men and five women, ranging in age from 24 to 62, with an average of 36. Five of them had some background in computer science, and all but one regarded themselves as familiar with the use of PCs. The estimated size of their personal collections ranged from 500 to 10000 photographs, with an average of about 4000. Of course, this sample is highly unlikely to be representative, but as we have already mentioned, our goal was to gain initial insight, not make generalisations about the population as a whole.

Only one of the participants was making any regular use of a digital camera at the time of this study (August 1998), although six of the others had tried using one. Most had not yet "gone digital" because of the cost of the cameras, and low image resolution. Some of the interviewees said that because they would always want to have prints of at least some of their photographs, they would not start using a digital camera until high-quality printing was cheap.

6.1.2 Present practice

In the first part of the interview, the participants were asked about their collection of personal photographs: what it is used for, how it is organised, and how they go about the task of searching through it for particular photographs.

Uses

Personal photographs tend to be of special events such as holidays or weddings, and are taken to help remember the events or people involved, and record them for posterity. The most common use is showing them to friends or family, when describing the events captured in them, for example a recent holiday. People who use slide film will do this with a slide show, usually focusing on a particular event, corresponding to one or more boxes of slides. A more diverse slide show is rare, because it would involve selecting slides from a number of different boxes, and then having to replace them later.

All of the interviewees said that the frequency with which they look at their pictures tends to decrease over time; recently taken photographs are kept handy for a short period, perhaps being shown to friends or family who visit, before being put away with the rest of the collection. Recent photographs could be said to fit into Barreau and Nardi's *working* category, because they are used frequently, with the rest of the collection being *archived*, and only browsed occasionally. Some photographs are usually regarded as being better or more special than others, and these may be placed in albums, or framed and displayed at home. Personal photographs appear to have no equivalent of the *ephemeral* category, and although they do have a reminding function, it is of a different nature to that of documents, far more long-term.

Some of the participants take photographs as part of their work or hobby (for example, prints to use as the basis of artwork, or slides to illustrate an academic talk) and these are normally kept with personal photographs, although their uses are different. Similarly, keen photographers will take some photographs for their aesthetic qualities rather than their personal meaning, and these photographs are also mixed up with the rest of the collection. It would be useful if these photographs could be marked, and perhaps filed separately.

Organisation

As expected, there were individual differences with regard to organisation of photographs. All but one of the interviewees had made *some* deliberate attempt at organising their collection, although only four had done this for *all* of their photographs. This includes the two participants who primarily use slide film, who both have a system involving numbering and labelling the slide boxes. The other participants all primarily use negative film, meaning that the majority of their photographs are prints, which are either organised into albums, or simply left in the packets in which they are returned from processing. Most participants felt that they should put their prints in albums, but often keep them in the original packets until they get around to the task, by

which time a number of packets may have accumulated. If albums are used, the photographs in the albums will be looked at more frequently than those left in the packets.

Albums are mostly classified by individual events, such as holidays, often with one album per event. However, if similar events tend to recur, such as a series of holidays in the same country, or a series of hill-walking trips, then some thematic organisation may be used. This was only observed in the collections of two interviewees.

Almost all of the interviewees said that they would only put the "good" photographs in albums, separating them from the "bad" ones as part of the process. Bad photographs may be technically poor, catching a person in an unflattering pose, or just "boring". Two participants said that they always throw away any bad photographs, regarding them as clutter. Seven said that they could never bring themselves to throw any away, and the others were somewhere in between. If bad photographs are not discarded, they are usually kept in the original packets.

Within an album, the photographs will usually be kept in rough chronological order, with perhaps some small adjustments to make the arrangement more meaningful or aesthetic. Similarly, in the packets, the photographs are usually kept in chronological order, as this is how they are returned from processing, although they often become out of order when browsed. One participant said that although he is lazy about organising photographs, he will usually move the best ones to the front of their packet, to make them easier to find.

The participants were asked if they write anything down about their pictures. Several said that they occasionally write notes (such as names, places, or dates) on the back of photographs, as a reminder, and for the benefit of other people who might inherit the pictures. When the photographs are in albums, they can be given labels. Others will only write a broad title on the album or packet, to remind them of the whole group of pictures it contains.

Searching

The interviewees said that if they searched their collection for something in particular, it was usually for the album or packet containing photographs of a certain event. They may remember its physical appearance (for example, the colour of the album, or the name of the developer), which can act as a search cue. Occasionally, the search will be for a particular photograph, and in this case an attempt will be made to remember which album or packet it is in, before starting to search. They might remember the rough point in time it was taken, and use that as a guideline to dip in to the collection and then move backwards or forwards. This is analogous to the way in which Barreau and Nardi's participants would locate a particular file by first trying to find the right directory, then browsing through it. The process is made more difficult if the albums, packets, or photographs are out of chronological order, or if photographs have been loaned or given to other people, or filed away elsewhere. There may also be problems if a photograph could be in a number of different

albums or packets, for example if more than one visit has been made to the same holiday destination.

The interviewees had difficulty remembering occasions when they had wanted to find photographs matching a more general requirement. In those situations which were recalled, the requirement was related to the presence of a particular person in the picture, or its perceived quality (for example, to find a good photograph of a certain relative, to put in a frame received as a gift). This will usually be attempted by thinking of a particular picture or pictures to look for. However, an exhaustive search of the collection may be carried out in the case of something major, like a death in the family. General requirements may also come from someone other than the owner of the collection, perhaps another family member. If they are not familiar with the collection, it becomes like a stock photograph library to them, so any organisation or annotation is invaluable.

6.1.3 Future possibilities

The participants were given a list of 19 possible features of a computer system for managing personal photographs, and were asked to rate their potential usefulness on a scale of 1 ("very useful") to 4 ("not at all useful"). They did this by circling numbers on a questionnaire sheet (Figure D.13 in Appendix D), while being asked to explain their ratings in more depth. We used a scale with an even number of points, so that the interviewees would have to make a decision one way or the other about the usefulness of each feature. They were given as much clarification of a feature's description as they felt they needed to be confident in giving a rating. However, an explicit "don't know" option was also available, in case they could not decide.

The tables in this section show the mean and median usefulness rating of each feature. For consistency with the other tables in this dissertation, the scale for the responses has been changed to 0 ("not at all useful") to 3 ("very useful").

Organisation and annotation

Table 6.1 shows the ratings that participants gave to possible system features for organisation and annotation of photographs. The ability to organise photographs into folders of some kind was perceived as very useful by all of the participants, and when asked how they would use this facility, participants said that they would arrange their photographs according to events, in chronological order. This is, of course, the way in which most of them store their photographs at present. Slide shows were also rated highly; these will be much easier to create with digital photographs than with physical slides, because the "slides" will not have to be returned to their original boxes.

Most current approaches to indexing of image data are based on associated annotations, which may be keywords or free text captions. If people entered notes about their photographs, these could be automatically indexed and used to facilitate querying. Many interviewees said that awareness of the existence

Possible system facility	*Mean*	*Med*
Organising photos into separate "folders"	3.0	3
Creating "slide shows" of selected photos	2.3	3
Adding a title to a photo	2.0	2
Typing notes to associate with a photo or group of photos	2.2	2.75
Speaking notes to associate with a photo or group of photos	1.5	1.5
Having spoken notes automatically "recognised"	1.8	2

Table 6.1: Participants' ratings of possible organisation and annotation features.

of such a facility would make them more likely to enter notes, and make those notes more detailed than normal.

Each of the interviewees had a different idea about how they would use annotations. Some, for example, would want to explicitly assign particular keywords to images, creating their own categorisation. Others would want to enter free text notes about a picture. As well as describing the actual content of the picture, these might explain more about its context, such as the events before and after those depicted. When constructing slide shows, extra notes could be added, specific to the context of that show.

Talking about photographs might seem more natural than typing in notes. It would then be possible to perform speaker-dependent speech recognition, to allow any annotations to be indexed in the same way as free text notes, enabling subsequent text-based retrieval. Although it would be impossible to recognise all words correctly, previous studies have shown that information retrieval can still be effective, despite recognition errors [17].

Opinions were divided about this possibility. Some interviewees expressed enthusiasm about it, saying that they could imagine themselves recording while talking to someone else about the photographs, thus taking notes as part of the normal process of browsing. However, others said that they would feel self-conscious about speaking to a computer, and would first have to plan what to say, and make sure their comments were appropriate for any context. Some said that they would not trust a computer to correctly recognise their speech.

One participant said that she would occasionally like to talk about pictures as she was taking them, because that would certainly be much easier than typing while carrying a camera, and she thought that she would not bother to do it retrospectively. Some digital cameras now offer an audio recording feature, although the recording conditions are likely to be less favourable than at the desktop, making accurate speech recognition even more difficult.

Searching

Table 6.2 summarises participants' responses with regard to potential search facilities. Unlike prints, digital photographs can be arbitrarily reduced in size, enabling a large number of thumbnails to be presented simultaneously, and this was perceived as very useful. Some participants did stipulate that the thumbnails would have to be big enough to get a sense of their content. The

Possible system facility	*Mean*	*Med*
Seeing all of the photos in a "folder" at once, reduced in size	2.8	3
Rapidly scanning through a group of photos (like video fast forward)	2.4	3
Seeing a very large number of photos at once, with "similar" photos clustered together	1.9	2
Searching for photos according to the date and time they were taken	2.3	3
Searching for photos based on the text of your notes	2.6	3
Searching for photos based on the colours present in them	0.8	1
Searching for photos based on the textures present in them	1.0	0.5
Searching for photos based on their layout/composition	1.3	1
Searching for other photos "similar" to a given one of your photos	1.7	2
Searching for other photos "similar" to another picture, e.g. a drawing	0.7	0
Choosing a region or regions of a photo and asking for other photos with similar regions	1.4	1.5
Specifying the position of a selected region in a photo	0.6	0
Specifying the relative positions of selected regions	0.3	0

Table 6.2: Participants' ratings of possible search features.

ability to flip rapidly through a series of full-size photographs would also prove popular, although at present this may be difficult to implement in practice.

Of the search features listed, searching based on the text of notes was the most popular, and, as mentioned above, the interviewees said that this facility would make them much more likely to write or record detailed notes.

The idea of "searching according to date" was popular, although participants had different interpretations of what it meant. Browsing the collection based on a chronological ordering seemed to be far more desirable than entering a query for a specific date. It is usually more difficult for people to make an accurate absolute judgement of when an event occurred than remember when it occurred relative to other events [65]. Digital cameras record the date and time that a picture was taken, and this will allow a number of ways of searching by date to be implemented.

Visual queries were far less popular, with many of the interviewees stating that they would only use them if they had to, for example if the collection was completely disorganised so that even directed browsing was difficult. The two participants who did express a definite interest both described themselves as "very visual" (an architecture student and an art student), and said that they would like to use colour, texture, and composition when searching for more abstract photographs. Generally, composition was considered more important than colour and texture, as it was perceived that this would allow classes of photographs to be identified: wide-angle views versus close-ups, group shots versus portraits.

With regard to searching for photographs similar to an example, several participants pointed out that similar-looking images are likely to be taken at the same event, and so browsing based on chronological ordering would give equivalent results. When asked how they would expect "similarity" to be defined, those with a computer science background mentioned visual similarity ("the combination of colour and composition", "areas of colour"), but most of the others expected semantic meaning ("all of the landscapes together, and all of the cityscapes", "Christmas photos", "pictures of two people kissing"). Many content-based image retrieval systems allow users to create a query image by drawing it, but this idea was almost universally unpopular, either be-

cause of the perceived effort involved, or a lack of faith in its potential effectiveness. One participant said "when I'm thinking of a picture, I do have a vague image of what it looks like, but the problem would be translating that to the computer". It is likely to be easier to use a browsing strategy to recognise a remembered photograph, than attempt to construct a visual query from memory.

Image segmentation techniques allow the user to select individual image regions to use as a query, rather than a whole image. Wood, Campbell, and Thomas [128] have incorporated this into a system which has been used to index a collection of personal photographs, although the group have not performed any studies to test how useful people might find region-based queries for this sort of collection. The interviewees could see the benefit of being able to specify one or more regions from an example image (particularly if these were faces), but specifying region positions was regarded as too "specific", "detailed", or "technical".

As part of their interview, the participants were presented with two printed examples of continuous MDS arrangements, based on visual similarity, containing photographs from the Corel collection. Because this study was carried out before the experiments described in the previous three chapters, we used a figure from a paper by Rubner, Tomasi, and Guibas [103] (created with the EMD similarity measure described in Section 3.1.3) and an arrangement produced using the demonstration system on their Web site. The second of these is reproduced in Figure D.14 in Appendix D.

Table 6.2 shows that participants gave the MDS arrangements a slightly higher rating on average than any of the potential visual query features. Many of them were interested in the idea, even if they had been sceptical about visual queries. One said that this might be an easier way of searching for images by colour, rather than trying to translate a vague requirement into a query. Other interviewees agreed that it might be interesting to see their photographs arranged in a different way, and that perhaps it would highlight visual connections between them that would not otherwise have been noticed. Several participants could not see any point in grouping unrelated photographs according to visual similarity, although some felt that it might be useful to lay out a group of related photographs, perhaps to assist in creating an aesthetically pleasing arrangement of them.

6.1.4 Discussion

In summary, the participants in the initial study would like to have their photographs categorised and ordered, usually in albums, but do not always make the effort to do it, often leaving them in the original packets. Malone found that one of the main reasons for leaving documents disorganised was feeling that organisation was not worth the cognitive effort involved, that is, fitting documents into an existing classification scheme, or creating a new one. Attempts at organising personal photographs are often half-hearted, because just keeping the packets related within themselves, roughly in order, is enough to make simple browsing fairly easy. Slides are more difficult to browse than prints,

however, and it is interesting to note that both of the slide photographers in this study had made the investment in organising them.

Having photographs available in digital form will immediately make many aspects of organising them easier, even without a dedicated system for doing so. For example, a large collection will no longer take up a lot of physical space. Multiple copies of a single photograph can be created easily and for free, making it easier to give pictures to friends, and to file the same item in more than one place. For example, a good portrait of someone could be kept in its original chronological order, as well as in a separate folder of pictures of that person, and another folder of good portraits in general. A system for managing personal photographs can potentially offer many more benefits, such as facilities for annotation, browsing, and querying.

If people wish to search for something within their photograph collection, it is usually a particular photograph, or a set of photographs from a single event. Fulfilling a more general requirement (for example "all the pictures of Dad") potentially involves looking through the entire collection, and it is not clear whether people would be more likely to carry out this sort of search if it could be done more easily. In a computer-based system the search criteria could be specified using a query, and then all of the images in the collection could be automatically matched against it. Participants did express some interest in specifying requirements textually, although to make this possible, of course, the photographs would need to be annotated. At present, few people make notes about their photographs, again because it may not seem worth the effort. However, the prospect of subsequent querying made the interviewees say they would be more likely to provide annotations. Since these are so valuable for querying, it seems clear that systems should aim to provide a range of different annotation features, to suit all preferences from assigning keywords to recording spoken comments.

However, given the importance of browsing for this type of collection, systems should aim to support this above all else. As Malone [75] states, computer systems may be able to help users by performing some automatic classification, which for digital photographs could be done using their date and time stamps.

An interesting question still to be answered is whether people will want to have digital versions of their existing print collections. One of the interviewees said that having all of his prints and slides easily accessible in digital form "would be like finding an old friend". Scanning each photograph individually would be too laborious a task, but it is likely that more advanced scanners (able to process prints, negatives, or slides in bulk) will be able to get around this problem, and it may be provided as a commercial service.

6.2 The Shoebox trial

The initial study considered how keen photographers manage their physical photograph collections, and the potential usefulness of certain features in a system for managing a collection of digital photographs. Of course, it is dif-

ficult for people to accurately predict how their practices will change as they adapt to new technology, and therefore the findings of the second part of the study should be regarded as provisional.

The Shoebox system [81], developed at AT&T Laboratories Cambridge, has basic facilities for organising photographs, and also allows users to create typed or spoken annotations (which can be automatically transcribed), and issue textual and visual queries. A group of volunteers were asked to use Shoebox over an extended period of time, and this allowed us to study how they organised and browsed their digital photograph collections, as well as get a more accurate picture of the usefulness of particular system features.

More specifically, we were interested in finding out how much effort the participants would put into organising their digital photograph collections, compared to their non-digital collections. Given the results of the initial study, we wondered if the Shoebox trial participants would be more likely to annotate their collections, either by typing or speaking, with the incentive of being able to issue queries based on the annotations. Browsing, particularly by date, seemed to be very important to the interviewees in the initial study, and we wanted to know if this would be the case with Shoebox, especially as chronological ordering can be provided automatically. We did not expect that visual queries would be very useful, but were interested to see what participants would think when given the chance to actually use them.

Because scanning their existing collections of prints would have been very time-consuming, the Shoebox trial participants were given digital cameras by AT&T, and were asked to start taking photographs using those. This meant that their collections of digital photographs would be recent and fairly small, probably making their organising and browsing practices unrepresentative of those that will be adopted by people in future, who will only ever take digital photographs. However, the situation is representative of the one in which most people will find themselves, a few months after purchasing a digital camera. We were also interested, therefore, in seeing how participants' photograph taking practices would be affected by switching to a digital camera, whether they would use their photographs differently, and if they would miss having printed versions of the photographs.

6.2.1 Digital photography

Digital cameras, such as the Canon PowerShot S10 (Figure 6.1), generally have the same features as other modern cameras, such as automatic focusing and exposure setting. However, the user does not have to buy film or pay for developing it, because the photographs are captured by an image sensor, usually a charge coupled device (CCD), and then stored using removable flash memory such as a CompactFlash card, typically in JPEG format. The number of photographs that can be stored depends on the size of the card, the image resolution, and the degree of compression used. The only other practical limitation is the battery life, which at present is typically fairly short. Photographs can be transferred to a computer via a number of methods: a special drive for the memory card, the USB port, or the serial port. Afterwards, the flash

(a) front view (b) top view (c) back view

Figure 6.1: The Canon PowerShot S10 digital camera.

memory can simply be erased and used again.

The cameras often have an LCD screen (such as the one visible in Figure 6.1(c)), which can be used as a viewfinder, or to look at the photographs currently in memory. Photographs can be reviewed immediately after they are taken, allowing a shot to be re-taken if necessary. The photographs in memory can also be viewed at a larger size by connecting the camera to the video input of a suitable television set.

Although the Shoebox trial concentrated on digital photographs produced by digital cameras, there are other ways of obtaining them. Schemes such as Kodak's PhotoNet, for example, allow people to have conventional film developed as normal, while also having the photographs automatically uploaded in digital form to a World Wide Web site.

6.2.2 The Shoebox software

Shoebox allows users to organise their photographs into **Shoebox databases**, each of which may contain a number of **rolls**; for example, in Figure 6.2, the **index view** on the left shows that the database sample.box is open, and that it contains three rolls. The first roll, AT&T Laboratories Cambridge has been selected, causing its contents to be displayed in thumbnail form in the **roll view** on the right. The index view also displays a list of the photographs in the roll. An individual photograph can be displayed by selecting it, either from the index view or the roll view. For example, Figure 6.3 shows the result of selecting the photograph gip2.

Rolls and photographs all have names (such as AT&T Laboratories Cambridge and plaque in Figure 6.2); by default these are the names of the directories or files from which they were originally imported, but can be changed within Shoebox. Keywords can be assigned by editing a photograph's properties. Longer descriptions can be entered as text annotations, or recorded as spoken annotations. A single annotation can be attached to a group of photographs, allowing the user to make a general comment about the group, and then a specific comment about each individual photograph, if required. The speaker icons at the bottom right of each thumbnail in Figure 6.2 indicate that these photographs all have spoken annotations. Speech recognition (using Mi-

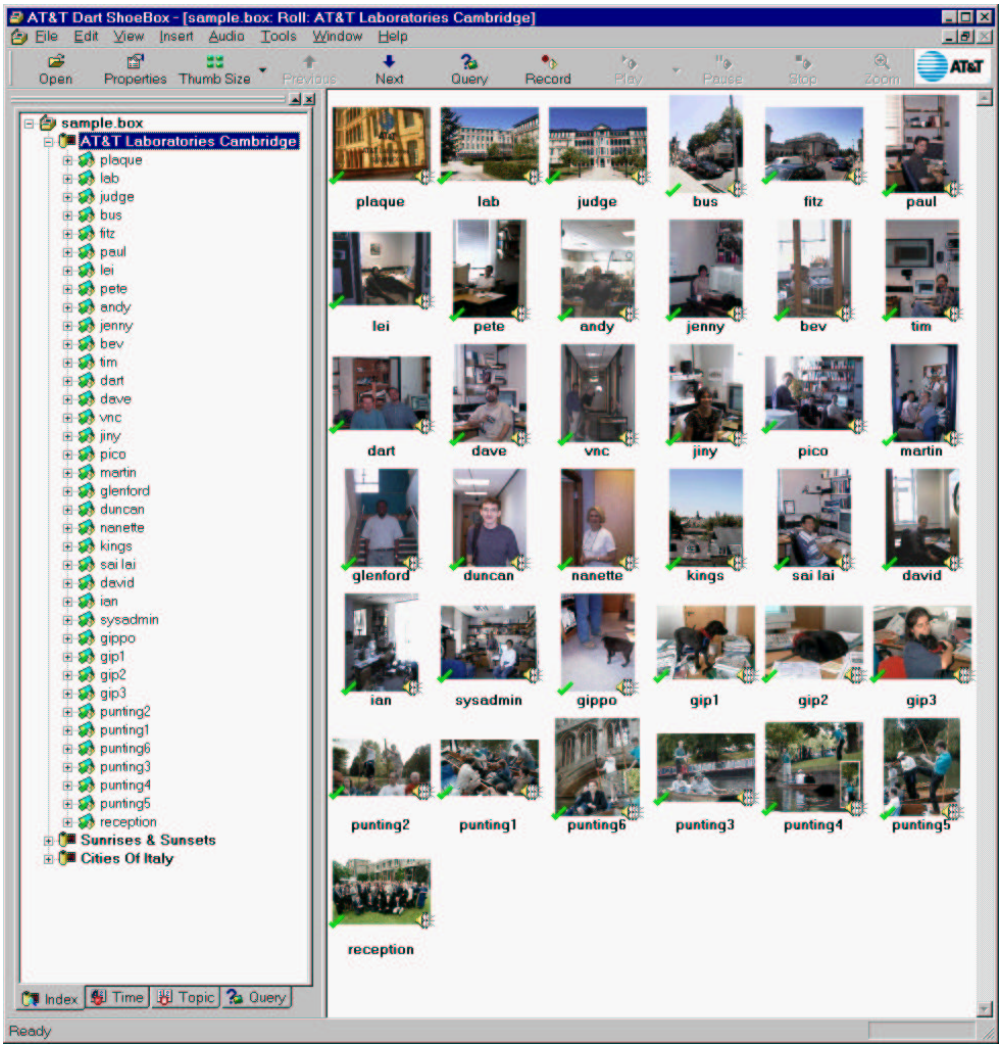


Figure 6.2: A Shoebox screenshot, showing the contents of the roll AT&T Laboratories Cambridge in thumbnail form.

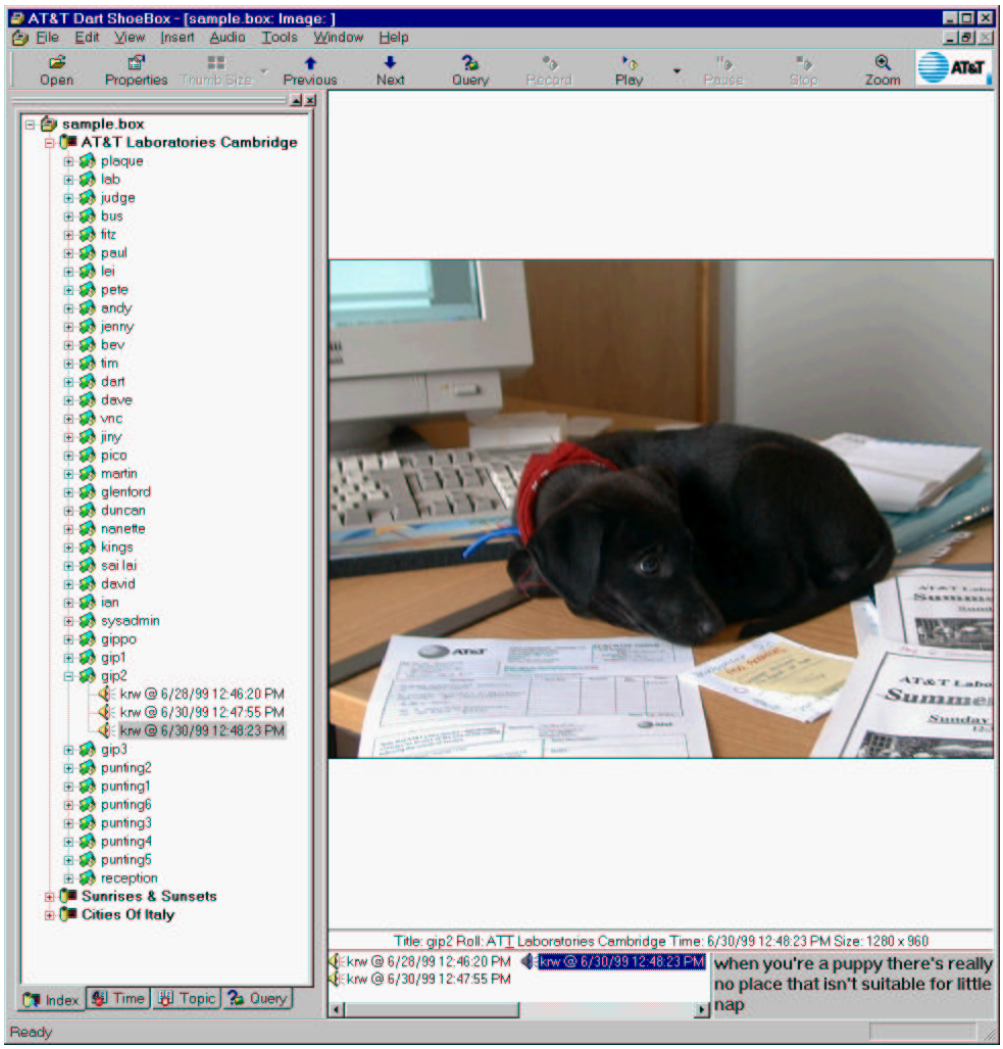


Figure 6.3: A Shoebox screenshot, showing a single image from the AT&T Laboratories Cambridge roll.

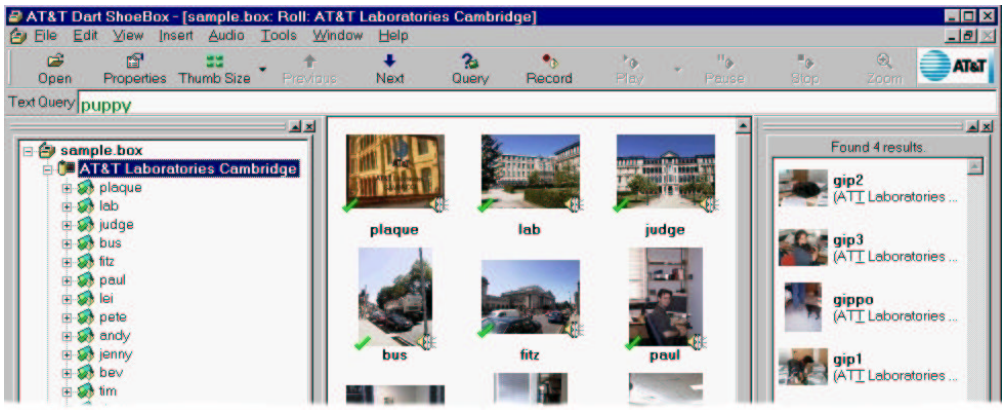


Figure 6.4: Issuing a text query in Shoebox.

crosoft's Speech API) can be applied to these annotations to generate a text transcript, once the user has trained the recogniser to her voice. The panel at the bottom of Figure 6.3 shows that the gip2 image has three spoken annotations associated with it; the transcript of the selected annotation is visible on the right of the panel.

