Mixing Research Methods in HCI: Ethnography Meets Experimentation in Image Browser Design

Ormerod, T.C.[1] Mariani, J. [1] Morley, N.J.[1] Rodden, T.[2] Crabtree, A. [2] Mathrick, J. [2]
Hitch, G. [3] & Lewis, K. [3]

[1] Lancaster University, Lancaster, LA1 4YD, UK
{t.ormerod;j.mariani; nicki_morley}@lancaster.ac.uk
[2] Nottingham University, Nottingham, NG8 1BB, UK
{tar; a.crabtree; jym}@cs.notts.ac.uk
[3] York University, York, YO10 5DD, UK
{g.hitch;k.lewis}@psych.york.ac.uk

Abstract. We report the specification and evaluation of a browser designed to support sharing of digital photographs. The project integrated outcomes from experiments, ethnographic observations, and single-case immersive observations to specify and evaluate browser technologies. As well as providing and evaluating new browser concepts, a key outcome of our research is a case study showing the successful integration of ethnography and experimentation, research and design methods that are often viewed as orthogonal, sometimes even mutually exclusive, in HCI.

Keywords: Ethnography, controlled experimentation, digital photographs, browser design and evaluation,

1. Introduction

1.1 Methods for Specifying Technologies

In the search for appropriate ways to specify and evaluate user-centered technologies, researchers and developers are increasingly turning away from laboratory-based controlled interventions towards more contextually-rich methods for studying user behaviours. This shift is exemplified by the emergence of ethnography as a method for informing systems design [1, 2]. Ethnography offers a non-invasive approach to observing rich social interactions around technologies in-situ. The approach facilitates the recognition of important exceptions and exemplars that inform technologies for supporting best practice, as well as revealing common patterns of activity. The shift in methods has partly been at the expense of controlled experiments that sacrifice detailed description of context and outliers in favour of factorial descriptions of user activity patterns. Indeed, proponents of ethnography [3, 4] cite limitations of experimentation as a key motivator for adopting an ethnographic stance.

Despite the advantages that accrue from ethnography, there is still a role for controlled empirical methods. Ball & Ormerod [5] point to the need for verifiability of observations to justify investments in technology, and the need for specificity and goal-directedness to focus upon the design imperative, as key reasons why designers need to supplement ethnographic data with controlled empirical studies. A further reason comes from the fact that people, both users and observers, are not always aware of or able to report the processes that influence their behaviour [6]. Hypothesis-driven experiments can reveal implicit influences on behaviour that affect user activities with information technologies.

Digital photography provides a domain that illustrates the relative merits of ethnographic and experimental approaches. Photographs are inherently social artifacts: the reasons for taking pictures, the uses we put them to, and the ways in which we handle, store and reveal them are guided by the context of use. To specify technologies for digital photography without conducting some form of ethnographic study risks underestimating the complex social activities that surround image handling. Yet, the ways in which individuals categorise, remember and subsequently recall information about photographs will also play a key role in determining the success of image handling technologies. Like many aspects of human cognition, these memory-based processes are not easy to observe or report.

We have previously argued [5] that ethnographic methods can and should be combined with other research techniques to properly inform user design. Other exemplars of research programmes that mix experimental and observational methods (e.g., case studies) in HCI exist [7]: This paper focuses upon mixing experimentation with an ethnographic approach to design and evaluation in HCI. In the remainder of the paper, we report empirical studies that use three research methods to inform the design of image handling technologies, and the development of a photo browser prototype that reflects the findings of these studies. The studies used *experimentation* to investigate the feasibility of interventions to reduce collaborative inhibition, *ethnography* to identify natural categories of shared encoding cue, and a detailed *case observation* to validate the feasibility of our chosen encoding approach. Evaluation of the browser again used experiments to assess the relative strengths of a prototype photo browser against a commercial alternative.

1.2 Digital image handling

There is a growing shift from chemical to digital photography, with mass-market and low-cost technology becoming commonplace within homes and families. As the digital camera grows in popularity, the number of images that individuals and groups store and handle can increase dramatically. An important consequence of digitalization is that photographs lose their physical availability. Physical artifacts provide retrieval cues for photographs (e.g., ‘the shoe box under the bed full of wedding photographs’) that are lost in digitalization [8]. From a situated perspective, methods for sharing non-digital photographs are central to how they are used. For example, traditional photograph albums serve as a constructed way of sharing information, often representing a collective familial resource. Methods for sharing images are likely to change greatly when photographs are stored on computers. Internet-based image transfer opens up new opportunities to share photographs across virtual communities, changing the nature of image communication and ownership in as yet poorly understood ways.

A number of different forms of software exist to manage digital images. Many commercial and research applications offer single-user query-based approaches to retrieval, with commands based on filename (i.e., a name of a photograph), user fields and keywords assigned by the user to photographs. Commercial browsers focus upon management of disk space for storing images (e.g., Thumbplus, Jasc). A number of research projects have also examined human-centred issues in image handling. For example, the Maryland PhotoFinder project [9] offers a browser for personal image management that supports encoding and retrieval through novel interface features for Boolean searches and visual overviews of search match results.

Other projects have focussed upon image sharing. For example, the Personal Digital Historian (PDH) is a table-based environment around which users collaborate to construct stories around a set of images [10]. One interesting feature of the PDH is the use of an image categorization scheme based around four dimensions that describe who the image pertains to, what the subject of the image is, where it was taken, and when it was taken. User selections under each dimension are combined automatically, providing an innovative solution to problems associated with constructing Boolean searches. Intuitively, the 'Who, What, Where and When' scheme captures the main episodic dimensions associated with the event portrayed by an image.

1.3. Psychological studies of memory

Studies of autobiographical memory suggest that 'Who, What, Where and When' dimensions play a key role in remembering. For example, Wagenaar [11] kept a diary in which he noted personal events over a period of some years. Subsequently he tested his ability to recall details of individual events by cuing himself with features such as who was involved, what happened, where and when the event took place or combinations of these cues. Among his findings were that 'when' is a poor cue and that combinations of cues are in general more effective than single cues.

There are other aspects of psychological research into human memory that might inform the development of image handling technologies. For example, a number of studies have demonstrated an effect of *collaborative inhibition*. In these studies, participants learn items individually, and subsequently recall the items either collaboratively (e.g., in pairs) or on their own. The effect is demonstrated when the total number of unique items recalled by groups is less than that recalled by nominal pairs made up of individuals recalling on their own [12]. The locus of the effect appears to be at retrieval: cues reflecting the subjective organization that one individual imposes upon information at encoding inhibit the subjective organization of a collaborating individual and so suppress their recall contribution [13]. If individuals who recall together have also encoded together, they tend to share the same subjective organization of the material, and an effect of inhibition is not found [14]. Collaboration at encoding reduces the incompatibility between cues generated by one individual and the subjective organization of the other individual. Technologies for sharing images that organize encoding and retrieval around individuals' categorisation preferences may provide precisely the conditions under which collaborative inhibition arises. The corollary to this argument is that image-sharing systems need to provide dimensions for encoding images that are common to collaborating users.

2. Experimental manipulations to reduce collaborative inhibition

The collaborative inhibition effect presents a challenge to the development of image handling technologies, since it suggests that an individual's organization of information at encoding may inhibit later retrieval of the same information by others. To address the problem, it was necessary first to find further evidence that collaborative inhibition effects can be reduced by appropriate interventions. If the effect arises because individuals impose different subjective organizations at encoding, then eliciting shared encoding categories might ameliorate the effect. Below we describe one experiment that investigated how self-determined categorization influences collaborative recall of image categories. It tested a prediction that partners who organise material similarly will show less collaborative inhibition than those who organise differently.

2.1 Method

Participants. Eighty undergraduate students from York University were paid £10 each to take part.
Design and materials. Participants were assigned to one of two groups, comprising either nominal pairs or pairs who collaborated at retrieval. Nominal pairs were made up by combining data from participants recalling alone to allow comparison with collaborating participants. Each of these groups was further divided, participants being paired with a partner who generated either the same or different categories when encoding the materials. Materials consisted of image labels of famous people (Elvis Presley, Margaret Thatcher, Britney Spears, etc.), which could be organised along various dimensions (e.g., gender, occupation, country).
Procedure. Encoding and retrieval phases were separated by approximately one week. In the encoding phase, participants sorted word sets into two self-determined categories. In the recall phase, participants recalled word sets collaboratively or alone (for nominal pairs).