Any text associated with photographs can be indexed, enabling text-based queries. At the top of Figure 6.4, the user has entered the query "puppy", and there are four results, shown in the **results view** on the right. As with the index view and roll view, photographs can be selected from the results view and displayed individually.

The first three tabs at the bottom left of Figure 6.2 (Index, Time, and Topic) allow the user to switch between different views of the database in the left-hand panel. The Index tab is selected by default, and displays the index view. Clicking on the Time tab replaces it with the **timeline view** (Figure 6.5), which sorts all of the photographs in the Shoebox database into chronological order, and groups them into days according to their date and time stamp, displaying them as tiny thumbnails. Selecting a date from the timeline view causes all photographs from that date to be displayed on the right, in a **date view**.

Clicking on the Topic tab displays the **topic view** in the left-hand panel. If the images have been annotated, this contains a list of keywords automatically extracted from the annotations. Selecting one of these keywords displays thumbnails of all of the photographs associated with it, in the main panel, and is analogous to issuing a query containing that keyword.

Shoebox is also capable of automatically segmenting the images into regions, allowing users to issue visual queries. A number of different segmentation methods are available, and users can select one from a list of options. The green ticks at the bottom left of each thumbnail in Figure 6.2 indicate that these photographs have all been segmented. Once this has been done, it is possible to issue a visual query-by-example by right-clicking on a photograph displayed individually (as in Figure 6.3) and selecting "Find similar" from the pop-up menu. The user can also click on the image to select regions from it, and then issue a query using only those regions. In accordance with the

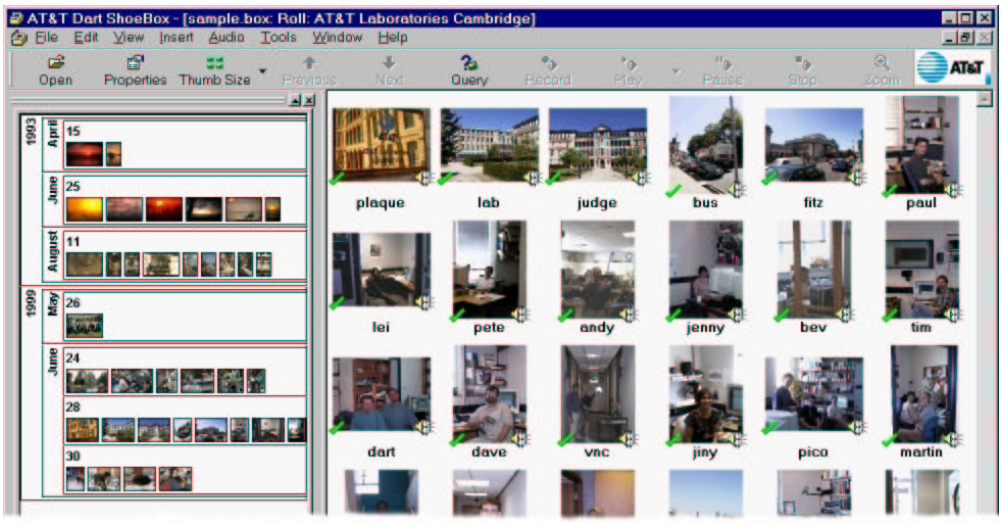


Figure 6.5: Shoebox's timeline view (left).

findings of the initial study, region positioning is not taken into account, and Shoebox does not have a query-by-sketch facility.

Photographs may be displayed as a slide show, where any spoken annotations associated with a photograph will be played back as it is shown. For the initial study, slide shows were classified as an organisation facility, but in Shoebox the user does not explicitly create and save slide shows; the set of photographs to be displayed is simply the contents of the current roll, by default, and although a set can be individually selected for a slide show, this cannot be saved, and there is no facility to create annotations specific to the context of the show. In the Shoebox trial, therefore, slide shows were considered to be a facility for using photographs, rather than organising them.

Shoebox can also generate a set of HTML pages, containing selected photographs and their accompanying annotations, for publishing on a World Wide Web site. A Shoebox database can be exported to XML, in order to back it up, and potentially allow it to be subsequently imported by other applications. Photographs can be printed or sent via e-mail from within Shoebox. Finally, users can rotate photographs, and perform simple colour correction; Shoebox can also launch other applications, enabling more sophisticated image editing to be performed.

6.2.3 Participants and method

There were fifteen participants, ten male and five female, with an age range of 24–38. None of them participated in the initial study. They were recruited from among the employees of AT&T Laboratories Cambridge, and volunteered to take part in the trial. Of course, people who were unwilling to switch to digital photography would not have volunteered to take part, and as in the initial study, we do not claim that these participants are representative of the population as a whole. The estimated size of their existing collections ranged from 300

ID	*Gender*	*Camera*	*Windows version*	*PC*
R1	Male	Canon PowerShot S10	Windows 98	laptop
R2	Male	Canon PowerShot S10	Windows 98	laptop
R3	Male	Canon PowerShot S10	Windows 98	laptop
R4	Female	Canon PowerShot S10	Windows 98	desktop
R5	Male	Canon PowerShot S10	Windows 98	desktop
R6	Male	Nikon Coolpix 950	Windows 98/2000	laptop
R7*	Male	Nikon Coolpix 950	Windows 98	laptop
R8*	Female	Canon PowerShot S10	Windows NT 4	desktop
S1	Male	Olympus C900	Windows 98	laptop
S2*	Male	Canon PowerShot S10	Windows NT 4	laptop
A1	Female	Canon PowerShot S10	Windows 98	laptop
A2	Female	Canon PowerShot S10	Windows NT 4	desktop
A3*	Female	Nikon Coolpix 950	Windows 98	desktop
D1	Male	Canon PowerShot S10	Windows NT 4	desktop
			Windows 2000	laptop
D2	Male	Canon PowerShot S10	Windows NT 4	desktop
			Windows 98	laptop

Table 6.3: The participants in the Shoebox trial, and the cameras and platforms they were using. Those marked with an asterisk (*) were not given the Shoebox part of the second interview, for reasons explained in the text.

to 3000, with an average of about 1000. This is smaller than in the initial study, where the participants were selected because they considered photography as a hobby.

The participants are listed in Table 6.3, with identifiers to indicate whether they are **R**esearchers (all with first degrees, and most with PhDs, in computer science or engineering), **S**upport engineers, or **A**dministrative staff, to give some background to their comments, which are quoted later in this chapter. Two researchers who were involved in the **D**evelopment of the Shoebox software have been classified separately; they were used as pilot participants, and their data is excluded from the tables, although they are occasionally quoted. A1 was the only participant not employed by AT&T Laboratories Cambridge; she is married to R1, and works as a finance manager. The most common camera in use was the Canon PowerShot S10, illustrated in Figure 6.1.

At an initial meeting for all participants, they were given a demonstration of Shoebox and its features[1]. They each received a short Shoebox manual, with instructions on how to obtain the software from the local network, and a headset microphone, for making voice annotations. We explained to them that their actions would be recorded in log files, and that they would be periodically requested to submit these (one for each Shoebox database) by copying them to a particular place on the network. It was made clear to them that we would not be able to actually look at their photographs. The log files allowed us to gauge the usefulness of different Shoebox features, by measuring how frequently they were used. In addition, we interviewed the participants twice, at the beginning and end of the trial, and they filled in two questionnaires at

[1]by Ken Wood of AT&T Laboratories Cambridge.

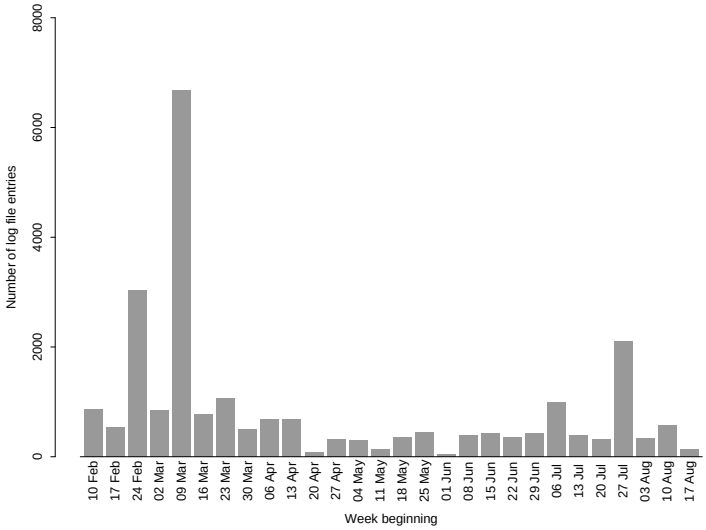


Figure 6.6: Overall usage of Shoebox during the period of the trial.

each interview.

The timetable of the trial was as follows:

Initial meeting: 08 February 2000
Request for log files: 09 March 2000
First interviews: 23 March 2000 to 04 April 2000
Request for log files: 24 April 2000
Request for log files: 15 June 2000
Request for log files: 03 August 2000
Second interviews: 22 August 2000 to 03 October 2000

Figure 6.6 shows the total number of log file entries, for all participants, for each week of the trial. There was a large burst of activity at the beginning, especially after the initial request for log files, but this evened out towards the end.

The first interviews

The first interviews were carried out after about the first or second month of usage. Participants were asked about their existing non-digital collections, so that their organising and browsing practices could be compared with those subsequently adopted for their digital collections. They were also briefly asked for their first impressions of digital photography and Shoebox, and about their fledgling digital photograph collections. The interviews were of a similar form to those in the initial study, with participants being asked a fixed set of questions, and answers recorded in the form of notes. Appendix D contains a list of the questions, starting on page 233.

They were also asked to fill in two questionnaire sheets: one to indicate their level of agreement or disagreement with a series of statements about their non-digital photograph collections (Figure D.15 in Appendix D), and another to indicate how useful they found each of a given set of Shoebox's features (Figure D.16 in Appendix D).

The second interviews

In the second interviews, at the end of the trial, participants were asked in more detail about their digital collections. The questions are listed in Appendix D, starting on page 235. These were based on the digital photography questions from the first interview, but were refined and expanded, based on the answers received then. Participants were also asked for their final verdict on Shoebox. These interviews were recorded on tape and subsequently transcribed.

Again, participants were asked to fill in two questionnaire sheets. This time, the first questionnaire asked them to indicate their level of agreement or disagreement with a series of statements about their *digital* photograph collections (Figure D.17 in Appendix D). Most of the statements were the same as those in the non-digital questionnaire, with a few digital-specific items added. The aim was to compare people's opinions about their non-digital and digital collections. Then, participants filled in the same Shoebox questionnaire used in the first interview, to establish whether their opinions about its features had changed in the course of the trial.

The participants marked with an asterisk in Table 6.3 (R7, R8, S2, and A3) were not given the Shoebox part of the second interview, and did not fill in a second Shoebox questionnaire. R7 and A3 were unable to get Shoebox to work reliably, but were still taking digital photographs, and so were forced to find another means of organising them. R8 and S2 had not taken any more digital photographs since the first interview: R8 had been too busy (moving house, getting married, and changing jobs), and S2 did not feel he had done anything else that was worth photographing.

As we have noted, Shoebox does not contain a facility to create arrangements of photographs using multidimensional scaling, so participants were asked to make a set of about 80 of their photographs available, to allow MDS arrangements to be demonstrated to them at their second interview; all of the participants listed in Table 6.3, except S2 and A3, agreed. These participants were shown a 10×10 MDS proximity grid arrangement of their photographs (based on visual similarity), with a chronological grid arrangement for comparison. Their comments were recorded and transcribed.

6.2.4 Results

This section describes the main findings of the Shoebox trial, concentrating primarily on the organisation and browsing of personal digital photographs, but first of all briefly noting how switching to a digital camera affected participants' practices with regard to taking and using their photographs.

Tables 6.4, 6.6, 6.9, and 6.10 show the frequency of use of the main Shoebox features, as recorded in the log files. A1 and R1 share a collection and therefore submitted combined log files, and they were also interviewed together, although each filled in separate questionnaires. In some of the comments quoted here, participants mention using a feature, but this usage is not represented in our tables. There are two possible reasons for this. One is that participants sometimes accidentally deleted log files when recreating a database after a system crash. Secondly, R1/A1, R3, and S2 were initially using early versions of Shoebox, prior to the official start of the trial, which did not have full logging enabled. All data from this period has been omitted, and as S2 did not submit any further log files, he is missing from these tables altogether. R7 is also missing; he was unable to provide any log file data because of the difficulties he had with Shoebox. As mentioned above, D1 and D2 are excluded, because they were involved in Shoebox's development. Therefore, eleven participants (all of them bar R7, S2, D1, and D2) are represented in these tables, with R1 and A1 grouped together.

Tables 6.5, 6.8, and 6.12 summarise participants' questionnaire responses with regard to the usefulness of selected Shoebox features, from both the first and second interviews. The scale used is 0 ("not at all useful") to 6 ("very useful"). Participants who did not fill in a Shoebox questionnaire at the second interview (those marked with an asterisk in Table 6.3: R7, R8, S2, and A3) are not represented in these tables, and, again, D1 and D2 have been omitted, leaving nine participants in total.

Tables 6.7 and 6.11 show participants' level of agreement or disagreement with a series of statements about their photograph collections. Again, the scale is 0 to 6, with 0 representing "strongly disagree" and 6 representing "strongly agree". The responses from the first interview relate to their non-digital photographs, and those from the second interview relate to their digital photographs, facilitating comparison between the two. All participants bar D1 and D2 are represented in these tables, thirteen in total.

Both types of questionnaire table indicate, for each item, how many participants selected a particular rating (from 0 to 6) as well as the mean and median rating. The *I* column indicates whether a set of ratings is from the first or second interview. The *p* column gives the *p* values of Wilcoxon signed-rank tests, used to determine whether differences in ratings between interviews were statistically significant.

The quotes in this section are all drawn from transcriptions of the second interviews, and were chosen to be representative of participants' opinions. In cases where many participants made similar comments, usually only one quote has been used.

Taking photographs

The participants had been much more prolific in taking photographs since starting to use a digital camera. By the end of the trial, the approximate size of participants' digital photograph collections ranged from 200 to 1000, with an average of about 500, which is half the average size of their existing non-

digital collections. Per participant, the proportion ranged from 20% to 200%. As we have already noted, it costs nothing to take a digital photograph, and typically many more images can be fitted into flash memory than on a film, so participants often take several digital photographs where they would only have taken one with a conventional camera. This perhaps makes it more likely that they will obtain a good photograph, but does tend to mean that they also have a larger number of bad photographs. It should be possible to automatically group together multiple shots of the same scene [72], assisting the user in selecting the best ones.

A3 *I'm more likely to take a lot more photographs now, in the hope that there's going to be one good one out of them all, whereas with a conventional camera I don't tend to do that, because I've got to pay for developing the ones that aren't very good, so I certainly take a lot more, and I probably discard a lot more.*

Professional photographers usually work in this way: a photojournalist, for example, may shoot many rolls of film in order to get a single news picture, and it is interesting to see that amateur photographers seem to adopt a similar style when it does not cost them anything to do so. People may also be more likely to take "risky" photographs (because if the picture does not turn out as intended, they have lost nothing) and "everyday" photographs (because they do not have to save the film for a special occasion). This can open up possibilities, allowing people to think of taking photographs in situations where they previously would not.

R1 *Most of the photos are the same things you would take photos of, you just take more of them. But there are some things where you wouldn't normally bother, but the fact that it's immediate, and that it doesn't cost anything means that you take a few extra.*

Photographs can be examined immediately on the camera's LCD screen, and any which turn out badly can simply be deleted or re-taken.

A3 *My daughter's now got to the point where, if someone gets a camera out, "can I see the picture?". She hasn't twigged that not all cameras can show you the picture now.*

R6 *I was at my girlfriend's graduation, and she had her camera with her, and she wanted me to take some shots with that as well, and I felt it was so weird, not being able to see that you'd got a good shot! I felt much more comfortable with my camera.*

Many participants had only taken digital photographs during the trial, as they found them to be a perfectly adequate replacement for the snapshot-style photographs they would normally take, and were happy to view them on a television or computer screen.

A1 *I hate looking at things on the screen normally, like at work, I print everything out, whereas with this, it just seems normal now, because the quality's so good on the screen, it's just as nice, and it's bigger.*

Those participants who were used to taking photographs with an SLR camera, however, still preferred to use it for occasions where photographs of higher quality than snapshots were required.

Using photographs

Being able to take photographs which do not cost anything, and see them immediately, allows people to use photography in a new way, as a means of information capture [16].

D2 *I use the digital camera for practical things that I wouldn't have used an ordinary camera for. For instance, we had a meeting today where we made notes all over the whiteboard, and I just went up afterwards and took digital photos of all the whiteboard frames, and sent them around to all the people that were at the meeting, so we had sort-of instant minutes, and that's something which you couldn't do with an ordinary camera.*

These photographs often have only a short-term purpose, and may be deleted once they have served that purpose, rather than be kept with the rest of the collection. This means that digital cameras enable photographs to be used like the ephemeral documents described by Barreau and Nardi [6]. For example, a user might take photographs while shopping for something expensive, such as furniture, or spectacle frames, and then use the pictures to help remember and reflect on the options later that day, perhaps showing them to a friend or partner. Photographs can also be taken of something before it undergoes a major change, such as a person about to get a haircut, or a room about to be painted, and then immediately compared to the final result. A recent field trial found that a group of children, when given digital cameras capable of transmitting photographs to each other across a wireless network, started using them as part of the games they were playing [74].

However, as in the initial study, the most important use of digital photographs is to record holidays or other significant events, and then show the pictures to friends or family. People who are present when the photographs are taken can be shown them immediately, like Polaroids, using the screen in the camera. Any older photographs in flash memory can also be looked at in this way.

A1 *If you take a couple of pictures, people are always interested, I don't know if it's because it's quite novel, but everyone's "oh, can I have a look?", so you tend to show them the photos that you've just taken of them, or whatever, but then also they'll want to look back through what you've got, so very frequently I hand round the camera and people are flicking back through with the little screen, and looking at what I've taken.*

Digital camera screens are only a few centimetres across, and it is desirable to be able to view photographs at a larger size. Once photographs have been transferred to a computer, its monitor can be used to view them. Within Shoebox, the user can browse through rolls, or set up a slide show to automatically

Command	*R1/A1*	*R2*	*R3*	*R4*	*R5*	*R6*	*R8*	*S1*	*A2*	*A3*	*Total*
Open a Shoebox database	30	20	8	24	14	32	18	15	155	18	334
Print out a photo	4	0	13	0	0	0	0	0	1054	0	1071
Publish a set of photos as HTML	0	2	0	0	0	10	3	0	0	0	15
Display a slide show of selected photos	0	11	1	2	4	12	0	1	26	1	58
Rotate a photo	0	21	0	0	0	0	0	5	124	0	150
Colour correct a photo	0	0	0	0	5	0	0	2	0	0	7

Table 6.4: Frequencies for Shoebox commands related to usage of photographs.

Feature	*I*	*Usefulness*							*Mean*	*Med*	*p*
		0	1	2	3	4	5	6			
Creating 'slide shows' of selected photos	1	0	0	0	2	1	1	5	5.0	6	0.321
	2	0	0	0	0	2	1	6	5.4	6	
Publishing a set of photos as HTML, to put on the Web	1	0	3	0	0	0	1	5	4.2	6	0.951
	2	1	1	0	1	2	0	4	4.0	4	

Table 6.5: Shoebox questionnaire responses, from both the first and second interviews, relating to usage of photographs.

display a series of photographs, one after the other. The slide show facility was rated slightly, but not significantly, higher at the second interview than at the first (Table 6.5), and was used at least once by all but two of the participants (Table 6.4). Of course, with conventional photography, it is impossible to create a slide show unless the photographs have been taken using slide film, and as noted in the initial study, the slides are usually only drawn from one box; with Shoebox they can be selected from any roll.

R2 *I use the slide show facility within Shoebox to quickly whizz through a roll of film, just showing people what I've been doing, on my laptop at home.*

Digital cameras can usually be connected to the video input of a television, allowing people to give a slide show without having to transfer the photographs to a computer. This is especially convenient when visiting friends or family, because cameras are smaller and more portable than laptop computers, and, as one participant (S2) noted, "there are more TVs than computers".

R3 *Our telly, the input's on the front, so it's dead easy to take the camera and just plug it into the telly, and do the slide show off the camera, a hell of a lot easier than getting it all onto the computer.*

Only the current contents of the camera's flash memory can be looked at in this way, but these are usually the most recent photographs, which, as noted in the initial study, are those most likely to be shown to others. Of course, if necessary, older photographs can be transferred from a computer back into the camera's memory.

Showing non-digital photographs to people elsewhere involves sending them through the post. If the photographer does not wish to give away the original prints, she must buy an extra set at the time of processing, or order individual reprints specially. Copying digital photographs is trivial, however, and thus may make people more likely to share their pictures with others. As

well as printing them out and posting them, they can be made available via the Internet, an increasingly attractive option as more people gain access to it. Photographs may be sent by e-mail, but because the file sizes are often very large, a World Wide Web site may be preferred for a whole set of photographs, especially if there are a lot of people who would be interested in seeing them. There are a number of Web sites which provide services to help people publish their photographs online. Participants were somewhat divided about the usefulness of Shoebox's facility to generate HTML, with only three of them (Table 6.4) using it. Its rating did not change significantly between the first and second interviews (Table 6.5).

R5 *When my family visited me, and we've taken pictures together, I e-mailed them back to them, which is more convenient than trying to print them. They decide what to do with the picture.*

R2 *There was a mountain bike weekend in Wales, and I took lots and lots of photos, and there were a lot of people from Cambridge but a couple from Bristol, and elsewhere, and I just was able to e-mail the list and say "hey guys, I've got the photos up [on the Web]".*

Participants still wanted to have prints of their photographs for certain purposes, for example to send to people who do not have Internet access, or to look at without having to switch on a computer or television. Usually, only prints of selected photographs, not the whole collection, are desired. Participants wanted to have their favourite photographs printed out at the highest possible quality, to be added to their existing permanent collection of special photographs, for example put in an album, or displayed at home. This is again analogous to professional photography, where a contact print is used to decide which of the photographs on a film are worth printing at full size. Several participants felt that not being forced to have a print of every photograph was a definite advantage of digital photography.

D2 *I'm coming to realise that I would like to have prints, certainly for the good photos that I've taken, just for having them in a box at home, a permanent non-electronic collection someplace. And back home to show people. I wouldn't necessarily print every single one of them.*

A3 *I have printed some out; obviously I don't print out as many as we would have developed. I pick what I consider to be the best ones and print out a few. They print out very well; the quality's very good.*

Although Table 6.4 shows that only three of the participants who submitted log files had printed photographs from within Shoebox, others simply printed the files directly. Many participants complained about the fact that it is currently quite difficult and expensive to get a high quality print of a digital photograph. Often, they had simply used a colour laser printer and ordinary office paper, which did not really seem adequate for display, or permanent storage.

Digital editing of photographs was relatively uncommon, despite the fact that scanning is no longer necessary. Shoebox provides some simple editing

facilities, with rotation being more popular than colour correction (Table 6.4). Some participants chose different applications to perform these, and other, editing functions.

Organisation and annotation

In accordance with the results of the initial study, participants in the Shoebox trial have so far organised their digital photographs in a similar way to their non-digital photographs, using Shoebox's facility to create rolls, which all of them rated very highly (Table 6.8). Some participants separate the photographs into rolls according to event, while others simply import the entire contents of the camera's flash memory into a single roll, meaning that the roll may contain photographs of a number of different events (as is often the case with conventional film).

R2 *I take a load of photos, and then I download them and I put them in a folder, and I label that with the date that I download it on, and that gives me as much organisation as I want, really. Actually, when I moved house, all my non-digital photos are literally in a shoebox now, and they just fit in one. That's as much organisation as I want.*

R1 *When your CompactFlash card's full, or you're going away and you want to make it empty, before you go, download all the files, and they end up in one directory because of that, and then when they get imported into Shoebox, they stay in that roll. The roll gets renamed, if it's lucky, to something that vaguely describes all the photos that are in it.*

Participants gave the statement "I am content with the way my photograph collection is organised" significantly higher agreement scores for their digital collections, and scores for the statement "my photograph collection is intentionally organised" were higher on average, only just missing significance at the 0.05 level (Table 6.7). However, from the interviews and log files, it did not appear that the participants had actually put much more effort into intentional organisation. They simply feel more organised, because all of their photographs are in one place, in folders that correspond roughly to events, in chronological order. By default, the photographs within a roll appear in order according to their date and time of insertion into Shoebox, which usually corresponds to the order in which they were taken, using the automatic time stamp on the file. This fits with Whittaker and Sidner's finding [124] that e-mail users often feel that they do not need any organisation for their messages, other than a chronological ordering of the inbox.

Some of the initial study participants took photographs as part of their work, or a hobby, and said that it would be useful to them to be able to file these separately from their more conventional holiday photographs. A digital photograph, of course, can be filed in more than one place by simply copying or linking to the file, without removing it from its original chronological context. Some of the Shoebox trial participants also had photographs they would like to file separately, but only one had actually done this. Table 6.6 shows

Command	*R1/A1*	*R2*	*R3*	*R4*	*R5*	*R6*	*R8*	*S1*	*A2*	*A3*	*Total*
Insert a new roll into the Shoebox	0	1	1	6	3	5	5	1	24	0	46
Insert new photo(s) into the Shoebox	1	7	5	32	4	35	13	15	41	11	164
Delete a roll, photo, or annotation	101	18	1	77	14	21	12	9	30	1	284
Move a photo between rolls	2	0	0	23	0	7	20	0	164	0	216
Copy a photo between rolls	0	2	1	0	0	2	37	0	2	0	44
Edit roll properties (e.g. name)	1	0	1	13	1	3	13	3	28	0	63
Edit photo properties (e.g. title or date)	0	0	0	30	7	0	0	0	656	0	693
Edit annotation properties (e.g. transcript)	33	0	0	0	0	0	0	0	1	0	34
Insert a typed annotation	0	0	0	0	0	0	24	0	90	0	114
Insert a spoken annotation	200	80	0	58	24	0	57	5	2	0	426
Play a spoken annotation	452	63	0	348	134	0	9	10	1	0	1017
Record a spoken annotation	139	80	0	57	24	0	57	5	1	0	363
Stop playing or recording	587	110	0	388	134	0	59	9	2	0	1289
Perform speech recognition on annotations	1	1	0	16	13	0	0	1	0	0	32

Table 6.6: Frequencies for organisation-related commands in Shoebox.

that creation of new rolls was fairly infrequent, as was moving or copying of photographs.