2.2 Results and discussion

Figure 1 illustrates the recall performance of each group. A two-way analysis of variance on these data showed significant effects of type of pair (nominal versus collaborating), F(1, 36) = 37.0, MSe=4.11, p<.01, and of coding category (same versus different), F(1, 36) = 6.4, p<.01. Most importantly, the interaction between these factors was significant, F(1, 36) = 6.4, p<.01. These results indicate that, while collaborative recall by pairs with the same encoding categories (17.3/40 items) was similar to nominal pair recall with both same and different encoding categories

(19.6/40), collaborating pairs who had different encoding categories showed the effect of collaborative inhibition (14.1/40).

A second experiment examined whether the same effects are found when the dimensions for sorting are imposed externally. The stimuli comprised words that could be organised into three-member groups, either associatively (e.g., shepherd, sheep, wool) or categorically (e.g., shepherd, chef, fisherman). Participants sorted items associatively or categorically. Individual recall was unaffected by sorting associatively or categorically. Collaborating pairs who sorted items according to different criteria recalled less (29/45 items) than nominal pairs (33/45 items). In contrast, collaborators who encoded items according to the same criteria showed no inhibition (34/45 items).

These experiments suggest that methods to increase the similarity of subjective organizations that individuals bring to encoding information will enhance collaborative retrieval. A reduction in collaborative inhibition was found both with explicit presentation of organizational schemes at encoding and when individuals with self-determined schemes were paired with like-minded participants. However, the experiments leave open the question as to which category labels might suit image sharing best. It appeared, from the results of both experiments, that there is no one semantic dimension that is superior to any other in enhancing retrieval. Thus, in the next phase, we turned to ethnographic studies to investigate whether natural accounts of image sharing yield dimensions appropriate for instantiation within image handling technologies.

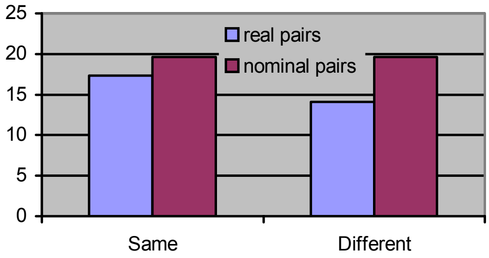


Fig. 1. Recall by collaborating and nominal pairs, sorts by partner having same or different categories.

3. Ethnographic studies of families and photographs

We undertook ethnographic studies of how photographs are handled and involved in everyday activity across a number of families. The studies build upon the work of Frolich et al [8], who used home-based interview and diary-keeping methods to examine how families manage photographs. Among the important observations made by Frolich et al was the multiplicity of archiving approaches adopted (e.g., special project mini-albums), and the social nature of co-sharing of physical photographs, a process that was not easily supported by digital media. The aim of our studies was to provide a broad background for on-going experimental investigations, illustrating the different forms of interaction that surround photographs within the home. Below we offer specific examples of issues that informed the refinement of an encoding approach within the TW3 browser prototype.

Photographs differ from other forms of record because of the cultural significance of photographs within family life. Perhaps the most significant thing to note is the ways in which photographs find their way into the set of everyday activities central to our family lives. One of the most visible aspects of photograph use in the home is the symbolic and decorative role they assume. Photographs of family members in particular are displayed around the home in prominent positions. They recall people that are important to us, significant events in our lives, places that visit and memories of past times.

The framed photographs made visible in our homes provide a public display of our family lives and the episodes that make up the family history are often placed on displace for public inspection. These photographs fine their way into the everyday fabric of our home. Figure 2 exemplifies the everyday settings within which photographs are routinely placed. With one family group we studied, photographs were kept in boxes, bags, and albums according to the *significance of particular ensembles*:

1. Pictures of a family wedding were kept in simple but ornate boxes.
2. Pictures of the householder's own wedding were kept in specially made album, which in turn was kept inside a white cloth cover to protect the album.
3. Pictures of children over the years were kept in another album.
4. An ongoing project (a photographic family tree) was kept in a folder of plastic wallets inside a shopping bag underneath the cupboard 'ready to hand'.