S1 *I've got separate folders for different types of things like motorbikes, racing. They could be duplicated from one to the other so it's just a case of copying them. So if I want to have a look at the motorbike shots I can just go straight into that.*

Although not quite significant at the 0.05 level, on average participants said that they were less likely to separate good and bad photographs in their digital collection, but more likely to throw away bad photographs (Table 6.7). If bad photographs are not deleted in the camera, they can be deleted at the organisation stage, once transferred to a computer. However, some participants do not tend to do this, and simply find themselves using the good photographs more. This reflects the individual differences with regard to throwing photographs away that were identified in the initial study.

D2 *I tend to keep them all, I tend to then just e-mail the good ones, or print the good ones, whatever I do with them I do just with the good ones, but I tend to keep them all anyway. [. . .] I don't have a good-bad folder or anything. I do treat them differently but I don't physically separate them.*

As mentioned by R1 and R2 above, a roll will usually be given a name, to help the user identify the set of photos it contains. This facility was generally rated highly (Table 6.8), and all but two of the ten participants listed in Table 6.6 had changed the name of a roll. Only three had changed the name of a single photograph, and this was considered a less important facility than changing the name of a roll (Table 6.8). Shoebox's default title for a photograph is the file name, and if this is assigned by the digital camera, it may not seem descriptive enough, a "meaningless number" (A3), which several participants said would motivate them to change it. With respect to longer annotations, only two participants had added typed notes to their photographs. Seven had recorded spoken annotations, eight including R3, whose attempts are not represented in Table 6.6 because he was using an earlier version of Shoebox at the time.

Statement	*I*	*Agreement* 0	1	2	3	4	5	6	*Mean*	*Med*	*p*
"My photograph collection is intentionally organised"	1	4	0	5	0	1	3	0	2.2	2	0.057
	2	1	1	1	1	3	1	5	4.1	4	
"I regard organising my photo collection as a chore"	1	3	1	1	2	0	4	2	3.2	3	0.752
	2	2	3	1	2	2	0	3	2.8	3	
"I am content with the way my photo collection is organised"	1	3	1	4	3	0	2	0	2.2	2	0.005
	2	0	1	1	0	4	2	5	4.5	5	
"I separate 'good' and 'bad' photos in my collection"	1	2	1	2	3	1	2	2	3.1	3	0.073
	2	5	4	1	0	0	2	1	1.7	1	
"I throw away (or delete) the 'bad' photos"	1	5	3	1	1	1	1	1	1.8	1	0.086
	2	3	1	1	2	2	1	3	3.1	3	
"I write notes about my photos (e.g. on the back, or in an album)"	1	5	1	2	1	2	1	1	2.1	2	–
"I change the names of my photos from the default ones"	2	5	1	1	1	0	2	3	2.6	2	0.673
"I add typed annotations to my photos"	2	7	2	2	2	0	0	0	0.9	0	0.137
"I add spoken annotations to my photos"	2	6	2	3	0	0	2	0	1.4	1	0.596
"Spoken annotations are useful even without accurate speech recognition"[a]	2	4	3	2	0	1	1	1	1.8	1	–

[a] In both interviews, one participant (R7) said that this statement was not applicable to him as he had not tried using spoken annotations, and therefore he did not want to give a rating.

Table 6.7: Comparing questionnaire responses about non-digital and digital photograph collections, with regard to organisation. The questionnaires were largely the same, allowing the responses for each collection type to be directly compared. However, the statement "I write notes about my photos" was present only in the non-digital questionnaire, and is compared to the three statements from the digital questionnaire which follow it, and refer to different types of annotation. The statement about the usefulness of spoken annotations was present only in the digital photography questionnaire.

Feature	*I*	*Usefulness* 0	1	2	3	4	5	6	*Mean*	*Med*	*p*
Allowing the organising of photos into separate rolls	1	0	0	0	0	1	2	6	5.6	6	0.184
	2	0	0	0	0	0	1	8	5.9	6	
Giving a title to a roll of photos	1	0	0	0	0	0	3	6	5.7	6	0.947
	2	0	0	1	0	0	1	7	5.4	6	
Giving a title to a single photo	1	0	0	0	3	1	2	3	4.6	5	0.432
	2	1	0	2	0	1	4	1	3.8	5	
Typing annotations to associate with a photo or group of photos	1	0	0	1	3	1	1	3	4.2	4	0.248
	2	1	1	1	2	2	1	1	3.1	3	
Speaking annotations to associate with a photo or group of photos	1	0	0	0	2	1	2	4	4.9	5	0.013
	2	1	0	1	3	2	2	0	3.2	3	
Having spoken annotations automatically transcribed	1	0	0	1	0	2	3	3	4.8	5	0.766
	2	0	0	2	1	1	1	4	4.4	5	
Listening to spoken annotations	1	0	0	0	0	4	2	3	4.9	5	0.021
	2	2	2	0	1	2	1	1	2.7	3	

Table 6.8: Shoebox questionnaire responses, from both the first and second interviews, relating to organisation.

In summary, there are three note-taking methods available in Shoebox (titles, typed annotations, and spoken annotations), but the participants were no more likely to use any of them than they were to write notes on their prints (Table 6.7). This contradicts what was predicted by the initial study. As with prints, the participants said they would use their notes to record the names of people and places, especially those which are unfamiliar, and might therefore be forgotten with time. The ability to assign the same annotation to a group of photographs was popular. Most participants said that they would only want to annotate some of the photographs, and that doing it for all of them would be too much work. When the photographs are recent, these details are still fresh in the photographer's mind, and therefore recording them may not seem worth the effort, because the photographs are self-explanatory. It may not begin to seem important until some time after the photographs have been taken, when some of the details may already have been forgotten. This may help to explain why so few of the participants have annotated their photographs in Shoebox: they are still too recent.

R7 *At the time that I'm taking photos off the flash card and into a directory somewhere, I just want to look at the pictures, and I'm not really thinking about, "ooh, I could type text saying what this is", because at the time, to me, it's blatantly obvious what it's a picture of. In a couple of years' time I might be looking at these pictures thinking "what the hell's that? I wish I'd written some notes". [...] So it's possibly something I will do when I've had them for long enough that I can't remember what was going on.*

R3 *I haven't taken the camera to anywhere that I haven't been before — the holidays we've been on this year are places we've been to previously, so when I look at the photos everything's very familiar, whereas if you go to some exotic island, or whatever, some funny waterfall, you very quickly forget, so as soon as you can, you actually write down where it is that you've been to. So I think that is important, but I haven't actually used it.*

As dates are one of the things that people would normally write on the back of prints, having an automatic date and time stamp may make them feel that annotation is less important.

R5 *With the print photos, [...] I used to do some degree of annotation, not always. With the digital I try to do that as well. But because most of the information I need, like the date, for example, is already there, I found that I needed to do annotation less on digital pictures than I would with a print picture.*

The idea of sitting in front of a computer in the evening to organise or annotate photographs may not be appealing to someone who has spent their day working on a computer, and as a result the task may get a very low priority. Participants were divided about whether they considered organisation to be a "chore" (Table 6.7). Some stated explicitly that they had added annotations because they were taking part in the trial, and had felt that they ought to put in some effort.

When showing the photographs to people elsewhere, via e-mail or a World Wide Web page, they usually need some explanation, and often there are stories associated with them. Other researchers have identified the importance of story telling, with one group producing a handheld device specifically for recording stories using personal photographs [5]. In Shoebox, spoken annotations can be played back with a slide show (for the benefit of people who are present) or included in generated HTML (for people looking at the photos on the World Wide Web). Similarly, titles and text annotations can be displayed next to the photographs. This might therefore be a motivation for participants to create annotations.

R6 *I understand that the annotations come out when you publish as a Web page, and I hadn't realised that was there, so I am more likely to start annotating, just so that when other people look at the photos, they know what they're looking at.*

However, some annotations may not be appropriate for everyone to read, because they may contain private comments, or insufficient explanation for someone unfamiliar with the pictures. Also, a single photograph may be presented in different contexts, either on its own, or as part of one or more stories or slide shows, and may require different annotations for each of these. Users should therefore be able to create multiple annotations, and select which ones they would like to accompany published photographs.

S2 *They were notes to me, that's the way I used it, they weren't for anyone else's benefit. I wasn't expecting anyone to view them without me around, I suppose.*

R2 *I'd like it to be optional whether you publish the annotations with the HTML. When I was making annotations about [a recent group trip], I did have to think, I'm going to have to be careful about what I say, because you don't always like everybody on these trips.*

Table 6.8 compares participants' responses at the first and second interviews with regard to Shoebox's features for organisation. The ratings for every feature (except creating rolls) fell over the course of the trial, but this was only significant for recording and listening to spoken annotations. As in the initial study, opinion was divided on the usefulness of spoken annotations in principle: some participants definitely disliked the idea, saying that they would have to plan what to say, and do not enjoy listening to their own speech.

R8 *I just don't like the sound of my voice. I don't like sitting there talking to the computer. I just felt a bit odd, and I always say stupid things because I run out of things to say.*

Others found that it was easier to speak annotations than type them.

A1 *I find it very easy, not bad at all to just sit there and just speak to the photos, "this is this, this is this", click through, it's much less of a chore than actually typing, or whatever, so I can see myself doing that.*

It may be frustrating, however, if the speech recognition does not work as well as expected: although other studies have shown that transcription does not have to be completely accurate for queries to be reasonably effective, all of the participants who tried Shoebox's speech recognition facility found that it produced an unacceptably high level of inaccuracy. This probably discouraged them from putting further effort into recording spoken annotations. Its rating did not fall significantly during the trial (Table 6.8), but this is perhaps because some participants still felt that the feature was useful in principle, even if Shoebox's implementation of it was unsatisfactory.

R4 *I tried very hard with [speech recognition] in the first few months of the trial, and I was very disappointed, because... I even tried to train using a different accent, a more American-like accent, and it didn't work. I was a bit disappointed. But I guess it can be done, I've got faith in it, but somehow it wasn't there this time.*

R1 *I think it's true that it doesn't have to be 100% accurate, because it's just to help, you don't need to find all the photos that have those things in them, but I wasn't sure if it was quite good enough.*

In particular, names of people and places are often wrongly transcribed, and may not even be in the vocabulary, but are usually the most important elements of the annotations. Therefore, if the user is relying on an automatic transcription of spoken annotations as the only means of indexing the content of the photographs, she will have to verify that the transcriptions are correct.

A1 *Can you train the speech recogniser on specific names? [...] My niece's and nephew's names, for example, it can't get that at all, so if we could go in and train those words, it would be quite useful. [...] I went through a few times on the training, the speech thing, and it gets the gist of most of it, apart from, the trouble is that you want it to recognise things like people's names, and place names, because that's normally the core of it.*

Most participants considered that there was no point in making spoken annotations unless there was an accurate speech recognition facility (Table 6.7), although in a few cases participants felt they could be useful in their own right, especially for telling a story. Table 6.6 shows that some participants did listen to their annotations, although as already noted, ratings for this facility fell significantly in the course of the trial.

R4 *For me it's very encouraging to have the opportunity to put down a story, put a collection of pictures together and be able to put a story, or just explain characters throughout, or explain the photos one by one.*

Two participants also mentioned that although they might not want to record spoken annotations for themselves, they could be very special to others in years to come, especially children or grandchildren.

R3 *I think it's one of these things that, you don't like hearing your own voice, but people in a hundred years' time, for instance, or your kids would like to have your voice on there. If I had my grandmother's collection, I'd quite like her to actually annotate it, because then when she's dead, then you've got her voice going through it. But she won't want to do that.*

Finally, some participants did not use spoken annotations at all, because they may not have had the right conditions in which to record them (a shared office, for example), or simply did not get around to plugging in the microphone and training the recogniser, especially if they expected that this would be difficult.

Browsing and querying

Table 6.11 shows that participants' level of agreement with the statement "I browse my photo collection often" was significantly higher for their digital collections, although the mean score for this statement was still only 2.9 for digital, which is on the "disagree" side of the scale. Participants may have browsed their digital photographs more often simply because they were the most recent photographs in the collection. Also, looking at them is generally easier, because one does not have to fetch a box or an album and physically look through it, and then put it back later, or worry about whether it is in the right order. Digital photographs were often more accessible, especially if they were stored on the computer that the participant normally works on.

R2 *I feel more organised now, because I know that the photos are in a particular order in a particular place, whereas if I want to show somebody non-digital photos I get the shoebox out and I flick through the envelopes, and you try and remember, "ooh, was the Africa trip a Truprint or a Jessop?"*

A2 *It just takes two seconds, just click on it and go and have a look, and then I don't have to put anything away. You know when you take photos out, then you have to put them all away, and you usually just plop them back in, and then they're in a worse state than before.*

Participants stated that, as before, their most common requirement was to locate the pictures of a particular event. This is fairly easy with digital photographs because, like prints, they are grouped into rolls, which may be labelled, and within a roll the photographs are generally in chronological order. However, digital photographs have the added advantage that a large number of them can be viewed at the same time by using thumbnails. Plus, the participants had a good memory of their digital photographs and the way in which they were organised, because of their recency. This ease of browsing is another reason why participants felt more content with the organisation of their digital collections than their non-digital collections.

R3 *It's easier to go through a larger volume, so going back, all the photos that I've got, all 200 or so of them, I can just whizz through them in an absolute instant, so*

Command	*R1/A1*	*R2*	*R3*	*R4*	*R5*	*R6*	*R8*	*S1*	*A2*	*A3*	*Total*
Perform image segmentation	0	1	0	0	0	10	0	0	0	0	11
Issue an image query	0	0	3	0	3	28	0	0	0	0	34
Issue a text query	0	0	0	0	0	0	1	0	84	0	85
Issue a text query via the topic view	0	0	0	0	0	0	0	0	14	0	14
Select photo(s) from results view	0	0	0	0	0	37	0	0	185	0	222

Table 6.9: Frequencies for query-related commands in Shoebox.

quickly that I don't even have to think about any other sort of indexing strategy. [...] I'm browsing for an event, that I roughly know when it is, even if I just know relatively when it was compared with other events, I might not know the date, but I can quickly zoom through and find them.

A2 *I wasn't organised at all before, and if I did have any photos in any kind of album, or even those little things that they give you when you have them developed, you'd have to remember maybe a particular colour: "oh yes, those photos, they were in that red album", or something like that. But here, obviously there's a little title to every roll: "Holiday in Italy, May 1998", so you know. And then also when you click on it and then you see all the little thumbnails, just scroll down and you can pick out the photo straight away, so that's really easy. [...] I feel super-organised; I know exactly where a photograph is, and I can find it straight away, it's really good.*

Participants felt that Shoebox's timeline view, which automatically classifies and orders all of the images in a Shoebox database according to date, was useful (Table 6.12), although it was not used very frequently (Table 6.10). This may be because normally, of course, the photographs within a roll are already presented in chronological order. Also, the timeline view uses the time stamp on a photograph's file, which is easily altered by rotation or editing, meaning that a few photographs may end up out of place, and several participants mentioned that they found this annoying. Emerging metadata standards for digital photography, such as EXIF, embed the original time stamp in the file itself, and future photograph management systems should take advantage of this, especially as many photograph editing tools are now capable of recognising and preserving metadata. There are a number of possible uses of such metadata to assist organisation and browsing. Platt [91], for example, has used clustering techniques to automatically partition a collection into folders, by assuming that photographs taken at about the same time are part of the same event. Location data, obtained via GPS, could also be used to group images according to where in the world they were taken.

It is immediately obvious from Tables 6.9 and 6.10 that Shoebox's browsing features were used far more often than its querying features. Those participants who had used queries all said that they had done so just to play with them.

The participants in the initial study expected that text queries would be useful. However, having a query facility did not seem to make the Shoebox participants any more likely to actually want to carry out this type of search. Obviously, textual queries are useless without annotations, and as we have

Command	*R1/A1*	*R2*	*R3*	*R4*	*R5*	*R6*	*R8*	*S1*	*A2*	*A3*	*Total*
Select database from index view	95	35	10	84	40	106	27	34	298	56	785
Select roll(s) from index view	94	101	21	114	56	335	233	160	1259	27	2400
Select photo(s) from index view	353	87	39	397	296	373	446	1238	3588	28	6845
Select annotation(s) from index view	324	17	0	105	43	0	16	6	13	0	524
Select roll from database view	2	10	6	11	10	8	0	7	92	1	147
Select photo(s) from roll view	239	459	62	79	18	371	546	86	2422	11	4293
Select date(s) from timeline view	0	0	0	0	16	2	0	23	24	0	65
Select photo(s) from date view	0	0	0	0	2	2	0	0	0	0	4

Table 6.10: Frequencies for selected browsing actions in Shoebox.

Statement	*I*	*Agreement*							*Mean*	*Med*	*p*
		0	1	2	3	4	5	6			
"I browse my photo collection often"	1	1	5	5	0	1	1	0	1.8	2	0.027
	2	1	2	2	5	0	1	2	2.9	3	
"When I look at my photos, it is to search for something in particular"	1	1	1	2	5	2	2	0	2.9	3	0.595
	2	1	3	2	2	3	1	1	2.8	3	
"When I am looking for something in particular, it is easy for me to find it"[a]	1	2	3	3	3	0	1	0	1.9	2	0.004
	2	0	0	0	2	3	5	2	4.6	5	

[a] In both interviews, one participant (R7) said that this statement was not applicable to him as he never searches for something in particular.

Table 6.11: Comparing questionnaire responses about non-digital and digital photograph collections, with regard to browsing and searching.

Feature	*I*	*Usefulness*							*Mean*	*Med*	*p*
		0	1	2	3	4	5	6			
Changing the size of the 'thumbnail' photos	1	0	0	2	0	0	3	4	4.8	5	0.277
	2	0	1	2	0	2	2	2	3.9	4	
Searching for photos according to the date and time they were taken (the timeline)	1	0	0	1	0	0	5	3	5.0	5	0.952
	2	1	0	0	1	0	2	5	4.8	6	
Using a query to search for photos based on the text of your annotations	1	0	1	0	0	1	4	3	4.8	5	0.025
	2	1	0	2	3	1	2	0	3.0	3	
Searching for other photos visually 'similar' to a given one of your photos	1	0	1	2	0	1	3	2	4.0	5	0.024
	2	2	4	2	0	0	0	1	1.6	1	
Choosing a region or regions of a photo and retrieving photos with similar regions	1	0	1	2	2	0	1	3	3.8	3	0.010
	2	3	5	0	0	0	0	1	1.2	1	

Table 6.12: Shoebox questionnaire responses, from both the first and second interviews, relating to browsing and searching.

already discussed, few participants had created these. It perhaps did not seem worth the effort when the need to use them for queries is so rare, and browsing is so easy. Participants did feel that it was significantly easier for them to find something in particular from their digital collections (Table 6.11), but were no more likely than before to actually want to do so.

R5 *I haven't queried them, because probably I don't have too many pictures, and they're in, to me, a logical-looking order. Getting back to them is quite easy.*

A2 *[Using text queries] was playing around, because I haven't really annotated that many more, I only did it with my India photos, really, and a few others, and that's it, but it just took too much time. And because I can find them so easily anyway, and because I look at them so often, I know where they are.*

R8 *I don't know, even if I could say "find all the photos of [my husband]", which I probably can't, I wonder if I'd ever use that. I don't know, I've never thought, "oh, I wish I had that".*

Queries (and therefore annotations) might start to seem more important as a collection grows larger and the photographs get older and less familiar. Also, requirements that can be answered by queries only occur occasionally, and the period of the trial was probably not long enough for many of them to arise. However, the bigger a collection gets, the longer it will take to annotate all of the photographs.

A3 *They tend to be grouped in events, and we tend to show the photographs of the events, so we don't tend to search for, say, "show me everything with Robin", or "show me everything with Jamie", it tends to be "this was this event, and these were the photographs that we took". But I suppose that's because they're fairly recent events, and if it was further down the line, and we wanted to say "look at all of the pictures of Jamie when he was a baby", then yes, you'd want to search for "Jamie as a baby", and they could span several events.*

R6 *Typing [annotations], although I haven't done it, it is something I definitely will do, and probably quite soon, actually. It's just because the collection's getting to the point where I'm going to need to do that.*

Often the results of a text query are disappointing, because if only part of the collection has been annotated, only results from that part will be returned, making the recall of the query lower than it should be.

A2 *I know that if I put some annotations in and then I call them up [with a query], I won't be calling them all up because I haven't annotated all of them. So then it annoys me.*

D2 *It's one of these things where you know that if you do it for all your photos, it will pay off, and that you will be able to search them, but unless you do it for all your photographs, it's not that useful, and it's hard to get into the habit of doing that, and I haven't.*

Even when the photographs have been given annotations, these may not be sufficiently detailed to match a query. If participants do want to issue queries, names of people are likely to be common query terms, in order to retrieve all of the photographs depicting them. But a member of a couple might simply annotate a photograph of them on holiday with "this is us at our hotel in Venice", without giving their own names, as these are obvious to them and to any family or friends who might view the photograph remotely. However, a subsequent query for one of their names would not return this photograph. Plus, as noted earlier, names in spoken annotations were often transcribed incorrectly by Shoebox, so unless the participants were willing to keep correcting the transcriptions, they had to get used to whichever names the system had chosen, as R1 and A1 discuss below.

R1 [sarcastically] *It's really easy to find pictures of Debbie, our friend, because we type in "deadbeat", and they come up.*

A1 *I thought text querying wasn't too bad, [...] particularly when we knew that Evie, my niece, was capital E, capital V, for example, you kind of got used to it, like you got used to "deadbeat". But, you know, "Mum", and "Dad", we did find pictures of people. Places weren't very good.*

Table 6.12 shows that participants' ratings of Shoebox's query facilities (textual query, visual whole-image query, and visual region query) all dropped significantly over the course of the trial. At the first interview, a number of participants who had still to try using visual queries rated them as being very useful, but as in the initial study, their expectations were rather high (for example, finding all photographs of the seaside, or of a particular person), and Shoebox did not live up to them. However, even when they tried visual queries with expectations that seemed realistic, the results were still disappointing.

R6 *What it considered to be similar was rarely what I considered to be similar. I deliberately, when I got my new car, I took a lot of photos of that, and I thought, right, it's blue, it's big, it's very obvious... and it missed loads of them, it only pulled up a very small handful of the car, which, I don't know anything about the subject, but it looked like quite an easy job to me. [...] I just found it to be pretty much useless, I'm afraid. But also, partly, I just didn't really know what was going on, and how I could improve on what I was getting.*

Shoebox offers a large number of image segmentation algorithms, and it wasn't clear to the participants which ones they should be using; the advantages and disadvantages of each, and how they compared to each other. If image segmentation facilities are to be provided, they should be simplified. For example, many algorithms had "fast" and "slow" versions, suggesting that the slow version would take longer but provide a better segmentation, but R6 said that he had tried the alternate versions and could not see much difference between the results they produced. The following exchange between R1 and A1 is particularly illustrative of the problems participants had with visual queries.

R1 *You pick an image processing scheme, OK, so I picked one down the middle of the list that I thought looked reasonably sophisticated, but I think maybe I made a bad choice… We've got a load of photos that I took at my nephew's third birthday party. He's a real Noddy fan, and my sister made him this Noddy cake, and there was a big Noddy theme. Noddy's a very noticeable character: he's got a blue hat, he's got a yellow scarf with red dots on it, he's got a big round pink face. So there's no getting away from it, there's Noddy everywhere, and we took plenty of photos of this, and we thought, no matter what image processing algorithm you're going to use, that's going to be Noddy.*

A1 *It came up with a picture of somebody in red wellies. Noddy was nowhere to be seen.*

R1 *But it didn't even pick out, you know, there's lots of primary colours, I just thought…*

A1 *But there were two photos of this cake, real close-up photos, all you've got is his round face, with a blue hat, red nose… and it didn't pick the other one! It got 582 matches for this photo, but not one…*

R1 *Yeah, "you want one that looks like it? Here, have a lot!" […] Obviously it orders them, but, there's a landscape, number two is a landscape, with an English country… you're just thinking "are there red spots?", "is there a bright blue hat?"*

A1 *Yeah, we've had absolutely no success whatsoever with that, at all.*

R1 *[…] I think we started off, the first few times we tried it, you were like "oh, well, I can kind of see…", and then it's just showing you the same photos every time, and you're just thinking, "nah…".*

Table 6.9 shows that only three participants tried image queries[2]. Many of the others simply realised that they had never wanted to use visual queries during the trial, and lowered their ratings for this reason. Participants could not think of occasions when visual queries might be useful to them, and so even if Shoebox's visual query tools had performed well for realistic requirements, they would probably not have been used much more often.

R8 *The trouble is, my idea of "get me something similar to that" is if I click on something with [my husband] on it, standing on a rock, and I know there's another photo where he's doing vaguely the same thing, I would want that, I wouldn't want pictures of anybody else standing on the rock, and there's just no way the retrieval system could know about that, that's the problem. I think humans have a lot higher expectations, you have something very specific in mind when you say "show me this, similar to it".*

R2 *I didn't think it would come up with matches that I would actually be interested in. There just seemed to be more obvious ways of finding photos.*

[2] R1/A1's usage of this facility was missing from the log files, for reasons discussed right at the beginning of this Results section.

Shoebox

With regard to Shoebox itself, many of the participants had technical difficulties with it, particularly those using Windows 98, where it crashed more often than was acceptable. Sometimes this resulted in corruption of the database file, and unless participants had kept a backup, they had to recreate it, losing any effort they had put into organisation, such as adding roll names and photograph names. The photographs themselves were not lost: Shoebox either links to the existing files or copies them, and any copies are stored outside the database. Similarly, the files containing the audio annotations were not lost, just the connections between photographs and annotations.

Unavoidably, Shoebox's unreliability affected participants' opinions and usage of it, often meaning that they did not trust it enough to invest much time and effort in organisation. Any organisation of non-digital photographs is unlikely to be lost so abruptly; for example, notes written on the back of a print are attached to the print forever.

R6 *Annotations-wise, there's almost nothing. I don't like to say this, but frankly, on Windows 2000 and 98 it wasn't stable enough to risk putting significant effort into that.*

S2 *I think if it was a wedding or something like that, I'd still feel more confident with a normal 35mm camera. But thinking about it, that's only because so many things went wrong, [...] I haven't actually lost the photos, as such, but the amount of effort I put in has disappeared, and it hasn't built confidence, for me.*

However, a number of participants said they were impressed by its speed and power, especially for fundamental functions like displaying thumbnails, and some said that they would continue using it for these reasons, despite the crashes.

S1 *It is a very quick program at doing [thumbnails], that's what I like about Shoebox, compared to these others. They tend to take a long time painting them on the screen. [...] And I will carry on using it because it's a very very easy program to use, very fast. I haven't found anything similar to that, any programs around. I know there are some, but not as quick. [...] The bugs don't bother me, I know it's beta release software, and that's what you expect.*

R3 *The image browsing interface, the thumbnail-y interface, is better than the software that came with the camera, and other bits of software I've tried. [...] It's one that's thought about volume, most of them haven't really considered volume, particularly.*

When asked if there were any extra features that they would like Shoebox to have, three participants said that perhaps it already had too many, suggesting that they would have preferred to use a basic system that was stable, rather than an unstable system with lots of features.

R2 *I think I will keep using it, but I will probably use a fairly limited feature set within it. Mostly just to organise it into rolls, mainly to use the slide show and publish as HTML, I find that I like that. But I probably won't use any of the other features. Probably because I don't really need them, and the bugs have made me less enthusiastic.*

A number of changes to Shoebox were suggested. Firstly, there were comments about the ordering of photographs within a roll: at present there are a fixed number of options (by insertion order, file date, or title). A "custom" option would be a useful addition for those participants who would like to be able to manually re-order the photographs, as they can with ordinary photograph albums. Also, as discussed earlier, if the photograph files contain metadata giving the date and time of exposure, this should be used in preference to the operating system time stamp when chronologically ordering the photographs.