The storage of photographs may seem haphazard, but it is possible to detect an organizing principle informing storage. Thus, wedding photos are kept in formal albums, pictures of a child over the years in a less formal, more sentimental album, pictures of another's wedding in simple decorative boxes, whereas ordinary photos are left in the packing they came in and may be thrown together in a large box, ongoing projects might be placed in a plastic bag, and so on. Each of these concrete storage arrangements reflects, for members, an order of significance such that the meaning of any particular ensemble can be seen-at-a-glance. Some orders of significance are thoroughly social; the use of special wedding albums is widespread for example, whereas others, such as storing photos of special occasions in simple but decorative boxes, are more personal and idiosyncratic.

By inference, one can interpret the arrangements of use we have observed as a physical instantiation of implicit categorization by Who, What, Where and When dimensions. However, the conceptual separators underlying these physically separate collections map onto Who, What, Where and When dimensions in interesting ways. For example, some events are clearly demarcated by all four dimensions (e.g., picture of a recent family celebration such as a Christening). Others lose one or more dimensions as organizing principles (e.g., collections of photographs of children over the years).

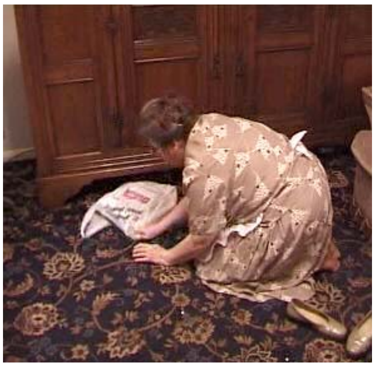


Fig. 2. Photographs retrieved from hiding place.

The majority of photographs, rather than being on public display, are brought out to be shown to visitors and friends, and in the showing to be used to explain the events surrounding then. A family member who puts the photographs away normally mediates this process. For example, in Figure 2, we see a collection of photographs (kept in a plastic carrier bag) being retrieved. Once retrieved from their normal place of storage, broad collection becomes a resource at hand to support the telling of stories.

Analysis of conversations shows how identifying the 'Who' of a photograph is built up from the physical manipulation of artifacts and from an emerging interactive discourse that relies on a specific family member, the mediator, to supply the recognition information, with new participants being drawn into the discourse as it unfolds. A unifying feature of the studies is the emphasis upon collaborative descriptions of images. What matters is not the taxonomic status of an image (as investigated in the experimental phase) but its situated characteristics, in terms of time, place, and involvement of people. These episodic cues are drawn upon as part of the storytelling surrounding the presentation of photographs across a grouping. This emphasis upon episodic descriptions is similar to that which is apparent in Wagenaar's [11] study of autobiographical memory.

4. Single-case observation of image encoding and retrieval

We conducted an in-depth study of the efficacy of a category scheme for photograph collections for one individual. The aim of the study was to validate design hypotheses for image browsers, notably the usability of a Who, What, Where, and When encoding and retrieval scheme. The study addressed three questions: first, can these dimensions be used effectively, and, in particular, how efficient is encoding? Second, do the categories discriminate well among items within a personal photograph album? Third, do the dimensions provide sufficient cues at recall?

The study focused upon the photograph collection of a married couple. The male member of the couple provided access to, and an overview of, a large set of photographs collected both before and since marriage. In the encoding phase, we elicited descriptive categories from his partner for 200 photographs selected from this collection. She then sorted photographs into categories under each of the Who, What, Where, and When dimensions. A week later, the participant gave each of the photographs a title.

Results of the encoding phases showed that sorting under the scheme was meaningful to the participant. The participant spontaneously chose no more than six categories on each of the four dimensions, with some overlap of subcategory label between different dimensions. Measures of fan size (the number of photographs that received exactly the same categorical assignment under the four dimensions) varied over the photographs, reflecting marked asymmetries in the use of the coding space (see Figure 3). In essence, the majority of photographs were categorised uniquely under the four dimensions, though some instances of large sets (up to 23 photographs) received identical categorisation under all four dimensions.

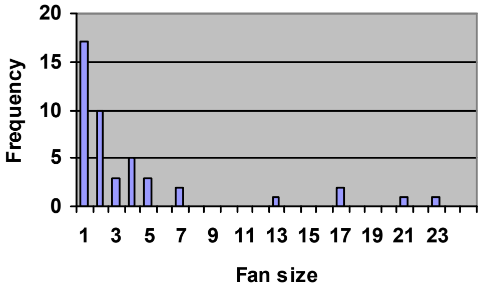


Fig. 3. Fan size during encoding phase (= no. of images encoded with same categories under Who, What, Where, and When dimensions; Frequency = instances of each fan size).

In the retrieval phase, four different procedures were used to vary retrieval cue and task (recall of photograph codes or titles versus recognition of photograph). Each procedure was evaluated using a different set of 24 photographs with varying fan sizes. Comprehensive *recall* of titles was poor (25% correct), as was recall of the codes used for each photograph (accurate recall of all 4 subcategories for only 54% of photographs). However, individual dimension recall was good (averaging 3 subcategories per photograph). Furthermore, code *recognition* was high (86% of photographs had all four codes accurately recognised). Overall the results suggest that the coding scheme was effective for recognition-based retrieval. Importantly, many of the errors in the retrieval phase were errors of commission (i.e. the participant including known photographs in her recall that were not among the 24 target items).

In summary, the case study provides some supportive evidence for a Who, What, Where and When scheme at both encoding and retrieval. The implication of the fan size results is that a browser must offer a categorization scheme that is extremely flexible, because the majority of photographs receive a unique categorisation. A two-level scheme such as that used in the case study, in which up to six categories are created under each dimension, allows 4^6 (or 4096) unique categorizations. Whether this space is sufficient to capture a large image set depends upon the extent to which images can be meaningfully categorized together. Further work is in progress to investigate the efficacy of the scheme for image sets of 1000+ that come from multiple sources (photographs from a decade of news articles).

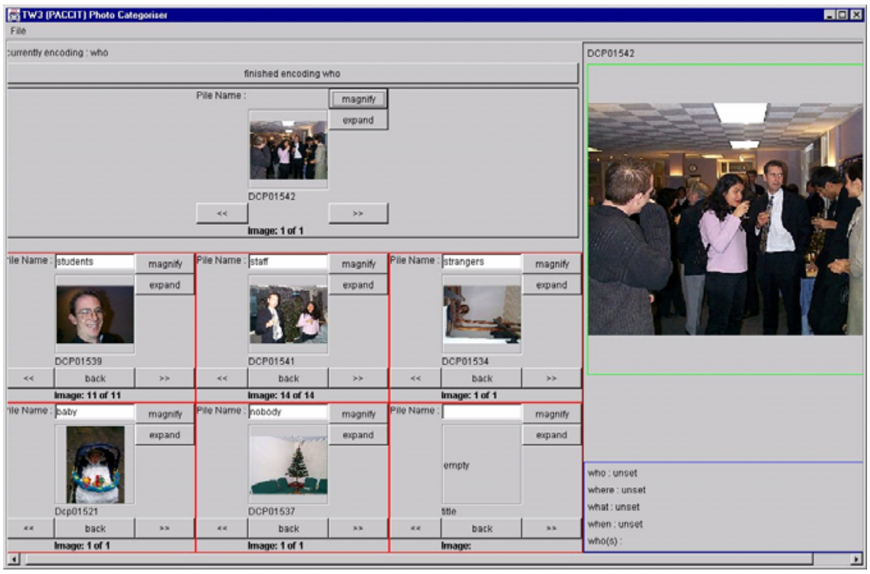
When errors were made they were errors of commission. The implication for a browser is that if only one sorting code is incorrectly recalled, the target photograph will not be found. However, we found that if any one of the four codes was ignored, a larger but manageable set of photos was retrieved with a high probability of containing the target. This pattern suggests a two-stage browser search mechanism in which the user can enter partial cues when not all of them can be remembered, and then visually scan the resultant set of retrieved photos for the target. A further implication is that while collaborative users will share a generic 'what' where' 'when' 'who' organizational scheme, they will typically differ in the categories they use within this shared scheme. We hypothesize that limiting categorization to four key dimensions, each with six categories to be specified by the user at encoding, will maximize the degree of overlap across the subjective organizations of multiple users. Where category systems differ among users, or where the search under four dimensions fails to yield a result, the gradual removal of one of the four dimensions will increase the degree of similarity among coding schemes and allow users to recover items for which one or more of the encoding categories has been forgotten.