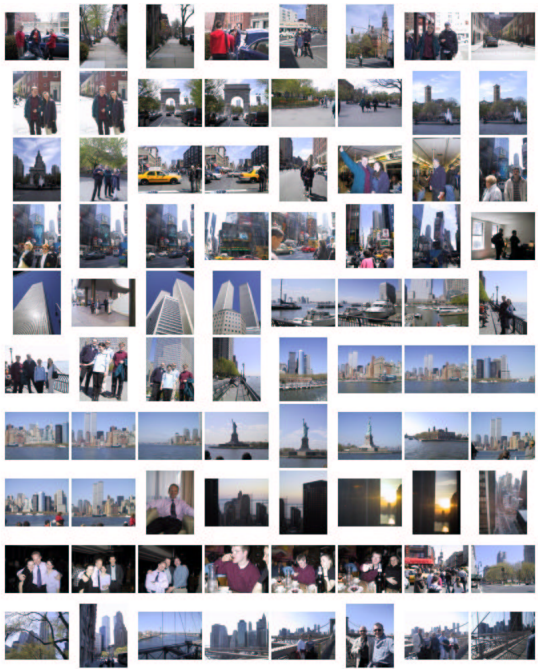
There were also some suggestions for improvements to Shoebox's annotation facilities. For example, a single annotation can be added to a whole group of photographs in one go, but if the user then wishes to delete the annotation, she has to delete it individually from each of the photographs in the group. Also, at present it is quite awkward to annotate a series of photographs one after the other, as the same sequence of mouse clicks has to be repeated for each photograph. It would be more efficient if users could select the whole series of photographs that they wish to annotate (as in StoryTrack [5]), so that the system can always move automatically to the next one in the series, and prompt the user for an annotation. The participants would also like to have more control over the inclusion or exclusion of annotations, for example being able to specify whether annotations should be exported into HTML, and whether spoken annotations should be played back with a slide show. Another helpful form of annotation would be the ability to mark photographs which the user feels are particularly good or bad, and then specify, for example, that the bad photographs should be excluded when viewing slide shows or publishing as HTML.

MDS arrangements

The developers of Shoebox had initially intended to include a facility to create arrangements of photographs using multidimensional scaling, but this was omitted due to time constraints on the project. We therefore asked the participants if they were willing to make a set of about 80 of their photographs available, to allow MDS arrangements to be demonstrated to them at their second interview. With the intention of simulating the typical content of a camera's flash memory, the only condition was that the photographs should be a continuous sequence, not necessarily all taken at the same event.

All of the participants listed in Table 6.3 contributed a set of photographs, except S2 and A3. We created a 10×10 proximity grid arrangement[3] of each participant's set (based on visual similarity, using the IRIS measure described

[3] using a version of MDS based on a genetic algorithm, described in Appendix A.



(a) Chronological order.

(b) 10×10 proximity grid, arranged according to visual similarity.

Figure 6.7: A set of 80 photographs from the joint collection of R1 and A1, in the two different arrangements in which it was shown to them.

in Section 3.1.4), as well as a chronological grid arrangement of the same photographs; examples of both are given in Figure 6.7. We did not consider a text-based arrangement, because so few of the participants had created annotations; a visual arrangement can be used regardless of the availability of annotations. Participants viewed the arrangements using the experiment software described in the previous chapter. If there was time available at the end of the interview, they were also shown the continuous version of the MDS arrangement[4], and were asked for their comments on that. Of the participants who saw the grid arrangement, only R3 and S1 did not see the continuous version.

Initially, most of the participants said that they found the visual arrangement interesting, or that it seemed to have done a good job of clustering the photographs.

R1 [describing the arrangement in Figure 6.7(b)] *It looks like it's done a pukka job, it's obviously got people, and it's got some dark backgrounds up here, and these are definitely all very similar shots, with sea, and river, and skyline and stuff, and then it's a bit park-y down here, and there's a couple of taxis that are quite close to each other, so I think it's done a good job.*

R4 *These two, I find this very interesting, these two together, I can make up a story using this. And these two cats have the same posture. It's fascinating, actually. Somehow it connects in different manners. Or maybe I make up these connections in my head!*

The comments of the participants in the initial study suggested that MDS might be a better way of searching by visual similarity than issuing a visual query. Three Shoebox trial participants explicitly said that they were more impressed with the MDS arrangements than they had been with visual queries in Shoebox. The results of a visual query are presented in a one-dimensional ranked list, which may contain idiosyncrasies. These are tempered in MDS configurations because they are generated based on the similarity of each image to every other image in the set, allowing clusters to form.

R6 *That's quite interesting. The similarities are much more obvious when you see them presented like that, actually. When you say "find similar", it's not that obvious, but yes, you can see a progression, feels to me almost diagonally across the screen. [...] It's actually done quite a good job, I much prefer it presented like that than individually, you do see the groupings much more obviously. [...] That's definitely a much better presentation of the same sort of data than Shoebox is giving.*

D1 *I believe this, whereas if I click on one of these images and get back a ranked list, and something appears wrong... there's no sense of "wrongness" in here, whereas with "find similar images" there always is.*

[4] created using the Newton-Raphson method, also described in Appendix A.

As with visual queries, however, it was difficult for the participants to imagine realistic requirements that an arrangement based on visual similarity could be used to satisfy. One participant said he could see himself using it to look for "sunsets, that sort of thing… if you wanted to find something for your backdrop" (R6).

D1 *I like the fact that it keeps similar things together, but whether it would actually help me to find something… I think it's doing a great job of ordering them visually, but it depends what the task is. If I'm looking for "blue sky, green grass" pictures, this is better, because I don't have to scan through linearly. But for organising my photos, I don't think I ever will want to find "blue sky, green grass" pictures. […] I think it's quite a fun thing to have, but it's probably not as useful as date and time.*

As was noted in the initial study, photographs that are taken at the same time are often visually similar, but they can be grouped together more reliably by simply using their time stamps, something which may be apparent from comparing the two arrangements in Figure 6.7. Visually similar photographs from different events may not have any meaningful relationship to each other.

R5 *Probably most of the ones that are similar would always be pictures of similar things, they are already in the same order. This sort of arrangement, pictures from very different things in the same place: interesting that they are similar, but I don't think provides any more insight.*

A number of participants said that they could imagine an MDS facility being useful for arranging a much bigger set of photographs, rather than the single "roll" that was used here, especially when their collections became very large. For example, photographs of the same generic type from the whole collection might be grouped together.

R7 *The fact that you've taken them all from the same group of photos means that the ones that it decides are visually similar tend to be ones that were taken at the same time, so that patch, that was at the same time, and all these ones, I guess they're pretty much the same time, so it would be more interesting to see that from diverse collections, to pick ones that are similar from totally different times.*

R8 *I think that would be interesting, because I've got a lot of photos like these ones, with the coastline and the sea, but we also took a lot of photos that would kind of fit into this corner, which is like trees from the Botanical Gardens, and stuff like that, so in terms of finding very specific things like that, that would be really useful.*

R1 *This is essentially a roll, but if you took all the photos in your collection, which is kind of what we were doing when we were trying to do the visual similarity searches… I think it would be more useful for when you're looking for a particular photo, maybe if you've got no idea when it was taken.*

Seeing a photograph out of its original chronological context may be "disorientating" (R8), because knowing which photographs were taken before and

after it is helpful when trying to work out what it depicts. For example, the set in Figure 6.7(a) contains photographs of visits to both New York and Boston, separated by time, but in Figure 6.7(b) the cities are mixed up together.

S1 *I actually preferred them in order. I think because of the time-scale... that was the Monday, and that was the Friday, on the way home. It could be useful, but I like to keep mine in order. That's confusing me now. It's not an order that I know them to be.*

R3 *It's sort of lost the context of the group of photos. With the [chronological] one, even though there's no annotations at all, and it's taken over a six month period, I can figure out, just by looking at a couple of photos in the sequence, what the whole set were. [...] Like that one, that's [my wife] and [son] with some foliage behind them, and if I just saw that on its own, I wouldn't have a clue where it was, it would need to be annotated, but because it's after a picture of the two French friends, and before the picture of [my son] on the Normandy beach, then I know that this is the botanic gardens in Cherbourg. Whereas if you go back to here [the visual arrangement], if you find that one, it's just lost.*

Depending on the set of photographs used, sometimes there was no obvious pattern to the visual arrangement, making it seem almost random. In others, this effect was only noticeable in the central area.

R8 *I'm really confused by the bit in the middle here, that I find a bit strange, because it's kind of all the photos jumbled in together. I noticed that, [in] the corners, I could kind of work out what was going on.*

Of the eleven participants who saw both the continuous and grid versions of the visual arrangement, seven said they preferred the grid version, two preferred the continuous version, and the remaining two had no preference. The continuous arrangement, although more accurate in showing the similarities between photographs, was described as "cluttered" (both R2 and R6), "a bit messy" (D1), and "bizarre" (R6). One participant (D2) said that the continuous arrangement seemed to contain more photographs than the grid arrangement. Some found the overlap "annoying" (D2), but others said that because the images that were on top of each other tended to be taken at about the same time, the overlap was helpful because clusters were more strongly emphasised than in the grid arrangement. It was also noted that a facility could be provided to bring overlapped images to the front.

R4 *This one seems to be in all directions, and it's got more dimensions than the other one, because you can somehow follow things, you can say maybe the more vivid colours are here in this corner and the dull ones are over here, and you can say, somehow, they tend to connect to each other from other directions than the ones that we had before. [...] I think I like this better, because it looks as if there's a middle point, and the relevance is radiating in all directions.*

As in the initial study, three participants felt that it might be useful if they wanted to create an unusual presentation of a related set of their photographs; two described it as "artistic" (R6 and A2).

A2 *I don't care in what order they are if I'm just sending them as something pretty, that you could paste them all together, like the best ones out of your holiday or something, and then get them all on one sheet, that'd be quite nice, like a collage or something.*

6.3 Discussion and conclusions

The most important use for personal photographs is the same whether they are digital or non-digital: looking at the most recent ones and showing them to friends and family. Browsing is mostly undirected, rather than directed. When people do search for older photographs from the collection, there are three types of requirement, listed here in decreasing order of frequency:

1. the set of photographs from a particular event, such as a holiday
2. an individual remembered photograph
3. a set of photographs taken at different events, but all matching a specification, such as containing a certain person

With non-digital photographs, all of these requirements are easier to satisfy when a collection is well organised (usually in albums), rather than disorganised. However, organisation involves effort, and people may take a long time to get around to doing it, sometimes simply leaving the prints in the original packets. The main motivation for organising prints is not that it facilitates searching, but that it results in an attractive presentation of the best photographs, for showing to other people, and then keeping as part of a family archive. Finding the photographs of a particular event, the most common requirement, is relatively easy regardless of whether the collection has been intentionally organised, as long as they have at least been kept in the same packet. Finding an individual photograph is also fairly easy, because this involves remembering the event at which it was taken, and then looking through the photographs from that event. It is helpful if the photographs have been kept in chronological order within the album or packet.

Digital photographs are easy to divide into named folders, such as Shoebox's rolls, so that finding photographs taken at a particular event is then simply a case of remembering which folder they are in, and clicking on it. Image management systems like Shoebox can reduce the photographs in a folder to thumbnail size, enabling users to look through a large number of them very quickly, making it even easier than before to find an individual photograph. In addition, digital photographs are stamped with the date and time of exposure, and therefore can be automatically presented in chronological order. With respect to the two most common types of requirement, then, intentional organisation is probably even less important for a digital photograph collection than a non-digital collection.

However, satisfying the third type of requirement, finding a group of photographs matching a specification, is a tedious task for both non-digital and

digital collections. This is because they are organised by event, a classification scheme that does not suit requirements such as finding photographs containing a particular person. The process usually involves repeatedly trying to remember a picture that matches the specification, and then looking for it. If *all* of the relevant pictures need to be found, the whole collection will have to be searched. Again, this task is probably easier with digital photographs, because a large number of thumbnails can be assessed at once. As we discussed earlier, the easiest way of satisfying this type of requirement is to specify it using a query, which the system can automatically compare to all of the items in the collection. Any text assigned to digital photographs can be indexed, allowing the user to construct queries using words from the annotations, something which cannot be done with handwritten notes on the back of prints. As a digital photograph collection gets bigger, queries should become more valuable, but the user will also have more photographs to annotate.

The results of the initial study suggested that people would find it useful to be able to annotate their digital photographs, and that having text queries available would make them more likely to create annotations. However, the results of the Shoebox trial seem to indicate that this was over-optimistic, and that although people like having the facility available, they do not often use it. As with non-digital photographs, people's main motivation for annotating their digital photographs is to record for posterity who and what is depicted in them, or perhaps to tell a story to family or friends elsewhere, and the fact that they enable text-based queries in a system like Shoebox is simply a useful side-effect. The ability to issue queries does not make people more likely to annotate their photographs, because the type of requirement that can be satisfied with a query is very rare.

When titles or free text annotations are created, Shoebox's topic view can automatically extract keywords from them, and present a list to the user. These keywords then become virtual categories, so that when the user selects one, the photographs whose annotations contain that keyword are displayed. This is just like typing in the keyword as a query, but to the user it is a simple browsing action. Each new keyword attributed to a photograph places it in an additional virtual category. However, it is difficult for people to be comprehensive in their annotations, and so it is unlikely that all of the relevant photographs will actually be retrieved when a keyword is selected from the topic view (or entered as a query). For example, the user may not always remember to name everyone who is present in a photograph, and may use names inconsistently: David Smith might be referred to as "Dave", "David", "Dave Smith", "Dave S", and so on. Like most information retrieval systems, Shoebox only indexes words, not entities, so a subsequent query for "David Smith" will not return any of the photographs where he is named as "Dave". This is even without taking into account potential errors in transcription of spoken annotations, where names of people and places are the words most likely to be mis-recognised.

A logical next step for a system like Shoebox would therefore be to allow the user to explicitly define her own set of attributes, such as names of people or places, and then assign them to the photographs to which they apply.

This would be like the controlled vocabularies typically used in annotating commercial image collections, as discussed in Section 2.3.2. The FotoFile system [62] adopts this approach: the user can select any number of attributes and assign them to one or more photographs at the same time, a process called "bulk annotation". PhotoFinder [109] allows the user to maintain a list of people, and then drag and drop a label containing a name onto a photograph, to indicate the presence of that person. Labels can be placed on the photograph so that each individual is identified directly, rather than having their relative positions described in a caption (for example, naming them from left to right). In both of these systems, retrieving all of the photographs containing a particular person is then simply a case of selecting his or her name from the list, like Shoebox's topic view. Names or other attributes could also be combined to produce the union or intersection of these virtual categories. With such a scheme in place, free text annotations could then be used primarily for telling stories, rather than faithfully recording who and what is present in each photograph.

Visual queries can be used for general requirements involving the visual properties of photographs, but in both the initial study and the Shoebox trial, the participants expressed little interest in constructing these. Current image processing techniques cannot extract enough semantics for people to find them useful. Further advances in image understanding may allow photographs to be automatically tagged with keywords indicating the presence of recognised objects, and if this could be done reliably, it would take some of the effort out of annotation. In particular, as mentioned by Kuchinsky and his colleagues [62], if the user could select a person's face, provide a name for them, and then have the system recognise and tag other photographs in which that person appears, this would probably be a very useful feature, integrating automation with the user's own annotations. It may also be possible to identify different generic types of photograph, for example indoor or outdoor (as discussed by Bradshaw [14], and Oliva and colleagues [84]), wide-angle or close-up, group shot or portrait, blurred or sharp, simple or complex.

The participants in the six month trial of Shoebox found that it is a powerful and fast piece of software which makes browsing easy. However, it crashed more often than was acceptable to many participants, and its more advanced features (speech recognition and visual queries) did not work as well as they had expected.

Finally, it seems that MDS arrangements based on visual similarity would be an interesting, but not essential addition to a system for managing personal photographs. It would perhaps be most useful to people for whom the aesthetic qualities of a photograph are important, and who wish to search their collection on this basis.

6.4 Further work

Both of these studies concentrated only on the needs of individuals, although they included members of the same household among the participants. It

would be interesting to carry out a further study to focus specifically on the collections and requirements of couples and families, as in many cases a photograph collection belongs to a household, not an individual. For example, different members of the family may take photographs at the same event, and then want to contribute annotations about each other's pictures, but still keep track of who took which photograph. In some cases the family member who was participating in the Shoebox trial was not the person who was normally responsible for organising the household's photographs, resulting in a change in practice. Shoebox does not explicitly take account of multiple users; its speech recognition engine, for example, can only be trained to the voice of a single annotator.

Long-term storage of digital photographs is a concern, and should be investigated further. Not having photographs in physical form may make them seem less tangible, more easily deleted or removed from existence. Some participants in the Shoebox trial, however, took the opposite view: that digital photographs can be backed up, whereas prints and negatives can easily be lost or damaged.

Finally, a full ethnographic study of how people use their photographs in the course of everyday life would be likely to yield very useful qualitative results; more quantitative studies, with larger, more representative samples, would provide a rigorous test of the findings reported here. Six months is a relatively short time in this application area, and future studies would ideally follow their participants over a longer period.

Chapter 7

Conclusions

This chapter summarises the main findings presented in this dissertation, and offers suggestions for further work.

7.1 Summary of main findings

Multidimensional scaling can be used to arrange a set of images according to their visual similarity, at the primitive level, such that those images which look alike are placed together. In our theoretical analysis in Chapter 3 we found that, in such arrangements, images of similar generic content were grouped much more closely than would be expected by chance. We compared a number of visual similarity measures, ranging from simple (based on average colour) to complex (based on image segmentation). Although there were significant differences between the measures when used for image retrieval, their corresponding MDS arrangements were almost equivalent, quantitatively and qualitatively. Generally, the simpler measures were lower-dimensional, and so reducing their configurations to two dimensions with MDS resulted in less error. We selected IRIS as the measure to be used in our later experiments, but it is likely that the others would have produced analogous results.

In Chapters 4 and 5, we described four experiments where the participants were asked to carry out directed browsing tasks, selecting an image or images from within a presented set of thumbnails. The conditions of these experiments are summarised in Table 7.1.

Both of the experiments in Chapter 4 had one condition where the set of thumbnails was arranged according to visual similarity, and one where it was arranged randomly. In the first experiment, we found that the participants were faster at locating a given target image in a visual arrangement (created with continuous MDS) than in a randomly arranged grid. The salience of the target image within the set, however, seemed to have more influence than the way in which the set was arranged. Even if the participants had not noticed that the images were arranged according to visual similarity, they were still able to take advantage of it. Many of the participants found that the overlap between thumbnails in a continuous MDS arrangement was annoying, and for all of our subsequent experiments we used arrangements created with prox-

Chapter	*Experiment*	#	*Random*	*Visual*	*Caption*
4	First	80	8×10	continuous	–
	Second	100	10×10	10×10	–
5	infodesign 99	100	–	12×12	12×12
	Anglia	100	10×10	10×10	–

Table 7.1: Summary of the four comparative experiments described in this dissertation. The column marked # indicates the number of images that were in each set. The *Random*, *Visual*, and *Caption* columns show which arrangement types were compared; normally proximity grid arrangements were used (and the grid size is given here), but the first experiment used a continuous MDS arrangement for the visual condition.

imity grid algorithms. Such arrangements have a regular structure, while retaining much of the shape of the continuous version (depending on the chosen grid size).

In the second experiment, the participants were asked to locate images matching a requirement based on generic content, such as "fruit" or "surfing". They were faster at finding relevant images in a visual arrangement (this time using a proximity grid) than in a random arrangement, but in contrast to the first experiment, they had to be aware of the organisation by visual similarity in order to take advantage of it. We found in our theoretical analysis in Chapter 3 that images with similar generic content are often placed close to each other in arrangements based on visual similarity. The results of the second experiment therefore showed that the participants could use this grouping to help them find the relevant images, especially if these had a predictable appearance.

In Chapter 5, we conducted two experiments and a follow-up study, all of which used a simulated work task involving picture selection, with designers as the participants. They were presented with a set of images of a particular place, and were asked to select three of them to accompany a given passage of text from a travel guide. All of the images in the set were already relevant to the given (specific-level) requirement, and it was left to the participants to develop and apply their own selection criteria. The images were drawn from a single category, whose contents would normally be displayed to the user in whichever order they are stored on the file system. The main Anglia experiment compared a visual arrangement of the set to a random arrangement; the designers were faster at making their selections from the random arrangement, but preferred the visual arrangement, and felt that they had made better choices when using it. Small differences in efficiency may not be important in practice, if the user is happier with the end result. We found individual differences in all of our experiments, and therefore decided to test the Anglia participants' spatial ability; it had no relationship to their performance or their preferences, however.

When a collection is annotated, an alternative way of measuring the similarity of a pair of images is to use their annotations, for example by counting

the words that they have in common. In the resulting MDS arrangements, the images are grouped according to their subject matter, depending on the level of detail present in the captions. In the infodesign 99 study and the follow-up to the Anglia experiment (both described in Chapter 5), the participants were able to compare caption-based arrangements to visual arrangements. They seemed to find both types useful when they had a requirement in mind, and wanted to identify the relevant subset within the presented set of images. In general, different views of the same set of images can support different selection criteria, and many of the designers liked having more than one arrangement available. Switching between arrangements can also be helpful because different images tend to stand out in each.

In Chapters 3–5 we used images drawn from a stock photograph library, but in Chapter 6 we considered personal photograph collections, which are familiar to their users. We found that arrangements based on visual similarity are less useful in this application area, because people simply do not normally want to look for their photographs using visual criteria. The most common requirement is for a set of photographs of a particular event, which can easily be found by browsing through thumbnails, with only the most rudimentary organisation (chronological ordering, and categorisation according to event). Queries are usually unnecessary, and are difficult to support because many people do not feel that annotating their photographs is worth the effort.

7.2 Recurring themes

In this section we discuss two themes that arose in more than one chapter: the structure and local contrast of each arrangement type, and continuous versus grid arrangements.

7.2.1 Structure and local contrast

Each of the arrangement types we have considered in this research can be classified according to their definition of similarity, the understandability of their structure, and the strength of their local contrast.

Visual arrangements group together images that are similar at the primitive level. This low-level grouping means that the structure of the arrangements is easy to perceive, making them helpful for directed browsing when the user's requirement is based on visual content. This is why the participants in the first experiment (Chapter 4) were generally able to find a given target image more quickly in a visual arrangement than in a random arrangement. As we found in Chapter 3, similarity at the primitive level has some correspondence with similarity at the generic level, and this meant that visual arrangements were helpful to the participants in the second experiment (Chapter 4), and also to the designers in the experiments in Chapter 5, when they had a generic requirement in mind. Their main drawback, however, is that the local contrast is reduced, making the individual images less salient because they are surrounded by

visually similar neighbours. This can hinder undirected browsing, as well as directed browsing when the user has already narrowed down her search to a certain area of the arrangement, and wants to pick out an individual image (as we found in the first experiment in Chapter 4).

Random arrangements typically provide a much stronger local contrast, because neighbouring images are not usually visually similar, making individual images more likely to stand out. Of course, random arrangements contain no grouping, and thus have no perceivable structure. In both of the experiments in Chapter 4, the difference in response times between visual and random arrangements was small, probably due to some combination of the greater local contrast in a random arrangement, and the fact that 80–100 images is a reasonably low number to have to search through. In the Anglia experiment in Chapter 5, the designers made their selections more quickly when using a random arrangement, probably because they could easily identify distinctive images.

Caption-based arrangements can group images at whichever levels of content are covered by the captions; in the Corel collection, this is usually the specific level. These arrangements also have good local contrast, provided that caption similarity does not have a close correspondence with visual similarity in the collection being used. This higher-level grouping, however, means that the structure is not immediately obvious to the viewer. If we repeated the first experiment in Chapter 4, but compared a visual arrangement to a caption-based arrangement, we would expect the visual arrangement to result in better performance because of the low-level nature of the task. It would be interesting to see if there would be any difference between a caption-based arrangement and a random arrangement for this task. Superimposing descriptive words (extracted from the captions) over related clusters of images helps to make the structure more explicit. Otherwise, the user has to discover the structure for herself by running the mouse pointer over the images to read their titles.

7.2.2 Continuous versus grid arrangements

Some comparison between continuous and grid arrangements was made in all of Chapters 4, 5, and 6. In the first experiment (Chapter 4) we discovered that the participants disliked the overlapping images in a continuous arrangement, and were also slower if they had to find a target image that was not fully visible; this prompted the development of the proximity grid algorithms. In the Anglia follow-up (Chapter 5) and the final Shoebox interview (Chapter 6), the participants were able to qualitatively compare continuous and grid versions of the same visual arrangement. The continuous arrangements were described as "messy" and "cluttered" by participants in both studies, and there was a general (although not unanimous) preference for the grid version.

Further experiments could compare the two types more formally. It would also be possible to consider the potential effect of having different grid sizes, and try to find the optimal density; this involves a trade-off between retaining

as much of the original structure as possible, and allowing thumbnail images to be as large as possible. For 100 images, a 12×12 grid seemed to be a good choice.

Continuous versions of caption-based arrangements are unlikely to be useful if the captions are short, because in some cases two captions will contain exactly the same terms, and thus the corresponding images will be placed almost directly on top of each other.

7.3 Future work

Possibilities for future work fall into two main categories: incorporating MDS arrangements into a real image browsing system, and carrying out further theoretical or experimental studies. Suggestions for the area of personal photography in particular are given separately, at the end of Chapter 6.

7.3.1 Similarity-based arrangements in practice

As we described in Section 2.4.3, a number of researchers have developed image browsing systems that incorporate visualisation features, and in this section we briefly outline some of the practical issues involved. If such a system could be used in a real setting, it would be possible to conduct a field study to investigate whether similarity-based arrangements are useful in practice.

There are four main stages in the construction of an MDS arrangement of a set of images:

- extract suitable features from the images (or their annotations)
- calculate a similarity matrix for a particular image set
- apply MDS to find 2D coordinates for each image in the set
- display thumbnails on the screen at those coordinates

To create the arrangements used in our experiments, these steps were all carried out separately, using different programs. This meant that our experiment software could only display arrangements whose content was predetermined. A real system, however, should be able to create arrangements of sets whose content is determined dynamically, such as query results.

The extraction of features from the images or their annotations would need to be done only once for the whole collection. These could be the same features that are used to index the images, facilitating visual and textual queries. Simple visual features like colour histograms take less time to extract than more complex features like regions. Terms can be extracted from the annotations, and perhaps weighted according to their frequency. Thumbnails could also be generated at this stage, if necessary.

When the user has selected a set of images to be arranged (perhaps by issuing a query), a similarity matrix for that set can be created, by using the features of the images to calculate their pairwise similarities, with whichever

measure has been chosen. Ideally, the measure would be quick to compute, to ensure that the arrangement can be created interactively. Alternatively, if enough storage is available, a similarity matrix for the entire collection could be precomputed, with parts of it extracted as necessary.