5. The TW3 browser prototype

The case study provided validation for the use of an episodic organization scheme based around Who, What, Where, and When dimensions. In principle, there are a large number of ways in which such a scheme might be delivered within a browser, and the remainder of the TW3 project is exploring how these approaches might be optimized. The first prototype embodies the scheme explicitly as a procedural encoding and retrieval task.

The prototype is implemented as a Java point-and-click interface to a MySQL database. Figure 4 illustrates the encoding interface. The TW3 browser requires users to work through categorization under four dimensions. The user can code all photographs under one dimension at a time, or code each photograph under all four dimensions in parallel. Usability tests to date suggest that users require the capability to switch between encoding modes in real time during a single encoding run. Initially they typically choose to step through categories one by one. Once categories under each dimension become stable, however, some users prefer to switch to a mode of encoding each photograph under all four dimensions at once.

The retrieval mode uses the category structure created at encoding as cues to guide photographic description under each dimension. We make no use of user-assigned descriptive titles or keywords, since the case study pointed towards the inadequacy of labeling or keyword approaches. Moreover, early file-naming studies showed that file-naming even among experts yielded little consistency [15], a finding echoed by our own results in the experiments reported above, suggesting that keyword and label approaches will not support collaborative retrieval.



currently encoding : who

Pile Name :
DCP01542
Image: 1 of 1

Pile Name : students
DCP01539
Image: 11 of 11

Pile Name : staff
DCP01541
Image: 14 of 14

Pile Name : strangers
DCP01534
Image: 1 of 1

Pile Name : baby
Dcp01521
Image: 1 of 1

Pile Name : nobody
DCP01537
Image: 1 of 1

Pile Name :
empty
title
Image:

DCP01542

who : unset
where : unset
what : unset
when : unset
who(s) :

Fig. 4. The TW3 encoding tool. Photographs are presented as a stack (top center) ready for classification. The user categorizes under Who, What, Where and When dimensions in turn. The user can assign photographs to up to six categories for each dimension. Photographs can be magnified and categories expanded overview membership.

Items are retrieved according to their degree of fit with the categories under Who, What, Where, and When dimensions. If the target photograph remains undetected, the user can step through the dimensions, investigating the effects of removing each dimension in turn. By expanding on the retrieved sets with one dimension missing, the user is able to see a 'best fit' selection and discover the missing picture. In this way, the scheme allows an option to use partial encoding cues that are likely to offer a close match to the target items. In this respect, our use of a Who, What, Where and When scheme differs from that of Shen et al [10], who manipulate each dimension separately. Wagenaar's [11] results suggest that additional power for retrieval might gained by allowing the user access to these dimensions in parallel, and that the systematic dropping of dimensions that are uninformative at recall can guide people to the correct target set.

Perhaps the key difference between the TW3 prototype and other (e.g., commercially available) browsers is in the role of *constraint*. For example, other browsers tend to allow unlimited expansion of coding dimensions and categories (e.g., using a folder and sub-folder metaphor, labeling individual photographs with category tags), whereas the TW3 browser constrains encoding to four dimensions, and allows only six categories under each dimension. Also, because encoding is relatively unconstrained in other browsers, there is no restriction on the kinds of dimensions that users may use: they are just as likely to classify photographs semantically as episodically. In contrast, the Who, What, Where and When approach of the TW3 browser effectively constrains the user to an episodic category scheme. Moreover, the ways in which users retrieve photographs in other browsers is typically unconstrained: users can search for named photographs by keyword, or add and change as many label tags to photograph searches that they wish, when they wish. In contrast, to retrieve a photograph in the TW3 prototype, users must select categories under each of four dimensions. If the required photograph is not found, users are constrained to dropping one dimension at a time.

While this level of user constraint is uncommon (indeed, arguably, it is generally frowned upon) in user-centered design, we hypothesise that it might prove crucial to successful sharing of digital images. For example, constraint on encoding increases the relative likelihood and degree of overlap between different peoples' subjective organizations of photograph collections. Also, the inclusive use of all four dimensions during retrieval, followed by their systematic removal to continue to search, provide a procedural structure to guide the process of recovering from error (i.e., knowing what to do next if your first attempt does not yield the desired photograph).

6. Experimental evaluation of the browser prototype

The TW3 browser prototype reflects a number of design hypotheses and assumptions. Perhaps the most fundamental assumption is the one derived from the psychological literature on collaborative remembering, namely that there might be a problem in retrieving photographs that are stored under someone else's coding categories. Then there is the issue of the Who What Where and When coding approach itself – it offers commonality between individuals at the level of dimensions under which categories are specified, but it is not clear whether this will hinder or help the process of photograph encoding and retrieval relative to browsers that do not fix the dimensions under which individuals categorise photographs. Another hypothesis concerns the restriction to six categories under each dimension. This limit was based upon empirical observation, yet its effects on browser performance cannot be easily predicted.

One approach to evaluating the prototype might be to employ an ethnographic approach, situating the browser in, say, a family context and observing over a number of weeks or months how peoples' activities around photograph handling are supported or changed by the imposition of the new technology. Indeed, we are adopting this approach in studies currently in progress on a substantially revised second prototype. However, we chose in the first instance to conduct a controlled experimental evaluation of the browser prototype, for three main reasons. First, an experimental evaluation allowed us to collect comparative data that pits our prototype against a commercially available browser, in this instance, the Adobe™ Jasc browser. Second, we were concerned that a situated evaluation of the browser might provide an unduly negative outcome for the simple reason that the TW3 browser was an early prototype with all the lack of functionality and irritations that early prototypes tend to have. In particular, we felt that users would be likely to abandon use of the browser prematurely, regardless of any merits that its key design features might bring, simply because of fixable prototype limitations. Third, we wanted to investigate whether the browser does address problems of shared encoding and retrieval using measures of search and retrieval which would simply not be observable using ethnographic methods.

The comparison between TW3 and Jasc browsers is not intended to be simply one assessing relative performance: we confidently expected the Jasc browser to outstrip our prototype on a majority of performance measures, if only because it is a properly-tested and fully-functional piece of commercial software developed for market by a team of designers, programmers, and testers. We were interested only in how the TW3 prototype compared with the Jasc browser in terms of *change* in performance, both across conditions (notably, when retrieving from ones own codes compared with retrieval using someone else's codes) and within conditions (notably, how the browsers fared in terms of recovery from failure to find photographs). In some respects, one might not expect major differences between the two browsers. In particular, the Jasc browser comes with three pre-configured tag dimensions, of People (i.e. who), Event (i.e., what) and Place (i.e., where), with only the time-based tag missing. Where differences emerge, they must then reflect user preferences to make use of the freedom within Jasc to create their own categories and ignore system-set ones.

6.2 Method

Participants. 28 undergraduate and postgraduate students from Lancaster University were paid £10 each to take part.

Design and materials. Materials consisted of 200 photographs of members of the British royal family or places and events relating to them, gathered from a trawl of Internet media sites. Participants were assigned to one of two groups. One group used the TW3 browser to encode and retrieve photographs, the other used the Adobe Jasc browser (the free demonstration version available on the Adobe web site). For the retrieval phase of the experiment, each participant was nominally paired with another participant from the same group, matched by average encoding time. A second (within-subjects) factor in the retrieval phase was whether participants retrieved photographs using their own codes or those of their nominal pair.

Procedure. Encoding and retrieval phases were separated by approximately one week. In the encoding phase, participants were first shown all 200 photographs at a rate of 2 seconds per image. They then encoded each of the 200 photographs. For participants using the TW3 browser, they coded each photograph in a category under each of the four dimensions before proceeding to the next photograph, the categories (maximum = 6) emerging during the encoding process. For participants using the Jasc browser, they encoded each photograph by assigning either system-set or new tags (i.e., category labels). In the retrieval phase, participants retrieved 30 photographs using their own codes and 30 different photographs using their nominal pairs codes. Each photograph to be retrieved was presented on paper, and the participant's task was to find the photo in the browser by selecting categories under each dimension (TW3) or tag sets (Jasc).