Coordinates would be found by sending the similarity matrix to the MDS algorithm, which would need to be incorporated into the system. The user could be given the the option of producing a continuous arrangement, a proximity grid created directly from the continuous arrangement, or a proximity grid created using a slower but more accurate method such as a genetic algorithm. The user could specify the grid size, although a sensible default would be the most dense grid possible for the number of images in the set. If the image set is likely to be used again, the coordinates could be stored once they have been computed.

Finally, the system would translate the output coordinates to current screen coordinates, placing thumbnail versions of the images at these points. A number of extra interface features could be implemented, to provide different views and adaptations of the display; our experiment software simply included an overview-plus-detail feature to allow the user to see a magnified version of part of the arrangement. The system created by Combs and Bederson [28] has an in-place zoom facility, which is a possible alternative. When viewing a continuous arrangement, the image currently under the mouse pointer could be moved to the front of any that are occluding it.

With caption-based arrangements, as we have already discussed, it would be helpful to superimpose keywords, allowing the user to toggle their display. It may also be useful to highlight any topic-related clusters, perhaps shading the background of the arrangements accordingly.

7.3.2 Further research

There are a number of possible theoretical or experimental studies that would extend the research described in this dissertation. Firstly, a similarity-based arrangement could be compared to a more meaningful alternative than a random arrangement; such a comparison could be done using any of the evaluation strategies employed in this dissertation.

Ranked lists Except for the personal photographs in Chapter 6, the image sets we have considered would not normally be presented in any particular order. The results of a query, however, are returned in a one-dimensional list, ranked in order of estimated relevance. In an annotated collection, MDS arrangements (both visual and caption-based) of the results of a textual query could be compared to the original ranked list. As we noted in Section 2.4.3, the ANVIL system [102] allows users to view the results of a text query in either a ranked list or a simple caption-based arrangement. It would also be possible to arrange the results of a visual query according to visual similarity, as in the system described by Rubner, Tomasi, and Guibas [103], and again, this could be compared to a

ranked list. The set may be too homogeneous, however, for a similarity-based arrangement to offer any real advantage.

Manually created arrangements Rogowitz and her colleagues [99] asked each of their participants to arrange a set of images on a table according to his or her own opinion of their similarity. It would be interesting to experimentally compare such arrangements to those created automatically, using either visual or text-based similarity. As we noted in Section 2.2, Lin [70] compared a self-organising map of a set of documents to a manually created arrangement, and found that there was no significant difference between them for the task of locating a given document.

We have concentrated primarily on arrangements constructed with visual similarity measures, and it would be interesting to carry out further evaluations of caption-based arrangements. The effectiveness of different caption-based similarity measures could be tested, as well as different methods of automatically generating labels. The Corel collection is unlikely to be adequate for such studies, however, because (as we noted in Chapter 5) its annotations lack sufficient detail.

Our four main experiments used sets of 80 or 100 images, assuming that the user had already carried out some restriction of the collection. It is likely that the organisation provided by a similarity-based arrangement would be even more helpful in a larger set of images, but of course this would mean smaller thumbnails (unless a proportionally larger screen was used), perhaps cancelling out any improvement. For example, the participants in the Anglia follow-up study in Chapter 5 felt overwhelmed when presented with arrangements containing 200 images, and said that the thumbnails were too small. This issue could be investigated further; a possible solution would be to allow the user to adjust the level of detail, as in the CIRCUS system [88] discussed in Section 2.4.3.

Our theoretical analysis of visual arrangements (Chapter 3) was limited to objectively measuring the degree of clustering of images with similar generic content. We assumed that the best arrangements were those which scored highest on this criterion, and did not attempt to quantify the understandability of an arrangement's structure, or the strength of its local contrast. It is likely that the structure of a visual arrangement will be less obvious if the images are of a lower technical quality than those in the Corel collection, or if the chosen set is visually homogeneous. For example, in Chapter 6 we noted that some visual arrangements of personal photographs had very little perceivable structure, making them seem almost random. It would therefore be interesting to attempt to predict, given a set of images and a visual similarity measure, whether they would result in an arrangement with understandable structure. The simplest visual similarity measures (such as those based on average colour) might produce the best results in this case. As we discussed in Section 4.1.5, it would also be possible to try to model the user's perception of an arrangement of thumbnails, in order to measure its local contrast, and predict the salience of the individual images.

In Chapter 5, we chose a simulated work task that involved selecting three images of a certain place to accompany a given passage of text. Of course, there are many other possible tasks: for example, the initial requirement could be for generic or abstract content, instead of a place name. The participants in the Anglia experiment (and its follow-up) said that they found a visual arrangement useful mainly because it tended to place images of the same generic type together. It would be interesting to investigate whether a visual arrangement would still be useful if the presented images were all of the same generic type; this would allow primitive-level similarity to be isolated from generic-level similarity.

Finally, throughout this dissertation we have concentrated on general photographic images, as used in the domains of graphic design, publishing, advertising, and home entertainment. The findings may apply to other domains, and other types of images, but this would need to be independently tested.

Appendix A

MDS algorithms

This appendix describes the multidimensional scaling (MDS) algorithms that were used in the course of this work. These were all written by Wojciech Basalaj, and are described in more detail in his PhD dissertation [7].

MDS is a technique that treats inter-object dissimilarities as distances in some high dimensional space, and then attempts to approximate them in a low dimensional (commonly 2D or 3D) output configuration. Basalaj's algorithms are based upon least squares metric MDS: they all attempt to minimise a **loss function** that measures the amount of error in the low dimensional representation of the dissimilarities. Each algorithm uses a different minimisation strategy in order to achieve this. A common loss function is **energy**, which is defined as follows:

$$E = \sum_{r<s} \frac{\left(d_{rs} - \hat{d}_{rs}\right)^2}{\hat{d}_{rs}^2}$$

where d_{rs} is the Euclidean distance between points p_r and p_s in the low dimensional arrangement, and $\hat{d}_{rs}$ is the (original) dissimilarity between objects r and s. The reported energy value is an average across all of the pairwise dissimilarities, that is, $E/\frac{n(n-1)}{2}$, where n is the number of objects in the configuration.

Because the algorithms have a randomly determined starting point, they tend to produce a slightly different solution (corresponding to a local minimum of the loss function) each time they are run. We therefore ran our chosen algorithm a fixed number of times (usually five or ten), and then selected the best configuration produced from these runs (that is, the one with the lowest value of the loss function).

The different algorithms are suitable for use in different circumstances. Not all of them were available for every experiment, because Basalaj's research was carried out in parallel with the work described in this dissertation.

There are two classes of algorithm, which can be distinguished by the nature of the output they produce.

A.1 Continuous

These algorithms (discussed in Chapter 3 of Basalaj's dissertation) produce a final configuration where the objects may be placed anywhere in a continuous space; the resulting coordinates can then be appropriately scaled for display on a screen. The first algorithm uses **Newton-Raphson iteration** to minimise the loss function, and we employed it to produce continuous arrangements for our first experiment, the Anglia follow-up, and the Shoebox study.

The second algorithm was developed later, and uses **simulated annealing**. In Basalaj's experiments, it proved to be more effective than the Newton-Raphson method, and was also slightly faster. We had originally used the Newton-Raphson method for the comparison of similarity measures in Chapter 3, but we subsequently repeated the analysis using the algorithm based on simulated annealing, and chose to report those results. There was very little difference between the two versions.

A.2 Proximity grid

In our first experiment, the participants expressed a dislike of the image overlap in the continuous MDS arrangements, and as a result, Basalaj developed this second group of algorithms, discussed in Chapter 5 of his dissertation. They produce a final configuration where the coordinates of the objects must lie on a square grid of a specified size. The simplest way of doing this is to take the output of a continuous MDS algorithm (we used the Newton-Raphson method) and snap it to a grid of the desired size, using a **greedy algorithm** (called *Greedy1* by Basalaj). This is fast, but only gives an approximate solution; it was used in our second experiment, and the infodesign 99 study.

Alternatively, an MDS algorithm can be adapted to work in the discrete rather than the continuous domain, with the prior knowledge that a grid configuration is desired. Basalaj chose to use a **genetic algorithm** for this purpose. He showed that although this was somewhat slower than the greedy approach, it produced proximity grid configurations with much lower values of the loss function. We used it in the Anglia experiment, and to create the grid arrangements shown to participants in the Shoebox study.

Appendix B

Caption-based similarity

This appendix describes the similarity measure that we used to create the caption-based arrangements for the experiments in Chapter 5.

The annotations in the *Corel Stock Photo Library* are not as comprehensive as those of most large commercial collections, containing only a caption and four general keywords. For example, Figure B.1 shows image 244093 from category 244000, "New York City". Its associated annotation is:

```
Liberty Island with sunset
;sunset;island;statue;people;
```

The keywords are applied very inconsistently (for example, sometimes images which contain sky have "sky" as a keyword, and sometimes they do not) and we decided to measure similarity based on the captions alone.

The simplest way to measure the similarity of two captions is to count the terms they have in common. For example, say that we wanted to measure the similarity between image 244093 (above) and image 244003, "Statue of Liberty". If X is the set of terms in image 244093's caption, and Y is the set of terms in image 244003's caption, their similarity would be $|X \cap Y|$, the number of terms in the intersection of the two sets. In this case, the only term the two captions have in common is "Liberty", and therefore $|X \cap Y| = 1$.

Figure B.1: Image 244093 from the Corel collection, *Liberty Island with sunset*.

After some experimentation, we decided to remove any *stopwords* (very common words, such as articles, pronouns, and prepositions) from the captions, with the remaining words being reduced to their *stems* via suffix stripping, using Porter's algorithm [92].

In information retrieval's vector model, each document in a collection is represented by a vector with n entries, where n is the number of unique terms in the collection as a whole; in this case, we treated each image set as a collection in its own right. Imagine for simplicity, however, that 244093 and 244003 are the only images in the set. Once the stop words "with" and "of" are removed, their captions contain a total of four unique terms: "liberty", "island", "sunset", and "statue". The vectors for each image would be as follows, with 1 indicating the presence of a term, and 0 the absence:

	liberty	island	sunset	statue
244093	1	1	1	0
244003	1	0	0	1

Our simple similarity measure can thus be thought of as the dot product of the term vectors. However, in its current form it is unsatisfactory, because it does not take into account the fact that the captions may contain different numbers of terms. In the vector model, the term vectors can be normalised, by dividing each component by the length of the vector. The similarity measure is then the dot product of the normalised term vectors. This is equivalent to reformulating the original similarity measure as follows:

$$similarity(X, Y) = \frac{|X \cap Y|}{\sqrt{|X|} \times \sqrt{|Y|}}$$

This is known as the *cosine coefficient* measure. In our example, because $|X| = 3$ and $|Y| = 2$, the similarity of the two images is $1/(\sqrt{3} \times \sqrt{2}) = 0.408$. The similarity will now always be between 0 and 1. In practice, the *dissimilarity* is used, which is found by simply subtracting the similarity from 1, so in this case it would be 0.592.

These methods are described in classic information retrieval textbooks [104, 118]. Because the captions are very short, we chose to use binary term weighting (as in the above example), and therefore the measure does not take into account the frequency with which words occur, either in an individual caption or in the collection as a whole.

Figure 5.1(a) on page 95 shows a caption-based arrangement of category 244000, in a 12×12 proximity grid. Image 244093 is circled, and appears immediately below 244003.

Appendix C

The Corel Stock Photo Library

This appendix lists the categories in the two volumes of the *Corel Stock Photo Library* that were used in the experiments described in this dissertation. Each volume contains 20,000 images, in 200 categories of 100. The photographs were supplied in Kodak Photo CD format, but were exported to a file system as 768×512 JPEG images, as well as 96×64 GIF thumbnails.

At the time of writing, the library is no longer available for purchase, and the rights to it have been transferred from Corel to Hemera. It can still be browsed via
`http://elib.cs.berkeley.edu/photos/corel/`

Corel Stock Photo Library 1

172000 Action Sailing
77000 African Antelopes
130000 African Specialty Animals
10000 Air Shows
44000 Alaska
173000 Alaskan Wildlife
38000 American National Parks
132000 Annuals For American Gardens
49000 Apes
113000 Arabian Horses
19000 Arizona Desert
69000 Australia
126000 Austria
21000 Auto Racing
150000 Autumn
34000 Aviation Photography
123000 Backyard Wildlife
135000 Bald Eagles
83000 Bali, Indonesia
97000 Barns & Farms
66000 Barnyard Animals
100000 Bears
141000 Beneath The Caribbean
8000 Birds
8100 Birds Volume II
231000 Bonny Scotland
93000 Brazil
22000 Bridges
52000 Butterflies
51000 Cactus Flowers
144000 California Coast
50000 California Parks
138000 Canada's East Coast
96000 Candy Backgrounds
155000 Canoeing & Kayaking
68000 Caribbean
194000 Caves
134000 Cheetahs, Leopards, & Jaguars
85000 China
15000 China & Tibet
72000 Christmas
24000 Churches
111000 Cities Of Italy
5000 Coasts
125000 Coins & Currency
63000 Commercial Construction
136000 Cougars
154000 Croatia
88000 Czech Republic
39000 Death Valley
121000 Denmark
32000 Deserts
156000 Divers & Diving
247000 Dogs
59000 Doors Of San Francisco
30000 Egypt
107000 Elephants
160000 Endangered Species
131000 English Country Gardens
29000 Exotic Cars
193000 Fantasy Backgrounds
28000 Fields
37000 Fighter Jets
101000 Fiji
78000 Finland
73000 Firework Photography
40000 Fireworks
185000 Fishing
124000 Flowering Potted Plants
13000 Flowers
13100 Flowers Volume II
25000 Food
109000 Foxes & Coyotes
116000 France
91000 Fruits & Vegetables
129000 Germany
114000 Glaciers & Mountains
162000 Grand Canyon
157000 Grapes & Wine
98000 Great Silk Road
67000 Greece
118000 Greek Isles
58000 Guatemala
46000 Hawaii
70000 Hawks & Falcons
179000 Helicopters
140000 Holland
120000 Hong Kong
197000 Horses
171000 Ice & Frost
184000 Ice & Icebergs
143000 Images Of Death Valley
102000 Images Of France
147000 Images Of Thailand
177000 Images Of The Grand Canyon
166000 Images Of Turkey
71000 India
189000 Indigenous People
35000 Insects
35100 Insects Volume II
17000 Ireland
122000 Israel
79000 Italy
65000 Japan
65100 Japan Volume II
55000 Korea
145000 Kyoto
26000 Lakes & Rivers
161000 Land Of The Pyramids
176000 Landscapes
105000 Lions
119000 Los Angeles
33000 Mayan & Aztec Ruins
90000 Meso America
64000 Mexico
53000 Mexico City
180000 Military Vehicles
181000 Models
146000 Monaco
2000 Mountains Of America
158000 Mushrooms
89000 Native American Ruins
178000 Nature Scenes
163000 Nesting Birds
54000 New Mexico
148000 New York City
36000 New Zealand
36100 New Zealand Volume II
104000 North American Deer
127000 North American Wildflowers

41000 North American Wildlife
48000 Northern California
164000 Ocean Life
211000 Oil Paintings
57000 Old Singapore
115000 Orchids Of The World
86000 Oregon
242000 Ottawa
75000 Owls
117000 Pacific Coasts
11000 Patterns
23000 People
239000 People Of The World
133000 Perennials In Bloom
142000 Peru
183000 Polar Bears
224000 Portrait Of Italy
42000 Predators
80000 Rajasthan, India
99000 Religious Stained Glass
87000 Reptiles & Amphibians
31000 Residential Interiors
112000 Rhinos & Hippos
149000 Rome
84000 Roses
16000 Rural Africa
174000 Rural France
128000 Russia, Georgia, & Armenia
56000 Sacred Places
7000 Sailboats
82000 San Francisco
196000 Sands & Solitude
92000 Scotland
61000 Ski Scenes
60000 Skiing In Switzerland
175000 Snakes, Lizards & Salamanders
170000 Soldiers
76000 Southeast Asia
20000 Spirit Of Buddha
1000 Sunrises & Sunsets
153000 Swimming Canada
137000 Textures
81000 Thailand
14000 The Arctic
74000 The Big Apple
139000 The Galapagos
108000 Tigers
9000 Trees & Leaves
152000 Tropical Plants
258000 Tulips
47000 Turkey
47100 Turkey Volume II
45000 Underwater Life
12000 Underwater Reefs
169000 Vegetables
95000 Wales
43000 Waterfowl
27000 Waterfalls
18000 Western Canada
168000 Wild Sheep & Goats
6000 Wild Animals
159000 Wildlife Babies
106000 Wildlife Of Antarctica
103000 Wildlife Of The Galapagos
62000 Windsurfing
151000 Winter
110000 Wolves
3000 World War II Planes
94000 Yellowstone National Park
167000 Yosemite

Corel Stock Photo Library 2

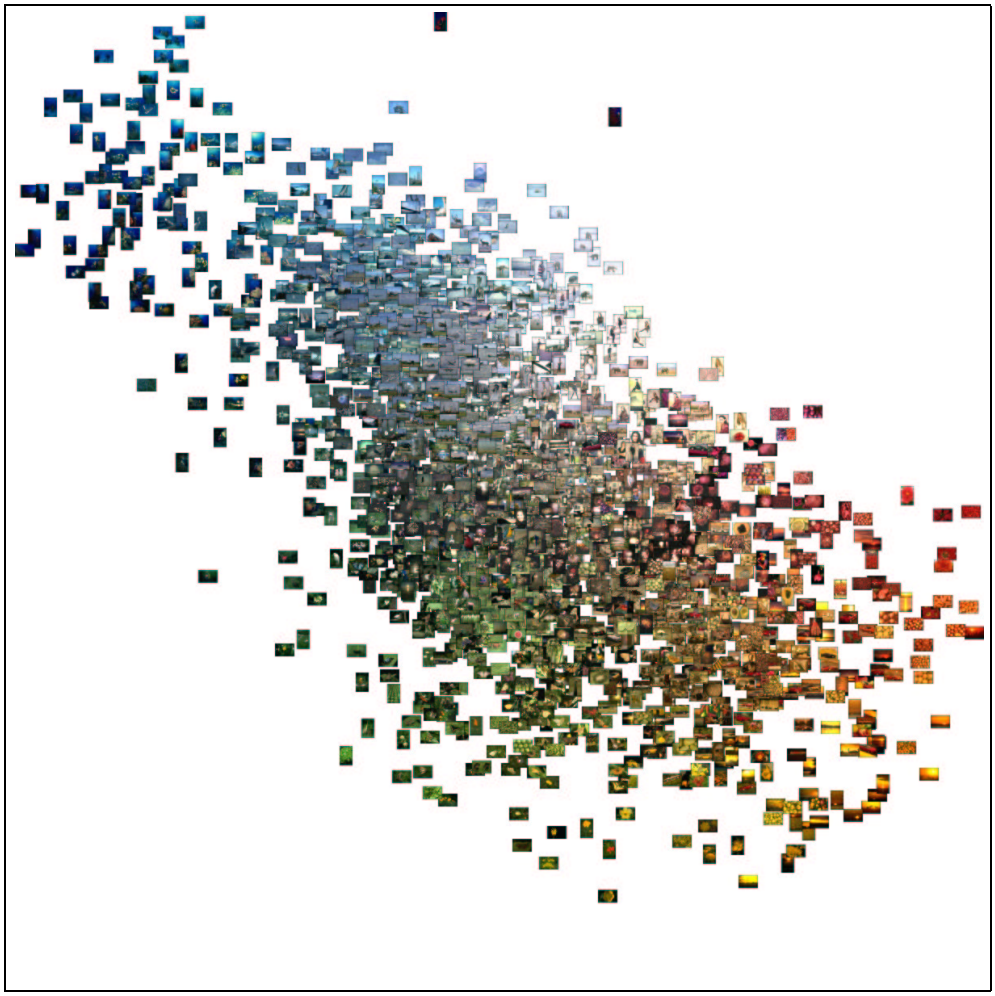
243000 Acadian Nova Scotia
268000 African Birds
396000 Air Force
226000 Amateur Sports
263000 Antique Postcards
240000 Arthropods
327000 Artist Textures
304000 Asian Wildlife
228000 Autumn In Maine
359000 Aviation Photography 2
294000 Barbecue And Salads
399000 Bark Textures
384000 Beaches
202000 Beautiful Bali
198000 Beautiful Women
369000 Belgium and Luxembourg
292000 Berlin
275000 Beverages
246000 Bhutan
412000 Bobsledding
353000 Bonsai And Penjing
394000 Botanical Prints
354000 British Motor Collection
230000 Canada
371000 Canada, An Aerial View
251000 Canadian Farming
344000 Canadian National Parks
276000 Canadian Rockies
238000 Canoeing Adventure
325000 Car Racing
382000 Castles
336000 Cats And Kittens
218000 Caverns
212000 Chicago
188000 Classic Antarctica
360000 Classic Aviation
191000 Clouds
219000 Coast Of Norway
255000 Colorado Plateau
403000 Colors And Textures
341000 Colors Of Autumn
237000 Construction
372000 Copenhagen, Denmark
220000 Cowboys
186000 Creative Crystals
367000 Creative Textures
333000 Cuisine
234000 Decorated Pumpkins
297000 Desserts
291000 Devon, England
310000 Dogsledding
314000 Dolphins And Whales
298000 English Countryside
340000 English Pub Signs
373000 Everyday Objects
279000 Exotic Hong Kong
281000 Exploring France
332000 Fabulous Fruit
302000 Fashion
285000 Fire Fighting
209000 Fish
282000 Fitness
221000 Flowers Close-Up
318000 Foliage Backgrounds
225000 Freestyle Skiing
350000 Frost Textures
322000 Fruits And Nuts
208000 Fungi
346000 Garden Ornaments And Architecture
214000 Gardens Of Europe
227000 Greek Scenery
363000 Highway and Street Signs
267000 Hiking
261000 Historic Virginia
308000 Holiday Sheet Music
378000 Horses in Action
329000 Hot Air Balloons
195000 Hunting
299000 Images Of Egypt
301000 In Shakespeare's Country
250000 Industry And Transportation
190000 Interior Design
303000 International Fireworks
385000 Ireland 2
328000 Jamaica
217000 Java
286000 Jersey Channel Islands
253000 Kenya
364000 Kitchens and Bathrooms
241000 Lake District, England
405000 Landscape Backgrounds
295000 Landscapes Of The World
368000 London, England
321000 Lost Civilizations
330000 Lost Tribes
349000 Marble Textures
280000 Martial Arts
204000 Mediterranean Cruise
264000 Mexican Holiday
260000 Middle East
245000 Military Aircraft
277000 Montreal
229000 Morocco
213000 Mountains Of Eurasia
335000 Namibia
366000 Nature's Textures
270000 Navy Seals
187000 Nepal
272000 New Guinea
244000 New York, NY
287000 Night Scenes
365000 North American People
347000 Northern Wilderness
293000 Northwest Africa
265000 Painted Textures
205000 Paris
376000 People 2
273000 Performance Cars
223000 Picturesque Paris
361000 Polo
309000 Portrait Of Alaska
256000 Portugal's Countryside
248000 Prague
393000 Prince Edward Island
232000 Quebec
315000 Rafting
313000 Rainy Nights
206000 Recreational Sports
252000 Reflections
343000 Reflective Effects

342000 Rocks And Gems
331000 Rodeo
374000 Rural England
274000 Russia
262000 Rustic Quebec
339000 Sailboarding
338000 Sailing
991000 Sampler BK1
390000 Sand and Pebble Textures
222000 Scenic Austria
249000 Scenic Japan
199000 Scenics
406000 Sculpted Light
305000 Sheet Music Cover Girls
355000 Shell Textures
397000 Sierra Nevada Mountains
271000 Sights Of Africa
233000 Solitude
323000 Space
283000 Space Scenes
377000 Spectacular Waterfalls
348000 Speedo Swimsuits
362000 Spice and Herb Textures
324000 Sports And Leisure
236000 Spring
182000 Steam Trains
269000 Studio Models
345000 Sunsets Around The World
300000 Surfing
201000 Sweden
257000 Tall Ships
192000 Textile Patterns
356000 Textures By James Dawson
404000 Textures 2
207000 The Alps In Spring
311000 The Everglades
357000 The Masters I
358000 The Masters II
401000 The Masters 3
402000 The Masters 4
235000 The Netherlands
216000 The Oregon Trail
215000 The Romance Of France
375000 Theater
389000 Tools
383000 Tour Through Europe
334000 Traditional Japan
351000 Trains
391000 Trains Of The World
210000 Tropical Sea Life
306000 Under The Red Sea
307000 Underwater Photography
254000 Utah, Color Country
316000 Valley Of Fire
320000 Victorian Houses
203000 Virgin Islands
289000 Wading Birds
266000 Washington State
259000 Washington D.C.
370000 Water Sports
312000 Waves
337000 Weddings
317000 Whitetail Deer
326000 Wildcats
319000 Wildlife Paintings
388000 Women In Vogue
288000 World Landmarks
290000 Yemen
296000 Zimbabwe
392000 Zion National Park

Appendix D

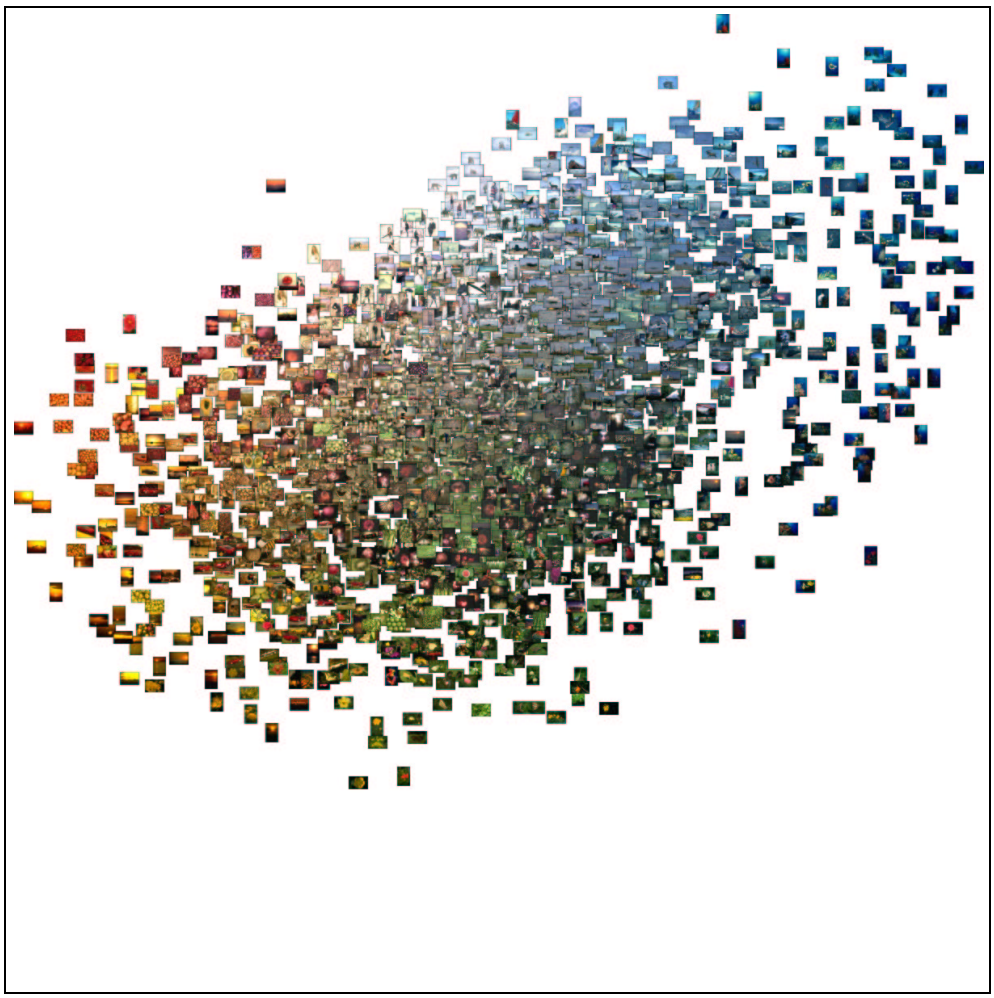
Materials

This appendix contains the instructions, questionnaires, and other materials used in the experiments described in this dissertation.



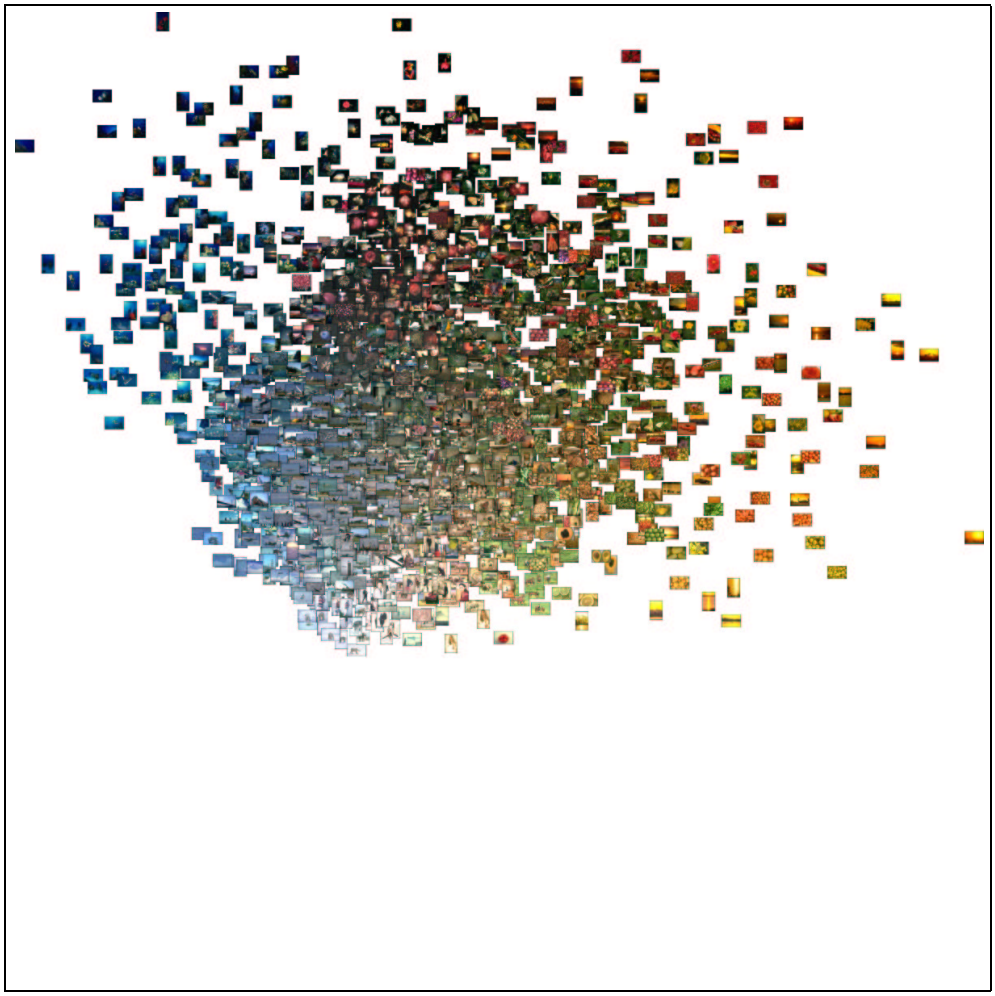
(a) Using the a.1 measure.

Figure D.1: MDS arrangements of the 2000 Corel 1 images.



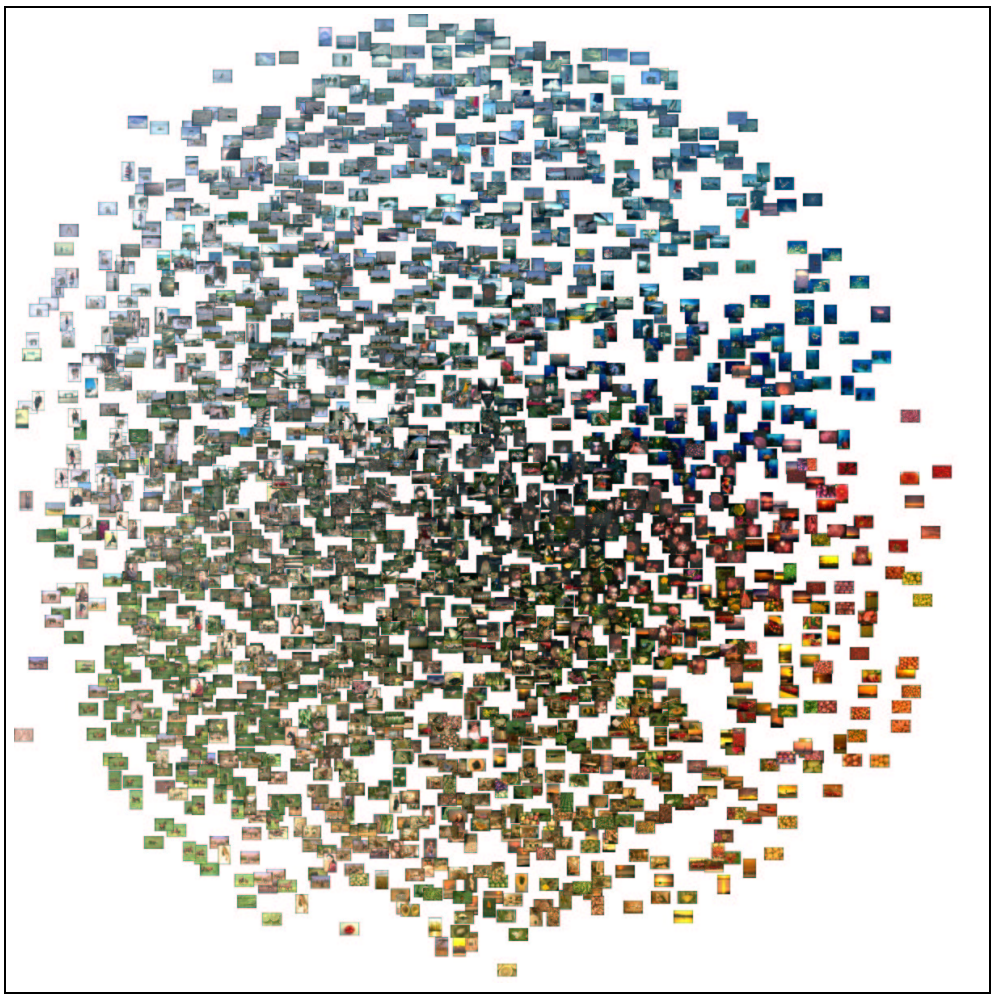
(b) Using the a.9 measure.

Figure D.1: MDS arrangements of the 2000 Corel 1 images.



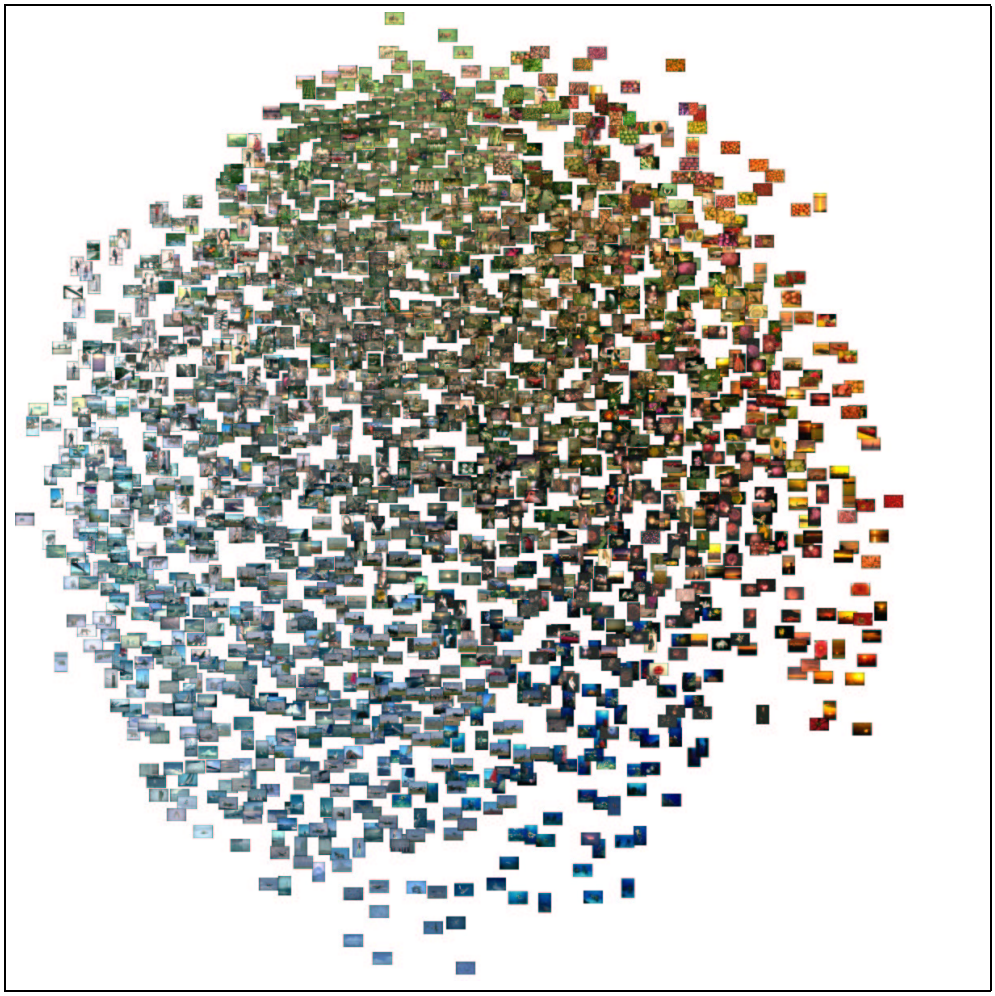
(c) Using the emd measure.

Figure D.1: MDS arrangements of the 2000 Corel 1 images.



(d) Using the h.jd measure.

Figure D.1: MDS arrangements of the 2000 Corel 1 images.



(e) Using the iris measure.

Figure D.1: MDS arrangements of the 2000 Corel 1 images.

Experiment instructions

Thank you for agreeing to take part in this experiment. Firstly, if you are colour blind, or have any problems with your eyesight (e.g. if you need glasses and are not wearing them) then please tell me now, as this may make you ineligible to participate!

Otherwise, here's how the experiment will work. It will take about an hour in total, and consists of a series of image searching tasks. In each of these, you will be shown a photograph for ten seconds, and then it will disappear, leaving you the task of finding it **as quickly as you can** in a set of 80 photographs. The photograph is always present and always visible, although sometimes it may be a little overlapped by others. You will only have twenty seconds. Once you have found it, you should click on it using the mouse. Please try not to move the mouse from the centre of the screen until you have found the photograph with your eyes. You can only click once, and accuracy is just as important as speed. As soon as you click, or once the twenty seconds is up, the correct photograph will be highlighted in red. You will have to click on buttons to move between each stage, and the timers only start when you click these buttons.

The experiment is broken up into eight separate "blocks" of twelve searches, and each of these blocks takes about five minutes to complete. You can take a break between blocks. The set of 80 images can be laid out in two different ways, and you will have four blocks of one method followed by four blocks of the other. Before you start each method, you will have one block of it to practice on. Hopefully all should become clear once you have your first practice.

That's it. If you have any questions, then please ask them now, or during/after the training blocks. However, if they are about the goals of the experiment, you'll have to wait until after you have completed it!

Figure D.2: The first experiment — instructions.

Post-experiment questions

Name: Subject id:

What sort of searching strategy did you use for the “grid” method?

What sort of searching strategy did you use for the “clustered” method?

Which of the two methods did you prefer? What were their advantages and disadvantages?

Were some kinds of photographs easier or more difficult to find than others?

Figure D.3: The first experiment — post-experiment questionnaire.

Experiment instructions

Thank you for agreeing to take part in this experiment. Firstly, if you are colour blind, or have any problems with your eyesight (e.g. if you need glasses and are not wearing them) then please tell me now, as this will affect the result of the experiment!

Otherwise, here's how the experiment will work. It takes less than an hour in total, and consists of a series of image searching tasks. In each of these, you will be given a description of a type of photograph to look for, along with a set of 100 photographs, and I'd like you to find as many as possible that match the description, **as quickly as you can**. Sometimes you may be uncertain of whether or not an image matches the description, so just use your own judgement, imagining, for example, that you are a journalist looking for possible pictures to illustrate a story on that topic.

The 100 images are quite small, and so the left of the screen shows a magnified version of the area currently under the mouse pointer, to let you see images in more detail. You should click on each suitable image, using the mouse. Each image you select will be highlighted in the main display, and a copy of it will appear at the bottom of the screen, to help you keep track of those you have found. You can de-select an image by clicking on it again, or by clicking on the copy at the bottom of the screen. Click on the "Done" button when you have found as many as you can. This will be timed, and there is a time limit of two minutes on each search. After you finish a search, the system will move on to the next one. Hopefully, all of this should become clear to you once you actually get started.

There are two different methods (M and R) of laying out the set of 100 images, and I am comparing these to each other, so you will do 10 searches using each method. You will also have four practice searches using each method, to get used to it and to the program being used to run the experiment, so in all that makes:
4 practice searches of the first method
10 "real" searches of the first method
4 practice searches of the second method
10 "real" searches of the second method

Be aware that the program may run quite slowly, as it is using a lot of the computer's resources. In particular, it will take a little while to load each set of images. The description will always appear before the images do.

The whole thing will take about 45 minutes. If you have any questions, then please ask them now, or during/after the practice searches. However, if they are about the goals of the experiment, you'll have to wait until after you have completed it!

Figure D.4: The second experiment — instructions.

Post-experiment questions

Name: Subject id:

Please describe how you went about searching for images in each of the two methods (e.g. how did you decide where to look?)

Method R (which you did first):

Method M (which you did second):

Did you have a preference for one of the two methods? If so, which, and why?

Please indicate your agreement or disagreement with the following statements:

	Agree strongly				Disagree strongly
a) The search descriptions were clear	1	2	3	4	5
b) The descriptions gave me a "mental image" of what to search for	1	2	3	4	5
c) It was difficult to decide if an image was correct	1	2	3	4	5
d) I was sure that I had found all of the images before I pressed "Done"	1	2	3	4	5
e) While searching, I mostly looked at the overview display	1	2	3	4	5
f) The images in the overview display were too small	1	2	3	4	5
g) The magnification tool was easy to use	1	2	3	4	5
h) The magnified images were too small	1	2	3	4	5

If you wish to comment on any of these statements, please do so here:

(a) Page one

Figure D.5: The second experiment — post-experiment questionnaire.

There follows a list of the searches you did in the main part of the experiment. Please place a tick next to any that you found particularly easy, and a cross next to any that you found particularly difficult (leaving the others blank):

fungi (e.g. toadstools)
fish
close-ups of flowers
skiers
small invertebrates (e.g. insects, spiders, moths, wasps)
birds
racing cars
surfing
dolphins or whales
deer
houses, from the outside
big cats (e.g. wildcats)
hot air balloons
fruit
sailing
sunsets
trains
planes or helicopters
the interiors of houses (e.g. kitchens, bathrooms)
female fashion models

Which factors made a search easier?

Which factors made a search more difficult?

(b) Page two

Figure D.5: The second experiment — post-experiment questionnaire.

The infodesign 99 experiment

Alaska

No other region in North America possesses the mythical aura of Alaska. Few who see this land of gargantuan ice fields, sweeping tundra, glacially excavated valleys, lush rainforests, deep fjords and occasionally smoking volcanoes leave unimpressed. All but three of the nation's highest peaks are found within its boundaries and one glacier alone is twice the size of Wales. Wildlife may be under threat elsewhere, but here it is abundant, with grizzly bears standing twelve feet tall, moose stopping traffic in downtown Anchorage, wolves prowling through national parks, bald eagles circling over the forests and fifty-pound-plus salmon leaping upstream. A mere 570,000 people live in this huge state — over forty percent of them in Anchorage — of whom only one-fifth were born here. As a rule of thumb, the more winters you have endured, the more Alaskan you are. Throughout this century tens of thousands have been lured by the promise of wealth, first by gold and then by fishing, logging and, most recently, oil. Experiencing Alaska on a low budget is possible, but requires a lot of planning. Winter, when hotels drop their prices by as much as half, is becoming an increasingly popular time to visit, particularly for the dazzling aurora borealis ("Northern Lights").

Kenya

With its long, tropical beaches and dramatic wildlife parks, Kenya has an exotic tourist image. Justifiably, for this is one of the most beautiful lands in Africa and a satisfyingly exciting and relatively easy place to travel, whether on a short holiday or an extended stay. Treating Kenya as a succession of tourist sights, however, is neither the best nor the most enjoyable way of experiencing the country. Travelling independently, you can enter the more genuine and very different world inhabited by most Kenyans: a ceaselessly active, contrasting landscape of farm and field, of streams and bush paths, of wooden and corrugated-iron shacks, tea shops and lodging houses, of crammed buses and pick-up vans, of overloaded bicycles, and of streets wandered by goats and chickens and toddlers. You'll find a rewarding degree of warmth, openness and curiosity in Kenya's towns and villages, especially off the more heavily trodden tourist routes. Out in the wilds, there is an abundance of authentic scenic glamour — vistas of rolling savannah dotted with Masai and their herds, high Kikuyu moorlands, dense forests bursting with bird song and insect noise, and stony, shimmering desert. And, of course, everywhere you go Kenya's wildlife adds a startling and rapidly addictive dimension.

New York City

New York City is the most beguiling place there is. You may not think so at first, but spend even a week here and it happens — the pace, the adrenaline take hold, and the shock gives way to myth. Walking through the city streets

is an experience, the buildings like icons to the modern age, and above all to the power of money. Despite all the hype, the movie-image sentimentalism, Manhattan — the central island and the city's real core — has massive romance: whether it's the flickering lights of the midtown skyscrapers as you speed across the Queensboro bridge, the 4am half-life in Greenwich Village, or just wasting the morning on the Staten Island ferry, you really would have to be made of stone not to be moved by it all. The city also has more straightforward pleasures. There is the architecture of corporate Manhattan and the more residential Upper East and West Side districts (the whole city reads like an illustrated history of modern design). You can eat anything, at any time, cooked in any style; drink in any kind of company; sit through any number of obscure movies. And for the avid consumer, the choice of shops is vast, almost numbingly exhaustive in this heartland of the great capitalist dream.

Paris

Paris is the paragon of style — perhaps the most captivating city in Europe. Famous names and events are invested with a glamour that elevates the city and its people to a legendary realm, and it still clings to its status as an artistic, intellectual and literary pacesetter. The city's history has conspired to create this sense of being apart. The supremely autocratic Louis XIV made Paris into a glorious symbol of the pre-eminence of the state, a tradition his successors have been happy to follow. Recent presidents have initiated the skyscrapers at La Défense, the Pompidou Centre in Beaubourg, Les Halles shopping precinct, the glass pyramid entrance to the Louvre, the Musée d'Orsay, and the Bastille opera house. Nowadays the most tangible and immediate pleasures of Paris are to be found in its street life and along the lively banks of the River Seine. Few cities can compete with the cafés, bars and restaurants — modern and trendy, local and traditional, humble and pretentious — that line every street and boulevard. An imposing backdrop is provided by the monumental architecture of the Arc de Triomphe, the Louvre, the Eiffel Tower, the Hôtel de Ville, the bridges and the institutions of the state.

Experiment instructions

Thank you for agreeing to take part in this experiment. Firstly, if you are colour blind, or have any problems with your eyesight (e.g. if you need glasses and are not wearing them) then please tell me now, as this will affect the result of the experiment! Otherwise, please read on…

You have been asked to choose photographs to illustrate three "destination guide" articles for a new "independent travel" World Wide Web site. Each article will be an overview of a different location, and is to appear on a separate page. The articles have not yet been written, so all you have are short summaries to indicate the general impression that each will convey. You also have 100 photographs of each location, and your task is to choose 3 of the photos (to be used together) for each article. **It is entirely up to you to decide on the criteria you use to make your selections – there are no "right" answers, and you are not bound by the given summaries.** No page layout has been decided yet, so the orientation of the photographs is not important.

You will be using a program to browse the photos, and before you start I will explain to you how it works, and give you a brief demonstration. You can then practice using it until you are comfortable, and ask any questions that you may have about the task or the program. You will then go on to choose illustrations for the three articles, one after the other, each from a different set of pictures. You will be timed, so **please work quickly** (imagine that you have a pressing deadline!) while still ensuring that you are happy with your selections.

At the end, I will ask you for your opinions about the task and the program. The whole thing should take about 20 minutes.

Figure D.6: The infodesign 99 study — instructions.

Post-experiment questions

Name: Experiment ID:

Please indicate your agreement or disagreement with the following statements:

	Agree strongly				Disagree strongly
a) I had a clear idea of what I was supposed to do	1	2	3	4	5
b) I thought the task was realistic	1	2	3	4	5
c) I have prior experience of carrying out picture selection	1	2	3	4	5
d) The articles gave me "mental images" of possibly suitable photos	1	2	3	4	5
e) The arrangement of photos by **caption** similarity was useful	1	2	3	4	5
f) The arrangement of photos by **image** similarity was useful	1	2	3	4	5
g) The task would have been just as easy with a random arrangement	1	2	3	4	5
h) It was useful to have two different views of the same set of photos	1	2	3	4	5
i) The two views complemented each other well	1	2	3	4	5
j) The magnified photos were too small	1	2	3	4	5

If you wish to comment on any of these statements, please do so here:

Please indicate the importance of these criteria to you when making your selections:

	Very important				Not at all important
a) Technically good, striking photographs	1	2	3	4	5
b) Photographs that were relevant to the given article	1	2	3	4	5
c) Photographs that worked well as a set of three	1	2	3	4	5
d) Photographs that fitted your own impressions of the place	1	2	3	4	5

If you were using any other criteria, please describe them below:

[over...]

(a) Page one

Figure D.7: The infodesign 99 study — post-experiment questionnaire.

There follows a list of the three articles you were asked to illustrate. Please indicate how satisfied you were with your final photograph selections for each of them:

	Very satisfied				Not at all satisfied
Paris	1	2	3	4	5
Kenya	1	2	3	4	5
Alaska	1	2	3	4	5

If you were particularly satisfied or dissatisfied with any of the three, why was that?

Did you have a preference for either of the two layout types? If so, which, and why?

Are there any extra facilities you would have liked the program to have?

Any other comments about the task or the program?

[that's it – thank you!]

(b) Page two

Figure D.7: The infodesign 99 study — post-experiment questionnaire.

(a) *Alaskan* (309034, chosen 6 times)

(b) *Northern Lights* (309032, chosen 4 times)

(c) *City Market, Nairobi* (253025, chosen 5 times)

(d) *Vegetable Market* (253058, chosen 5 times)

(e) *The Cafe Aux Deux Magots, across from Saint Germain des Pres* (223074, chosen 5 times)

Figure D.8: The most popular images selected in the infodesign 99 experiment, for Alaska, Kenya, and Paris. Of 162 images selected in total (18 participants $\times$ 3 choices $\times$ 3 places), 103 were unique. 10 further images were selected 3 times.

The Anglia experiment

Brazil

Brazilians often say they live in a continent rather than a country, and that's an excusable exaggeration: the landmass is bigger than the United States if you exclude Alaska. Brazil has no mountains to compare with its Andean neighbours, but in every other respect it has all the scenic — and cultural — variety you would expect from so vast a country.

Despite the immense expanses of the interior, roughly two-thirds of Brazil's population live on or near the coast; and well over half live in cities — even in the Amazon. It's fair to say that nowhere in the world do people know how to enjoy themselves more — most famously in the annual orgiastic celebrations of Carnaval, but this national hedonism also manifests itself in the country's highly developed beach culture. And if you needed more reason to visit, there's a strength and variety of popular culture, and a genuine friendliness and humour in the people that is tremendously welcoming and infectious.

Canada

Canada is almost unimaginably vast. It stretches from the Atlantic to the Pacific and from the latitude of Rome to beyond the Magnetic North Pole. Its archetypal landscapes are the mountain lakes and peaks, the endless forests and the prairie wheatfields, but the country also holds landscapes that defy expectations. Great tracts of Canada are completely unspoiled — ninety percent of the country's 28.5 million population lives within 100 miles of the US border.

Like its neighbour, Canada is a spectrum of cultures, a hotchpotch of immigrant groups who supplanted the continent's many native peoples. Alongside the French and British majorities live a host of communities who maintain the traditions of their homelands. The typical Canadian might be an elusive concept, but you'll find there's a distinctive feel to the country. Some towns might seem a touch too well-regulated and unspontaneous, but against this there's the overwhelming sense of Canadian pride in their history and pleasure in the beauty of their land.

Death Valley, USA

Death Valley is utterly inhuman: the hottest place on earth and almost entirely devoid of shade. Throughout the summer, the air temperature averages 112°F, and the ground can reach near boiling point. Better to come during the spring, when the wild flowers are in bloom, or from October to May, when it's generally mild and dry.

Its sculpted rock layers form deeply shadowed, eroded crevices at the foot of sharply silhouetted hills, their exotic mineral content turning million- year-old mudflats into rainbows of sunlit phosphorescence. A good first stop is

the Artist's Palette, an eroded hillside covered in an intensely colored mosaic of reds, golds, blacks and greens. It is overlooked by Zabriskie Point, from which the view is best during the early morning, when the pink and gold Panamint Mountains across the valley are highlighted by the rising sun. In the west spread fifteen rippled and contoured square miles of ever-changing sand dunes.

Denmark

Delicately balanced between Scandinavia proper and mainland Europe, Denmark is a difficult country to pin down. It has prices and drinking laws that are broadly in line with those in the rest of the EU, but social benefits and the standard of living are high. It is the easiest Scandinavian country in which to travel, both in terms of cost and distance, but the landscape itself is the region's least dramatic: largely farmland, interrupted by innumerable pretty villages.

The vast majority of visitors make for Zealand, and, more specifically, Copenhagen, the country's one large city and an exciting focal point, with a beautiful old centre, a good array of castles, gardens, and museums, and a boisterous nightlife. It completely dominates Denmark, but is one of Europe's most user-friendly cities, with a compact, strollable centre largely given over to pedestrians. Only the peninsula of Jutland is far enough away from Copenhagen to enjoy a truly individual flavour, as well as Denmark's most varied scenery.

Devon and Cornwall

This part of England's West Country encompasses everything from genteel, cosy villages to vast Atlantic-facing strands of golden sand and wild expanses of granite moorland. Warmed by the Gulf Stream, and enjoying more hours of sunshine than virtually anywhere else in England, it can sometimes come fairly close to the atmosphere of the Mediterranean.

With its rolling meadows, narrow lanes and remote thatched cottages, the region has long been the urbanite's ideal vision of a pre-industrial, "authentic" England. Zealous care is taken to preserve the undeveloped stretches of countryside and coast in the condition that has made them so popular. Pockets of genuine tranquillity are still to be found all over these counties, from villages with an appeal that goes deeper than mere picturesqueness, to quiet coves on the spectacular coastline. The full elemental power of the ocean can be appreciated here too, with splintered cliffs resounding to the constant thunder of the waves.

Ireland

Landscape and people are what bring most visitors to Ireland — south and north. And once there, few are disappointed by the reality: the green, rain-hazed loughs and wild, bluff coastlines, the inspired talent for conversation,

the easy pace of life. Ireland is becoming increasingly integrated with the industrial economies of western Europe, yet the modernisation of the country has to date made few marks.

It's a place to explore slowly, roaming through agricultural areas scattered with farmhouses, where the pastures, low wooded hills, and wide peat bogs are the classic landscapes of Ireland. Most visitors are drawn to the endlessly indented coastlines, which combine vertiginous cliffs and boulder-strewn wastes with mystical lakes and glens. In town, too, the pleasures are unhurried: evenings over Guinnesses in the snug of a pub, listening to the chat around a turf fire. And Dublin is an extraordinary mix of youthfulness and tradition, a human-scale capital of decaying Georgian squares and vibrant pubs.

Jamaica

Rightly famous for its beaches, beautiful, brash Jamaica is much more besides. There's certainly plenty of enchanting white sand, turquoise sea and swaying palm trees, and the resorts have become well-suited for those who want to head straight from plane to beach, never leaving their hotel compound. But to get any sense of the country at all you'll need to get into exploring mode. It's undoubtedly worth it, as this is a country packed with first-class attractions, oozing with character.

You can hike in the tropical rainforest or take a river rafting trip, and Jamaica even has mountains, whose cool woodlands, dotted with coffee plantations and often shrouded in mist, offer a welcome break from the heat of the coast. Despite the island's immense natural allure, it's not just the physical aspect that makes the country so absorbing. Notwithstanding the invasion of tourists, Jamaica retains an attitude — a personality — that's more resonant and distinctive than you'll find in any other Caribbean nation: it's a country with a swagger in its step.

Kenya

Kenya justifiably has an exotic tourist image, for it is one of the most beautiful lands in Africa and a satisfyingly exciting and relatively easy place to travel, whether on a short holiday or an extended stay. Out in the wilds, there is an abundance of authentic scenic glamour: vistas of rolling savannah are dotted with the wildlife that, everywhere you go, adds a startling and rapidly addictive dimension.

Treating Kenya as a succession of tourist sights, however, is neither the best nor the most enjoyable way of experiencing the country. Travelling independently, you can enter the more genuine and very different world inhabited by most Kenyans: a ceaselessly active, contrasting landscape. You'll find a rewarding degree of warmth, openness and curiosity in Kenya's towns and villages. Of all the country's peoples, the Maasai and Samburu have received the most attention, and the tourist industry has given them a major spot in its repertoire. For the people themselves, however, the rewards are fairly scant.

Nepal

Nepal forms the very watershed of Asia. Landlocked between India and Tibet, it spans terrain from subtropical jungle to the icy Himalaya, and contains eight of the world's ten highest mountains. Its cultural landscape is every bit as diverse: a dozen major ethnic groups, speaking as many as fifty languages and dialects, coexist in this narrow, jumbled buffer state, while two of the world's great religions, Hinduism and Buddhism, overlap and mingle with older tribal traditions — yet it's a testimony to the Nepalis' tolerance and good humour that there is no tradition of ethnic or religious strife.

Unlike India, Nepal was never colonised, a fact which comes through in fierce national pride and other, more idiosyncratic ways. Founded on trans-Himalayan trade, its dense, medieval cities display a unique pagoda-style architecture, not to mention an astounding flair for festivals and pageantry. But above all, Nepal is a nation of unaffected villages and terraced hillsides — more than eighty percent of the population lives off the land.

Yellowstone National Park, USA

Millions of visitors each year come to Yellowstone National Park, America's oldest national park and the largest in the lower 48 states, to glory in its magnificent mountain scenery and abundant wildlife, and to witness hydrothermal phenomena on a unique scale. The park contains more than half the world's geysers, plus hot springs, and thousands of fumaroles jetting plumes of steam. Close by are the classic triangular peaks of Grand Teton National Park. These sheer-faced cliffs make a magnificent spectacle, rising abruptly to tower 7000ft above the valley floor. A string of gem-like lakes is set tight at the foot of the mountains; beyond them the broad, sagebrush-covered valley is broken by the winding river.

Yellowstone amounts to an extraordinary experience, combining the colours of the wild flower meadows and the rainbow-hued geyser pools with the sounds of subterranean rumblings and steam hissing from the mountainsides, alongside the presence of browsing bull moose, shambling bears, heavy-bearded bison, and herds of elk.

Experiment instructions

Thank you for agreeing to take part in this experiment. Firstly, if you are colour blind, or have any problems with your eyesight (e.g. if you need glasses and are not wearing them) then please tell me now, as this will affect the result of the experiment! Otherwise, please read on…

The Task

You have been asked to choose photographs to appear on a set of nine "destination guide" pages, for a new "independent travel" World Wide Web site. Each page will contain an overview of a different location. The text for the pages has not yet been written, but the project director has provided you with short pieces of background information on each place, which are intended to give you the general impression of what the finished text will convey.

You have 100 photographs of each location, all of which have been approved as suitable by the project director. Your task is to choose **three** of the photos (to be used together) for each web page. Please remember that **there are no "right" answers**. The project director would like three interesting photos that work well together for each location, and has employed you for your creativity. No page layout has been decided yet, so the orientation of the photographs is not important.

You will be using a computer program to browse the photos. It is capable of arranging them on the screen using two different organisation methods, and before you start I will explain to you how it works. You can then practice using it, with an example location, and ask any questions that you may have about the task or the program. You will then go on to choose photos for nine web pages, each from a different set of pictures. You will be timed, so **please work quickly, spending not more than a few minutes on each location** (imagine that you have a pressing deadline) while still ensuring that you are happy with your selections.

What to do

1. Fill in the "About You" sheet, up to but not including the "Test Answers" section.
2. Once everyone has done this, I will explain the experiment software, and you can practice using it. On the computer, double-click on the file called "practice" to start it up, and follow the instructions on the screen.
3. When you are ready to start Part 1 of the experiment, double-click on the file called "person _part1". In Part 1 you have to choose photos for six locations. The first three will be with one of the organisation methods (on its own), and the second three will be with the other (on its own).
4. Let me know when you have finished Part 1. I will then give you out a short test to do. Please do not write on the question sheets – put your answers in the "Test Answers" section of the "About You" sheet. Please work through it as quickly as you can.
5. When you have finished the test, you can start Part 2 of the experiment, by double-clicking on the file called "person _part2". In Part 2 you have to choose photos for three locations, and this time you have both organisation methods available. You have to choose one to start with, and then you can switch between them as often as you like.
6. Let me know when you have finished Part 2. The last thing I will ask you to do is to fill in a second questionnaire, giving your opinions on the task and the program.
7. Claim your five pounds!

If you have any problems or questions at any point, then please just ask me.

Figure D.9: Anglia experiment — instructions. The blanks in the file names were filled with the hand-written ID number of each participant.

About you [Reference number:]

Name:

Email:

Year and course:

Please circle a number to indicate your agreement or disagreement with the following statements:

	Disagree strongly						Agree strongly
a) I am familiar with using a mouse	0	1	2	3	4	5	6
b) I am familiar with Microsoft Windows	0	1	2	3	4	5	6
c) I play computer games frequently	0	1	2	3	4	5	6
d) I have prior experience of carrying out picture/photo selection	0	1	2	3	4	5	6

Test answers:

1.
2.
3.
4.
5.
6.
7.
8.
9.
10.
11.
12.

Figure D.10: Anglia experiment — "About you" questionnaire.

Post-experiment questions [Reference number:]

There follows a list of the nine locations for which you were asked to choose photos. Please circle a number to indicate how satisfied you were with your final photograph selections for each of them:

	Not at all satisfied						**Very satisfied**
Brazil	0	1	2	3	4	5	6
Canada	0	1	2	3	4	5	6
Death Valley	0	1	2	3	4	5	6
Denmark	0	1	2	3	4	5	6
Ireland	0	1	2	3	4	5	6
Jamaica	0	1	2	3	4	5	6
Kenya	0	1	2	3	4	5	6
Nepal	0	1	2	3	4	5	6
Yellowstone National Park	0	1	2	3	4	5	6

If you were particularly satisfied or dissatisfied with any of them, why was that?

Please indicate the importance of these criteria to you when making your selections:

	Not at all important						**Very important**
a) Technically good, striking photographs	0	1	2	3	4	5	6
b) Photographs that worked well as a set of three	0	1	2	3	4	5	6
c) Photographs that were relevant to the given text	0	1	2	3	4	5	6
d) Photographs that fitted your own impressions of the place	0	1	2	3	4	5	6

If you were using any other criteria, please describe them below:

[please turn over…]

(a) Page one

Figure D.11: Anglia experiment — post-experiment questionnaire.

Please indicate your agreement or disagreement with the following statements:

	Disagree strongly						Agree strongly
a) I thought the task was realistic	0	1	2	3	4	5	6
b) It was easy to find suitable photos from those available	0	1	2	3	4	5	6
c) I browsed the photos with "mental images" of what I wanted	0	1	2	3	4	5	6
d) I liked having two different arrangements of the same set of photos	0	1	2	3	4	5	6
e) It made no difference to me how the photos were arranged	0	1	2	3	4	5	6
With regard to the *Library* organisation of photos:							
f) It was enjoyable to use	0	1	2	3	4	5	6
g) I thought it was useful	0	1	2	3	4	5	6
h) It made it easy for me to find the photos I wanted	0	1	2	3	4	5	6
i) It made it easy to find photos that complemented each other	0	1	2	3	4	5	6
With regard to the *Visual* organisation of photos:							
j) It was enjoyable to use	0	1	2	3	4	5	6
k) I thought it was useful	0	1	2	3	4	5	6
l) It made it easy for me to find the photos I wanted	0	1	2	3	4	5	6
m) It made it easy to find photos that complemented each other	0	1	2	3	4	5	6

If you wish to comment on any of these statements, please do so here:

Please describe any advantages or disadvantages that you found with regard to the different organisations of photos:

Library Advantages:

Disadvantages:

Visual Advantages:

Disadvantages:

Using both Advantages:

Disadvantages:

(b) Page two

Figure D.11: Anglia experiment — post-experiment questionnaire.

Did you prefer having the Library arrangement on its own, the Visual arrangement on its own, or having both arrangements available?

Please try to describe how you went about choosing photos, generally, from forming an idea to making your actual selections, e.g.

- Did your ideas evolve as you looked through the photos?
- Was the process different between the two types of organisation?

Are there any extra facilities you would have liked the program to have?

Any other comments about the task or the program?

Finally, would you be willing to participate in a follow-up experiment (for similar payment) at a later date?

[that's it – thank you!]

(c) Page three

Figure D.11: Anglia experiment — post-experiment questionnaire.

(a) *Sunset* (39044, chosen 6 times)

(b) *Indian Paintbrush* (39062, chosen 3 times)

(c) *Windmill, Skagen* (121048, chosen 5 times)

(d) *Front View Of Boats And Buildings On The Nyhavn, Copenhagen* (121064, chosen 3 times)

(e) *Seven Mile Beach, Negril* (328001, chosen 5 times)

(f) *Halifax Harbor Bridge, Nova Scotia* (230087, chosen 4 times)

(g) *Grizzly* (94080, chosen 3 times)

(h) *Hang-Glider Soaring Over Rio de Janeiro* (93009, chosen 3 times)

Figure D.12: The most popular images selected in the Anglia experiment (for Death Valley, Denmark, Jamaica, Canada, and Brazil). Of 270 images selected in total (10 participants $\times$ 3 choices $\times$ 9 places), 197 were unique. A total of 39 images were chosen twice.

(i) *Amboseli Plain and Mount Kilimanjaro* (253026, chosen 3 times)

(j) *Small Hindu temple, Katmandu* (187065, chosen 3 times)

(k) *Always a friendly smile* (385031, chosen 3 times)

(l) *Dawn over the Emerald Isle* (385052, chosen 3 times)

(m) *Giant's Causeway, Northern Ireland* (385086, chosen 3 times)

Figure D.12: Continued (Kenya, Nepal, and Ireland).

The Anglia follow-up

The texts for Brazil, Canada, Denmark, and Kenya were the same as those used in the original experiment.

Czech Republic

Czechoslovakia's "Velvet Revolution" in November 1989 was probably the most unequivocably positive of eastern Europe's anti-Communist upheavals, as the Czechs and Slovaks shrugged off 41 years of Communist rule without a shot being fired. But just three years on, the country split into two separate states: the Czech Republic and Slovakia. The Czechs — always the most urbane, agnostic and liberal of Slav nations — have fared best, enjoying political and economic stability, and attracting more Western investment than their former eastern bloc rivals.

Almost untouched by this century's wars, the capital, Prague, is justifiably one of the most popular destinations in Europe. An incredibly beautiful city with a wealth of architecture from Gothic cathedrals and Baroque palaces to Art Nouveau cafés and Cubist villas, it's also a lively meeting place for young people from all over Europe. The lush, surrounding countryside is studded with well-preserved medieval towns. The country's eastern province, Moravia, is every bit as beautiful, only less touristed.

France

Straddling the continent between the Iberian peninsula and the nations of central Europe, France is a core country on any European tour. It would be hard to exhaust its diversity in a lifetime of visits. Each area looks different, feels different, has its own style of architecture and food and often its own patois or dialect. There is an astonishing variety of things to see, whether it's Gothic cathedrals, châteaux, or Roman monuments. The countryside, too, has its own appeal, seemingly little changed for hundreds of years.

Travelling in France is easy. Budget restaurants and hotels proliferate; the rail and road networks are efficient; and the tourist information service is highly organized. As for where to go, it's hard to know where to begin. If you arrive in the north, you may pass through the Channel ports to Paris, one of Europe's most elegant and compelling capitals. To the west lie rocky coasts and, further south, châteaux, although most people push on south to limestone hills, canyons, and the glorious Mediterranean coastline.

Greek Islands

Greece's islands have enough appeal to fill months of travel. They range from outcrops where the boat calls once a week to resorts as cosmopolitan as they come. The historic sites span four millennia of civilization, encompassing the renowned and the obscure. The people's traditional culture lives on in the songs and dances, costumes, embroidery, and woven bags and rugs of popular

image. Its vigour may be failing rapidly under the impact of western consumer values, but much survives, especially in remoter regions.

Many of the islands are arid and rocky, with brilliant-white, cubist architecture. Mykonos is visited by nearly a million tourists a year, but if you don't mind the crowds — or you come out of season — this is one of the most beautiful of all island towns. Dazzlingly white, it's the archetypal postcard image, sugar-cube buildings stacked around a cluster of seafront fishermen's dwellings. Rhodes is also among the most visited of Greek islands. The core of its capital is a beautiful and remarkably preserved medieval city.

Mexico

Mexico enjoys a cultural blend that is wholly unique: among the fastest growing industrial powers on earth, its vast cities boast modern architecture to rival any in the world, yet it can still feel, in places, like a half-forgotten Spanish colony, while the all-pervading influence of native American culture, five hundred years on from the Conquest, is extraordinary. Each aspect can be found in isolation, but far more often, throughout the Republic, the three co-exist — indigenous markets, little changed in form since the arrival of the Spanish, thrive alongside elaborate colonial churches in the shadow of the skyscrapers of the Mexican miracle.

Despite encroaching Americanism, and close links with the rest of the Spanish-speaking world, the country remains resolutely individual. Mexico is still a country where timetables are not always to be entirely trusted, and where any attempt to do things in a hurry is liable to be frustrated. But for the most part, this is an easy, a fabulously varied, and an enormously enjoyable and friendly place in which to travel.

New York City

New York City is the most beguiling place there is. You may not think so at first — for the city is admittedly mad, the epitome in many ways of all that is wrong in modern America. But spend even a week here and it happens — the pace, the adrenaline take hold, and the shock gives way to myth. Walking through the city streets is an experience, the buildings like icons to the modern age, and above all to the power of money. The whole city reads like an illustrated history of modern design.

Manhattan — the central island and the city's real core — to many, is New York. Certainly, whatever your interest in the city it's here that you'll spend the most time. Despite all the hype, the movie-image sentimentalism, it has massive romance: looking at the flickering lights of the midtown skyscrapers as you speed across one of the bridges, you really would have to be made of stone not to be moved by it all.

Paris

Paris is the paragon of style — perhaps the most captivating city in Europe. Famous names and events are invested with a glamour that elevates the city and its people to a legendary realm, and it still clings to its status as an artistic, intellectual and literary pacesetter. The city's history has conspired to create this sense of being apart. The supremely autocratic King Louis XIV made Paris into a glorious symbol of the pre-eminence of the state, a tradition his successors have been happy to follow. Recent presidents have initiated the skyscrapers at La Défense, the Pompidou Centre at Beaubourg, Les Halles shopping precinct, the glass pyramid entrance to the Louvre, and the Musée d'Orsay.

Nowadays the most tangible and immediate pleasures of Paris are to be found in its street life and along the lively banks of the River Seine. An imposing backdrop is provided by the monumental architecture of the Arc de Triomphe, the Louvre, the Eiffel Tower, the Hôtel de Ville, the bridges and the institutions of the state.

San Francisco

America's favourite city sits at the edge of the Western world, a location that lends even greater romance to its legend. Pastoral, cosmopolitan and surprisingly small, San Francisco is a ravishing city, acting as a magnet for nearly three million visitors a year. For all its nostalgic, even provincial feel, San Francisco is an affluent, world-class city — its visitors alone pour nearly $2 billion into its coffers every year, and as the financial centre of the West Coast, its business community thrives.

The city is indisputably beautiful, a unique confection of switchback hills, wooden Victorian houses, open green spaces and the shimmering bay that surrounds. With a moody weather pattern that has San Francisco alternately drenched in sunshine or bathed in swirling fogs, the romance factor is high. An easy city to negotiate, one of the few US centres where you do not need a car, it also has an excellent public transport system that will whisk you round its appealing, diverse neighbourhoods.

Turkey

Turkey is a country with a multiple identity, poised uneasily between East and West. The country is now keen to be accepted on equal terms by the West, and has aspirations to EU membership. But it is by no stretch of the imagination a Western nation, and the contradictions persist: mosques co-exist with churches, and remnants of the Roman Empire crumble alongside ancient Hittite sites.

It's a vast country and incorporates large disparities in levels of development, but it is an immensely rewarding place to travel. Its ancient sites have been a magnet for travellers since the eighteenth century, and large numbers of visitors are drawn to the resort towns. Western Turkey is the most visited and economically developed part of the country. Istanbul is touted as Turkish

mystique par excellence, and understandably so: it would take months even to scratch the surface of the old imperial capital, still the cultural and commercial centre of the country.

The initial personal photography study

Questions asked in the structured interview:

- How big would you say your collection of photos is? How often do you look at them?
- How do you organise your photos? (e.g. date, place, event, subject, original roll, shoebox)
- How did you develop this way of organising them, and has it evolved from other ways of organising them in the past?
- Do you write anything down about them, on the back or elsewhere? If so, what sort of things do you write?
- Do you organise "good" and "bad" photos separately? What counts as "good" and "bad"? Do you throw any away?
- What do you use your photos for? (e.g. slide shows, work, art, memories)
- If you have a number of uses for them, do you have separate schemes for these? (e.g. different type of film, different way of organising)
- How often do you select a group of your pictures to show to others? Can you think of an example? How do you go about this?
- How do you go about finding a particular photo? Can you think of a typical example of doing that? How easy is it for you, usually?
- When you look, do you always do it with a specific remembered photo in mind, or is there ever a more vague need? How do you go about searching in that case?
- Have you ever felt restricted when organising by having only one copy of a photo, so that it can only be filed in one place? (or, the fact that copying a photo involves a bit of hassle and costs money)
- Which format(s) of film do you use? (e.g. 35mm negative, 35mm slide, medium, large, APS, digital)
- What do you do about developing?
- Do you use black and white film at all? If so, roughly what proportion of your photos are in black and white? Do you have a particular use for it?
- Have you used a digital camera at all? If you have, why, how much of the time, and what do you think of it? If you don't, why not, and is there anything that could change about them to make you want to use one?
- Do you scan any of your photos into a computer? If yes, what proportion of them?

- Do you use any photo editing software such as Photoshop? If so, which one, and what do you use it for?
- Have you ever used any computer system for managing photos? If so, which one, and what did you think of it?
- How familiar are you (1="very", 4="not at all") with: Windows-based PCs, Apple Macs, Unix-based machines, other computers (specify)?
- Assuming that all of your photos had been digitised, how do you think you might use a computer to organise them?
- I'm going to read out a list of features that it should be possible to provide in a computer-based image management system, and I'd like you to tell me how useful you think you would find a particular feature if it was available (1="very useful", 4="not at all useful"), commenting on your answers. Please ask for clarification of any of these if you're not sure what I mean, and if you have no opinion or don't know, just say.

This is a list of features that it should be possible to provide in a computer-based image management system, and I'd like you to tell me how useful you think you would find a particular feature if it was available (from "very useful" to "not at all useful"). Please circle a number to indicate your opinion.

Please ask for clarification of any of these if you're not sure what I mean, and if you have no opinion or don't know, just say, and circle the question mark.

Storing and organising:

very			not at all		
1	2	3	4	?	Organising photos into separate "folders"
1	2	3	4	?	Creating "slide shows" of selected photos
1	2	3	4	?	Adding a title to a photo
1	2	3	4	?	Typing notes to associate with a photo or group of photos
1	2	3	4	?	Speaking notes to associate with a photo or group of photos
1	2	3	4	?	Having spoken notes automatically "recognised"

Browsing and searching:

very			not at all		
1	2	3	4	?	Seeing all of the photos in a "folder" at once, reduced in size
1	2	3	4	?	Rapidly scanning through a group of photos (like video fast forward)
1	2	3	4	?	Searching for photos according to the date and time they were taken
1	2	3	4	?	Searching for photos based on the text of your notes
1	2	3	4	?	Searching for photos based on the colours present in them
1	2	3	4	?	Searching for photos based on the textures present in them
1	2	3	4	?	Searching for photos based on their layout/composition
1	2	3	4	?	Searching for other photos "similar" to a given one of your photos
1	2	3	4	?	Searching for other photos "similar" to another picture, e.g. a drawing
1	2	3	4	?	Choosing a region or regions of a photo and asking for other photos with regions similar to them
1	2	3	4	?	Specifying the position of a selected region in a photo
1	2	3	4	?	Specifying the relative positions of selected regions
1	2	3	4	?	Seeing a very large number of photos at once (see example) with "similar" photos clustered together

Figure D.13: Initial personal photography study — features questions.

Figure D.14: One of the example MDS layouts presented to the interviewees in the initial personal photography study.

The Shoebox study

The first interview

About non-digital photographs:

- How big would you say your existing (non-digital) collection of photos is? Are they prints, slides, or both?
- How often do you look at them?
- How do you organise them? (and does anyone else apart from you do this?) Are they all in the same place or mixed up with other things? For example, albums, slide boxes, date, place, event, subject, original roll, shoebox.
- What do you use your photos for? If you have a number of uses for them, do you have separate organisation schemes for these?
- How often do you need to find a particular photo? How do you go about it? How easy is it for you, usually?
- When you are searching your collection (as opposed to browsing), is it always to look for a specific remembered photo or photos, or is there ever a more general need? How do you go about searching in that case?
- Had you used a digital camera before starting the Shoebox trial?
- Had you scanned any of your photos into a computer? If yes, what proportion of them, and why?
- Had you used any photo editing software such as Photoshop? If so, which one, and what did you use it for?
- I'd like you to tell me how much you agree with each of the following statements (from 0="strongly disagree" to 6="strongly agree") with regard to your existing (non-digital) photograph collection. Please circle a number to indicate your opinion, commenting on each of your answers. Please ask for clarification of any of these if you're not sure what I mean.

About digital photography and Shoebox:

- How many photos have you taken so far?
- What's changed now that you're using a digital camera and Shoebox? (with regard to taking photos, using your photos, organising your collection, browsing and searching your collection)
- What's better about digital photos, and what's worse?
- Are you still taking non-digital photos? Will you go back to them when the trial is over?

Name: Date:

I'd like you to tell me how much you agree with each of the following statements (from 0="strongly disagree" to 6="strongly agree") with regard to your existing (non-digital) photograph collection. Please circle a number to indicate your opinion, commenting on each of your answers.

Please ask for clarification of any of these if you're not sure what I mean.

About your existing (non-digital) photograph collection:

strongly disagree						strongly agree	
0	1	2	3	4	5	6	My photograph collection is intentionally organised
0	1	2	3	4	5	6	I regard organising my photo collection as a chore
0	1	2	3	4	5	6	I am content with the way my photo collection is organised
0	1	2	3	4	5	6	I write notes about my photos (e.g. on the back, or in an album)
0	1	2	3	4	5	6	I separate "good" and "bad" photos in my collection
0	1	2	3	4	5	6	I throw away the "bad" photos
0	1	2	3	4	5	6	I browse my photo collection often
0	1	2	3	4	5	6	When I look at my photos, it is to search for something in particular
0	1	2	3	4	5	6	When I am looking for something in particular, it is easy for me to find it

Figure D.15: Shoebox study — non-digital photography questionnaire.

- What are your first impressions of Shoebox?
- Are there any features you'd like Shoebox to have that it doesn't already?
- I'm going to read out a list of some of Shoebox's features, and I'd like you to tell me how useful you think each feature is (0="not at all useful", 6="very useful"), commenting on your answers. Please ask for clarification of any of these if you're not sure what I mean, and if you haven't used the feature, just say. I'm interested both in the feature's usefulness "in principle", and its specific implementation "in practice" in Shoebox, so if you think your scores for these would be substantially different, please let me know.

The second interview

About digital photographs:

- How many photos have you taken? How many annotations have you added?
- How does using a digital camera and Shoebox compare to conventional photography?
  - Taking photos:
    - The number of photos you take, and the proportion that are "good" or "bad"
    - The subjects you take photos of
    - Using the screen in the camera to take photos, or see them immediately
  - Using your photos:
    - The way you show them to other people (laptop, PC, TV, Web, email, in the camera)
    - Giving copies to people (Web, email)
    - Editing them
  - Organising your collection:
    - The amount of intentional organising you do, and how organised you feel
    - How likely you are to make notes about the photos, or change the titles, or roll names (does having queries help?)
    - How you feel about typing annotations versus speaking them
  - Browsing and searching your collection:
    - How often you look at them
    - How easy or difficult it is to find what you want

Name: Date:

This is a list of some of Shoebox's features, and I'd like you to tell me how useful you think each feature is (from "not at all useful" to "very useful"). Please circle a number to indicate your opinion. If you have not used the feature, just tell me.

Please ask for clarification of any of these if you're not sure what I mean.

Storing and organising:

not at all						very	
0	1	2	3	4	5	6	Allowing the organising of photos into separate rolls
0	1	2	3	4	5	6	Giving a title to a roll of photos
0	1	2	3	4	5	6	Giving a title to a single photo
0	1	2	3	4	5	6	Typing annotations to associate with a photo or group of photos
0	1	2	3	4	5	6	Speaking annotations to associate with a photo or group of photos
0	1	2	3	4	5	6	Having spoken annotations automatically transcribed
0	1	2	3	4	5	6	Creating "slide shows" of selected photos
0	1	2	3	4	5	6	Publishing a set of photos as HTML, to put on the Web

Browsing and searching:

not at all						very	
0	1	2	3	4	5	6	Changing the size of the "thumbnail" photos
0	1	2	3	4	5	6	Listening to spoken annotations
0	1	2	3	4	5	6	Searching for photos according to the date and time they were taken (the timeline)
0	1	2	3	4	5	6	Using a query to search for photos based on the text of your annotations
0	1	2	3	4	5	6	Searching for other photos visually "similar" to a given one of your photos
0	1	2	3	4	5	6	Choosing a region or regions of a photo and retrieving photos with similar regions

Figure D.16: Shoebox study — Shoebox features questionnaire.

    - The type of search you do (do you make use of queries?), and how you go about looking

- Are prints important to you? Why, or why not?
- Are you still taking non-digital photos? Would you use them for a particular purpose? Will you go back to them when the trial is over?
- I'd like you to tell me how much you agree with each of the following statements (from 0="strongly disagree" to 6="strongly agree") with regard to your digital photograph collection. Please circle a number to indicate your opinion, commenting on each of your answers. Please ask for clarification of any of these if you're not sure what I mean.

About Shoebox:

- What is your final verdict on Shoebox? Will you keep using it? If not, would you keep using it if the bugs were fixed, or is there more that you dislike?
- Are there any features you'd like Shoebox to have that it doesn't already?
- I'm going to read out a list of some of Shoebox's features, and I'd like you to tell me how useful you think each feature is (0="not at all useful", 6="very useful"), commenting on your answers. Please ask for clarification of any of these if you're not sure what I mean, and if you haven't used the feature, just say. I'm interested both in the feature's usefulness "in principle", and its specific implementation "in practice" in Shoebox, so if you think your scores for these would be substantially different, please let me know.

Name: Date:

I'd like you to tell me how much you agree with each of the following statements (from 0="strongly disagree" to 6="strongly agree") with regard to your digital photograph collection. Please circle a number to indicate your opinion, commenting on each of your answers.

Please ask for clarification of any of these if you're not sure what I mean.

About your digital photograph collection:

strongly disagree						strongly agree	
0	1	2	3	4	5	6	My photograph collection is intentionally organised
0	1	2	3	4	5	6	I regard organising my photo collection as a chore
0	1	2	3	4	5	6	I am content with the way my photo collection is organised
0	1	2	3	4	5	6	I change the names of my photos from the default ones
0	1	2	3	4	5	6	I add typed annotations to my photos
0	1	2	3	4	5	6	I add spoken annotations to my photos
0	1	2	3	4	5	6	Spoken annotations are useful even without accurate speech recognition
0	1	2	3	4	5	6	I separate "good" and "bad" photos in my collection
0	1	2	3	4	5	6	I delete the "bad" photos
0	1	2	3	4	5	6	I browse my photo collection often
0	1	2	3	4	5	6	I miss having prints of my photos
0	1	2	3	4	5	6	When I look at my photos, it is to search for something in particular
0	1	2	3	4	5	6	When I am looking for something in particular, it is easy for me to find it

Figure D.17: Shoebox study — digital photography questionnaire.

Bibliography

[1] J. R. Anderson. *Learning and Memory: An Integrated Approach*. Wiley, New York, second edition, 2000.

[2] L. H. Armitage and P. G. B. Enser. Analysis of user need in image archives. *Journal of Information Science*, 23(4):287–299, 1997.

[3] A. J. Ashford, L. R. Conniss, and M. E. Graham. The user in image retrieval: a developing framework. In *The Challenge of Image Retrieval*. Electronic Workshops in Computing, `http://www.ewic.org.uk`, 2000.

[4] A. Baddeley. *Human Memory: Theory and Practice*. Psychology Press, Hove, revised edition, 1997.

[5] M. Balabanović, L. L. Chu, and G. J. Wolff. Storytelling with digital photographs. In *Proceedings of CHI 2000*, pages 564–571. ACM, 2000.

[6] D. Barreau and B. A. Nardi. Finding and reminding: File organization from the desktop. *ACM SIGCHI Bulletin*, 27(3):39–43, 1995.

[7] W. Basalaj. *Proximity Visualisation of Abstract Data*. PhD thesis, University of Cambridge Computer Laboratory, 2000.

[8] M. J. Bates. The design of browsing and berrypicking techniques for the online search interface. *Online Review*, 13(5):407–424, 1989.

[9] N. J. Belkin, P. G. Marchetti, and C. Cool. BRAQUE: Design of an interface to support user interaction in information retrieval. *Information Processing and Management*, 29(3):325–344, 1993.

[10] I. Borg and P. Groenen. *Modern Multidimensional Scaling*. Springer-Verlag, New York, 1997.

[11] C. L. Borgman. All users of information retrieval systems are not created equal: An exploration into individual differences. *Information Processing and Management*, 25(3):237–251, 1989.

[12] P. Borlund. Experimental components for the evaluation of interactive information retrieval systems. *Journal of Documentation*, 56(1):71–90, 2000.

[13] P. Borlund and P. Ingwersen. The development of a method for the evaluation of interactive information retrieval systems. *Journal of Documentation*, 53(3):225–250, 1997.

[14] B. Bradshaw. Semantic based image retrieval: A probabilistic approach. In *Proceedings of ACM Multimedia 2000*, pages 167–176. ACM, 2000.

[15] D. Brodbeck, M. Chalmers, A. Lunzer, and P. Cotture. Domesticating Bead: Adapting an information visualization system to a financial institution. In *Proceedings of the IEEE Symposium on Information Visualization (InfoVis'97)*, pages 73–80. IEEE, 1997.

[16] B. A. T. Brown, A. J. Sellen, and K. P. O'Hara. A diary study of information capture in working life. In *Proceedings of CHI 2000*, pages 438–445. ACM, 2000.

[17] M. G. Brown, J. T. Foote, G. J. F. Jones, K. Spärck Jones, and S. J. Young. Open-vocabulary speech indexing for voice and video mail retrieval. In *Proceedings of ACM Multimedia '96*, pages 307–316. ACM, 1996.

[18] I. Campbell. Interactive evaluation of the Ostensive Model using a new test collection of images with multiple relevance assessments. *Information Retrieval*, 2(1):85–112, 2000.

[19] S. K. Card, J. Mackinlay, and B. Shneiderman, editors. *Readings in Information Visualization: Using Vision to Think*. Morgan Kaufmann, San Francisco, 1999.

[20] C. Carson, M. Thomas, S. Belongie, J. M. Hellerstein, and J. Malik. Blobworld: A system for region-based image indexing and retrieval. In *Proceedings of the Third International Conference on Visual Information and Information Systems (VISUAL'99)*, volume 1614 of *Lecture Notes in Computer Science*, pages 509–516. Springer, 1999.

[21] M. Chalmers, R. Ingram, and C. Pfranger. Adding imageability features to information displays. In *Proceedings of UIST'96*, pages 33–39. ACM, 1996.

[22] S.-J. Chang and R. E. Rice. Browsing: A multidimensional framework. In M. E. Williams, editor, *Annual Review of Information Science and Technology*, volume 28, chapter 6, pages 231–276. Learned Information, Medford, New Jersey, 1993.

[23] C. Chen and M. Czerwinski. Spatial ability and visual navigation: An empirical study. *The New Review of Hypermedia and Multimedia*, 3:67–89, 1997.

[24] C. Chen and M. Czerwinski. Empirical evaluation of information visualizations: an introduction. *International Journal of Human–Computer Studies*, 53(5):631–635, 2000.

[25] F. Chen, U. Gargi, L. Niles, and H. Schütze. Multi-modal browsing of images in web documents. In *Document Recognition and Retrieval VI*, volume 3651 of *Proceedings of SPIE*, pages 122–133, 1999.

[26] H. Chen, A. L. Houston, R. R. Sewell, and B. R. Schatz. Internet browsing and searching: User evaluations of category map and concept space techniques. *Journal of the American Society for Information Science*, 49(7):582–603, 1998.

[27] J.-Y. Chen, C. A. Bouman, and J. C. Dalton. Similarity pyramids for browsing and organization of large image databases. In *Human Vision and Electronic Imaging III*, volume 3299 of *Proceedings of SPIE*, pages 563–575, 1998.

[28] T. T. A. Combs and B. B. Bederson. Does zooming improve image browsing? In *Proceedings of Digital Libraries '99*, pages 130–137. ACM, 1999.

[29] W. Conover. *Practical Nonparametric Statistics*. Wiley, London, second edition, 1980.

[30] W. S. Cooper. On selecting a measure of retrieval effectiveness, part I: The 'subjective' philosophy of evaluation. *Journal of the American Society for Information Science*, 24(2):87–100, 1973.

[31] I. J. Cox, J. Ghosn, M. L. Miller, T. V. Papathomas, and P. N. Yianilos. Hidden annotation in content-based image retrieval. In *Proceedings of the IEEE Workshop on Content-based Access of Image and Video Libraries*, pages 76–81. IEEE, 1997.

[32] D. R. Cutting, D. R. Karger, J. O. Pedersen, and J. W. Tukey. Scatter/Gather: A cluster-based approach to browsing large document collections. In *Proceedings of SIGIR'92*, pages 318–329. ACM, 1992.

[33] M. Das, R. Manmatha, and E. M. Riseman. Indexing flower patent images using domain knowledge. *IEEE Intelligent Systems*, 14(5):24–33, 1999.

[34] S. W. Draper and M. D. Dunlop. New IR – New Evaluation: The impact of interaction and multimedia on information retrieval and its evaluation. *The New Review of Hypermedia and Multimedia*, 3:107–121, 1997.

[35] M. Dunlop. Reflections on Mira: interactive evaluation in information retrieval. *Journal of the American Society for Information Science*, 51(14):1269–1274, 2000.

[36] J. P. Eakins and M. E. Graham. Content-based image retrieval. JISC Technology Applications Programme, Report 39, 1999.

[37] D. E. Egan. Individual differences in human–computer interaction. In M. Helander, editor, *Handbook of Human–Computer Interaction*, chapter 24, pages 543–568. North-Holland, Amsterdam, 1988.

[38] P. G. B. Enser. Query analysis in a visual information retrieval context. *Journal of Document and Text Management*, 1(1):25–52, 1993.

[39] P. G. B. Enser. Pictorial information retrieval. *Journal of Documentation*, 51(2):126–170, 1995.

[40] J.-D. Fekete and C. Plaisant. Excentric labeling: Dynamic neighborhood labeling for data visualization. In *Proceedings of CHI'99*, pages 512–519. ACM, 1999.

[41] R. Fidel. The image retrieval task: implications for the design and evaluation of image databases. *The New Review of Hypermedia and Multimedia*, 3:181–199, 1997.

[42] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker. Query by image and video content: The QBIC system. *IEEE Computer*, 28(9):23–32, 1995.

[43] E. Frøkjær, M. Hertzum, and K. Hornbæk. Measuring usability: Are effectiveness, efficiency, and satisfaction really correlated? In *Proceedings of CHI 2000*, pages 345–352. ACM, 2000.

[44] C. O. Frost, B. Taylor, A. Noakes, S. Markel, D. Torres, and K. M. Drabenstott. Browse and search patterns in a digital image database. *Information Retrieval*, 1(4):287–313, 2000.

[45] G. W. Furnas. Effective view navigation. In *Proceedings of CHI'97*, pages 367–374. ACM, 1997.

[46] G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais. The vocabulary problem in human–system communication. *Communications of the ACM*, 30(11):964–971, 1987.

[47] S. R. Garber and M. B. Grunes. The art of search: A study of art directors. In *Proceedings of CHI'92*, pages 157–163. ACM, 1992.

[48] A. Gupta and R. Jain. Visual information retrieval. *Communications of the ACM*, 40(5):70–79, 1997.

[49] K.-A. Han and S.-H. Myaeng. Image organization and retrieval with automatically constructed feature vectors. In *Proceedings of SIGIR'96*, pages 157–165. ACM, 1996.

[50] D. Harman. What we have learned, and not learned, from TREC. In *Proceedings of the BCS IRSG 22nd Annual Colloquium on Information Retrieval Research*, pages 2–20. BCS, 2000.

[51] M. A. Hearst. User interfaces and visualization. In R. Baeza-Yates and B. Ribeiro-Neto, editors, *Modern Information Retrieval*, chapter 10, pages 257–323. Addison Wesley, Harlow, 1999.

[52] M. A. Hearst and J. O. Pedersen. Reexamining the cluster hypothesis: Scatter/Gather on retrieval results. In *Proceedings of SIGIR'96*, pages 76–84. ACM, 1996.

[53] A. Hiroike, Y. Musha, A. Sugimoto, and Y. Mori. Visualization of information spaces to retrieve and browse image data. In *Proceedings of the Third International Conference on Visual Information and Information Systems (VISUAL'99)*, volume 1614 of *Lecture Notes in Computer Science*, pages 155–162. Springer, 1999.

[54] P. Holland. 'Sweet it is to scan...': Personal photographs and popular photography. In L. Wells, editor, *Photography: A Critical Introduction*, chapter 3, pages 103–150. Routledge, London, 1997.

[55] D. Hull. Using statistical testing in the evaluation of retrieval experiments. In *Proceedings of SIGIR'93*, pages 329–338. ACM, 1993.

[56] D. Hull. Stemming algorithms — a case study for detailed evaluation. *Journal of the American Society for Information Science*, 47(1):70–84, 1996.

[57] G. W. Humphreys and V. Bruce. *Visual Cognition: Computational, Experimental, and Neuropsychological Perspectives*. Lawrence Erlbaum Associates, Hove, 1989.

[58] L. Itti and C. Koch. Computational modelling of visual attention. *Nature Reviews Neuroscience*, 2(3):194–203, 2001.

[59] C. E. Jacobs, A. Finkelstein, and D. H. Salesin. Fast multiresolution image querying. In *Proceedings of SIGGRAPH'95*, pages 277–286. ACM, 1995.

[60] J. M. Jose, J. Furner, and D. J. Harper. Spatial querying for image retrieval: a user-oriented evaluation. In *Proceedings of SIGIR'98*, pages 232–241. ACM, 1998.

[61] M. Koskela, J. Laaksonen, S. Laakso, and E. Oja. The PicSOM retrieval system: Description and evaluations. In *The Challenge of Image Retrieval*. Electronic Workshops in Computing, `http://www.ewic.org.uk`, 2000.

[62] A. Kuchinsky, C. Pering, M. L. Creech, D. Freeze, B. Serra, and J. Gwizdka. FotoFile: A consumer multimedia organization and retrieval system. In *Proceedings of CHI'99*, pages 496–503. ACM, 1999.

[63] Y. Kural, S. Robertson, and S. Jones. Deciphering cluster representations. *Information Processing and Management*, 37(4):593–601, 2001.

[64] T.-S. Lai, J. Tait, and S. McDonald. A user-centred evaluation of visual search methods for CBIR. In *The Challenge of Image Retrieval*. Electronic Workshops in Computing, `http://www.ewic.org.uk`, 2000.

[65] M. Lansdale and E. Edmonds. Using memory for events in the design of personal filing systems. *International Journal of Man–Machine Studies*, 36(1):97–126, 1992.

[66] M. W. Lansdale, S. A. R. Scrivener, and A. Woodcock. Developing practice with theory in HCI: applying models of spatial cognition for the design of pictorial databases. *International Journal of Human–Computer Studies*, 44(6):777–799, 1996.

[67] A. Leuski and J. Allan. Evaluating a visual navigation system for a digital library. In *Proceedings of the Second European Conference on Research and Technology for Digital Libraries (ECDL'98)*, pages 535–554, 1998.

[68] A. Leuski and J. Allan. Improving interactive retrieval by combining ranked lists and clustering. In *Proceedings of RIAO 2000*, pages 665–681, 2000.

[69] A. Leuski and J. Allan. Lighthouse: Showing the way to relevant information. In *Proceedings of the IEEE Symposium on Information Visualization (InfoVis 2000)*, pages 125–130. IEEE, 2000.

[70] X. Lin. Searching and browsing on map displays. In *Proceedings of the 58th ASIS Annual Meeting*, pages 13–18, 1995.

[71] X. Lin. Map displays for information retrieval. *Journal of the American Society for Information Science*, 48(1):40–54, 1997.

[72] A. C. Loui and M. D. Wood. A software system for automatic albuming of consumer pictures. In *Proceedings of ACM Multimedia '99*, volume 2, pages 159–162. ACM, 1999.

[73] J. MacCuish, A. McPherson, J. Barros, and P. Kelly. Interactive layout mechanisms for image database retrieval. In *Visual Data Exploration and Analysis III*, volume 2656 of *Proceedings of SPIE*, pages 104–115, 1996.

[74] A. Mäkelä, V. Giller, M. Tscheligi, and R. Sefelin. Joking, storytelling, artsharing, expressing affection: A field trial of how children and their social network communicate with digital images in leisure time. In *Proceedings of CHI 2000*, pages 548–555. ACM, 2000.

[75] T. W. Malone. How do people organize their desks? Implications for the design of office information systems. *ACM Transactions on Office Information Systems*, 1(1):99–112, 1983.

[76] G. Marchionini. *Information Seeking in Electronic Environments*. Cambridge University Press, Cambridge, 1995.

[77] M. Markkula and E. Sormunen. Searching for photos — journalists' practices in pictorial IR. In *The Challenge of Image Retrieval*. Electronic Workshops in Computing, `http://www.ewic.org.uk`, 1998.

[78] M. Markkula and E. Sormunen. End-user searching challenges indexing practices in the digital newspaper photo archive. *Information Retrieval*, 1(4):259–285, 2000.

[79] J. Marks, B. Andalman, P. A. Beardsley, W. Freeman, S. Gibson, J. Hodgins, T. Kang, B. Mirtich, H. Pfister, W. Ruml, K. Ryall, J. Seims, and S. Shieber. Design Galleries: A general approach to setting parameters for computer graphics and animation. In *Proceedings of SIGGRAPH '97*, pages 389–400. ACM, 1997.

[80] J. E. McGrath. Methodology matters: Doing research in the behavioral and social sciences. In R. M. Baecker, J. Grudin, W. A. S. Buxton, and S. Greenberg, editors, *Readings in Human–Computer Interaction: Toward the Year 2000*, pages 152–169. Morgan Kaufmann, San Francisco, second edition, 1995.

[81] T. J. Mills, D. Pye, D. Sinclair, and K. R. Wood. Shoebox: A digital photo management system. Technical Report 2000.10, AT&T Laboratories Cambridge, 2000.

[82] J. Nielsen. *Usability Engineering*. Academic Press, London, 1993.

[83] V. L. O'Day and R. Jeffries. Orienteering in an information landscape: How information seekers get from here to there. In *Proceedings of INTERCHI'93*, pages 438–445. ACM, 1993.

[84] A. Oliva, A. B. Torralba, A. Guérin-Dugué, and J. Hérault. Global semantic classification of scenes using power spectrum templates. In *The Challenge of Image Retrieval*. Electronic Workshops in Computing, `http://www.ewic.org.uk/`, 1999.

[85] S. Ornager. The newspaper image database: Empirical supported analysis of users' typology and word association clusters. In *Proceedings of SIGIR'95*, pages 212–218. ACM, 1995.

[86] P. Over. The TREC interactive track: an annotated bibliography. *Information Processing and Management*, 37(3):369–381, 2001.

[87] T. V. Papathomas, T. E. Conway, I. J. Cox, J. Ghosn, M. L. Miller, T. P. Minka, and P. N. Yianilos. Psychophysical studies of the performance of an image database retrieval system. In *Human Vision and Electronic Imaging III*, volume 3299 of *Proceedings of SPIE*, pages 591–602, 1998.

[88] Z. Pečenović, M. Do, M. Vetterli, and P. Pu. Integrated browsing and searching of large image collections. In *Proceedings of the Fourth International Conference on Advances in Visual Information Systems (VISUAL 2000)*, volume 1929 of *Lecture Notes in Computer Science*, pages 279–289. Springer, 2000.

[89] P. Pirolli, S. K. Card, and M. M. Van Der Wege. Visual information foraging in a focus + context visualization. In *Proceedings of CHI 2001*, pages 506–513. ACM, 2001.

[90] P. Pirolli, P. Schank, M. Hearst, and C. Diehl. Scatter/Gather browsing communicates the topic structure of a very large text collection. In *Proceedings of CHI'96*, pages 213–220. ACM, 1996.

[91] J. C. Platt. AutoAlbum: Clustering digital photographs using probabilistic model merging. In *Proceedings of the IEEE Workshop on Content-Based Access of Image and Video Libraries*, pages 96–100. IEEE, 2000.

[92] M. F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

[93] J. Puzicha, Y. Rubner, C. Tomasi, and J. M. Buhmann. Empirical evaluation of dissimilarity measures for color and texture. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1165–1173. IEEE, 1999.

[94] M. C. Ramsey, H. Chen, B. Zhu, and B. R. Schatz. A collection of visual thesauri for browsing large collections of geographic images. *Journal of the American Society for Information Science*, 50(9):826–834, 1999.

[95] E. M. Rasmussen. Indexing images. In M. E. Williams, editor, *Annual Review of Information Science and Technology*, volume 32, chapter 3, pages 169–196. Information Today, Medford, New Jersey, 1997.

[96] J. C. Raven. Raven's Advanced Progressive Matrices. Available from `http://www.psychcorp.com`.

[97] E. S. Raymond, editor. *The New Hacker's Dictionary*. MIT Press, Cambridge, Massachusetts, third edition, 1996.

[98] J. Ritchie and L. Spencer. Qualitative data analysis for applied policy research. In A. Bryman and R. G. Burgess, editors, *Analyzing Qualitative Data*, chapter 9, pages 173–194. Routledge, London, 1993.

[99] B. E. Rogowitz, T. Frese, J. R. Smith, C. A. Bouman, and E. Kalin. Perceptual image similarity experiments. In *Human Vision and Electronic Imaging III*, volume 3299 of *Proceedings of SPIE*, pages 576–590, 1998.

[100] M. Rorvig. Images of similarity: A visual exploration of optimal similarity metrics and scaling properties of TREC topic-document sets. *Journal of the American Society for Information Science*, 50(8):639–651, 1999.

[101] M. Rorvig, K. Jeong, C. Suresh, and A. Goodrum. Exploiting image primitives for effective retrieval. In *The Challenge of Image Retrieval*. Electronic Workshops in Computing, `http://www.ewic.org.uk`, 2000.

[102] T. Rose, D. Elworthy, A. Kotcheff, A. Clare, and P. Tsonis. ANVIL: a system for the retrieval of captioned images using NLP techniques. In *The Challenge of Image Retrieval*. Electronic Workshops in Computing, `http://www.ewic.org.uk`, 2000.

[103] Y. Rubner, C. Tomasi, and L. J. Guibas. Adaptive color-image embeddings for database navigation. In *Proceedings of the Asian Conference on Computer Vision*, pages 104–111. IEEE, 1998.

[104] G. Salton. *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley, Reading, Massachusetts, 1989.

[105] S. Santini and R. Jain. Integrated browsing and querying for image databases. *IEEE Multimedia*, 7(3):26–39, 2000.

[106] L. Schamber. Relevance and information behavior. In M. E. Williams, editor, *Annual Review of Information Science and Technology*, volume 29, chapter 1, pages 3–48. Learned Information, Medford, New Jersey, 1994.

[107] M. M. Sebrechts, J. Vasilakis, M. S. Miller, J. V. Cugini, and S. J. Laskowski. Visualization of search results: A comparative evaluation of text, 2D, and 3D interfaces. In *Proceedings of SIGIR'99*, pages 3–10. ACM, 1999.

[108] S. Shatford Layne. Some issues in the indexing of images. *Journal of the American Society for Information Science*, 45(8):583–588, 1994.

[109] B. Shneiderman and H. Kang. Direct annotation: A drag-and-drop strategy for labeling photos. In *Proceedings of the International Conference on Information Visualisation*, pages 88–95. IEEE, 2000.

[110] D. Sinclair. Voronoi seeded colour image segmentation. Technical Report 1999.3, AT&T Laboratories Cambridge, 1999.

[111] J. R. Smith and S.-F. Chang. VisualSEEk: a fully automated content-based image query system. In *Proceedings of ACM Multimedia '96*, pages 87–98. ACM, 1996.

[112] E. Sormunen, M. Markkula, and K. Järvelin. The perceived similarity of photos — seeking a solid basis for the evaluation of content-based retrieval algorithms. In *Proceedings of the Final Mira Conference on Information Retrieval Evaluation*. University of Glasgow, 1999.

[113] R. Spence. *Information Visualization*. Addison-Wesley, Harlow, 2001.

[114] L. Stroebel, H. Todd, and R. Zakia. *Visual concepts for photographers*. Focal Press, London, 1980.

[115] L. T. Su. The relevance of recall and precision in user evaluation. *Journal of the American Society for Information Science*, 45(3):207–217, 1994.

[116] R. C. Swan and J. Allan. Aspect windows, 3-D visualizations, and indirect comparisons of information retrieval systems. In *Proceedings of SIGIR'98*, pages 173–181. ACM, 1998.

[117] J. Tague-Sutcliffe. The pragmatics of information retrieval experimentation, revisited. *Information Processing and Management*, 28(4):467–490, 1992.

[118] C. J. van Rijsbergen. *Information Retrieval*. Butterworths, London, second edition, 1979.

[119] C. J. van Rijsbergen and K. Spärck Jones. A test for the separation of relevant and non-relevant documents in experimental retrieval collections. *Journal of Documentation*, 29(3):251–257, 1973.

[120] A. Veerasamy and R. Heikes. Effectiveness of a graphical display of retrieval results. In *Proceedings of SIGIR'97*, pages 236–244. ACM, 1997.

[121] J. Vendrig, M. Worring, and A. W. M. Smeulders. Filter image browsing: Exploiting interaction in image retrieval. In *Proceedings of the Third International Conference on Visual Information and Information Systems (VISUAL'99)*, volume 1614 of *Lecture Notes in Computer Science*, pages 147–154. Springer, 1999.

[122] E. M. Voorhees. The cluster hypothesis revisited. In *Proceedings of SIGIR'85*, pages 188–196. ACM, 1985.

[123] C. Ware. *Information Visualization: Perception for Design*. Morgan Kaufmann, San Francisco, 2000.

[124] S. Whittaker and C. Sidner. Email overload: exploring personal information management of email. In *Proceedings of CHI'96*, pages 276–283. ACM, 1996.

[125] P. Willett. Recent trends in hierarchic document clustering: A critical review. *Information Processing and Management*, 24(5):577–597, 1988.

[126] J. A. Wise, J. J. Thomas, K. Pennock, D. Lantrip, M. Pottier, A. Schur, and V. Crow. Visualizing the non-visual: Spatial analysis and interaction with information from text documents. In *Proceedings of the IEEE Symposium on Information Visualization (InfoVis'95)*, pages 51–58. IEEE, 1995.

[127] J. M. Wolfe. Visual search. In H. Pashler, editor, *Attention*, pages 13–73. Psychology Press, Hove, 1998.

[128] M. E. J. Wood, N. W. Campbell, and B. T. Thomas. Iterative refinement by relevance feedback in content-based digital image retrieval. In *Proceedings of ACM Multimedia '98*, pages 13–20. ACM, 1998.