6.3 Results and discussion

The average time taken to encode each image was significantly greater with the Jasc browser (38.4s) than with the TW3 browser (20.6s), t=7.85, p<.01. The fact that encoding times were nearly twice as long with the Jasc browser is probably a function of the number and complexity of tags assigned to images compared with the limited categories used with the TW3 browser.

Table 1 shows the average number of tags/categories created under each dimension. Interestingly, tags under the Event and Place dimensions created with Jasc are comparable, quantitatively at least, with those created under What and Where with the TW3 browser. The People dimension appears to have been encoded at a much greater level of detail with Jasc than with TW3. This may result from the use of multiple overlapping tags in Jasc (e.g., “Charles”, “Diana”, “Charles with Diana” as separate categories), a strategy that is effectively blocked by the category limit within TW3. The ‘other’ dimension of Jasc is not comparable with the ‘when’ dimension of TW3, since the former refers to all tags created by the user that did not fall within the system-set dimensions whereas the latter refers to the time dimension. What is clear is that users were making use of the flexibility inherent within Jasc to create many personalized coding categories.

Table 1. Mean number of tags/categories created under each dimension using Jasc/TW3 browsers at encoding.

	Who/Person	What/Event	Where/Place	When/Other
TW3	6.0	5.6	5.2	4.2
Jasc	24.2	6.2	6.9	19.3

Table 2 shows retrieval performance with the two browsers under a number of measures. A significant interaction was found between Browser and Code factors in the number of photographs retrieved at the first attempt, F(1, 26) =8.94, MSe=5.61, p<.01. The Jasc browser gave the highest level of retrievals at the first attempt, particularly with own codes. This result suggests that, as long as you find a photograph first time and you are the sole user of a collection, the Jasc browser is the better of the two.

A significant interaction was also found for the number retrieved overall, F(1, 26) =9.87, MSe=6.09, p<.01. It appears that, while there is no advantage for either browser when retrieving using ones own codes, the TW3 browser leads to greater retrieval using someone else’s codes. Indeed, performance is comparable with using ones own codes with the TW3 browser. Thus, the main advantage of the TW3 browser appears to be in recovering from a failed first attempt to find a photograph using someone else’s codes.

A main effect of Browser was also found with retrieval times, F(1, 26) =5.44, MSe=209.8, p<.05, though the interaction between Browser and Code factors was not significant. It seems likely that the advantage for the TW3 browser is a result of different strategies for finding a photograph after a failed first attempt. With the TW3 browser, users were limited to dropping each dimension in turn in order to inspect whether the required photograph had been mis-categorised or mis-recalled under that particular dimension. With the Jasc browser, users were also able to drop tags, but a much more common strategy was to add another tag in order to combine the results from tag categories. As well as taking longer to execute, this strategy was limited in effect. While it could deal with errors of omission (photographs not classified under a particular tag dimension), it was less successful in dealing with errors of commission (i.e. photographs wrongly classified or mis-recalled under a particular tag dimension).

Table 2. Mean number of photographs retrieved (N = 30) on first attempt, and overall (i.e. after dropping categories or adding extra tags), and mean time to retrieve image.

	No. found at first attempt	No. found overall	Mean retrieval time (s)
TW3 with own codes	14.4	24.4	35.5
TW3 with others codes	10.7	23.2	36.8
Jasc with own codes	18.2	23.6	40.9
Jasc with others codes	10.8	18.4	47.0

The results of the study confirm our key hypotheses. First, there is a detrimental effect of trying to retrieve photographs using another persons coding scheme. This result is not surprising in theoretical terms, but it has important practical implications for the design of collaborative browsers. Second, the Who, What, Where and When scheme seems to provide an efficient and effective set of dimensions and procedure around which to configure a browser. The study is, of course, limited to a particular observation and set (and size) of materials. It may be, for example, that a less favorable outcome would be found with less familiar materials (e.g., archeological shards) and with larger sets of photographs, especially when they are encoded over a longer and more fragmented time frame.

Of key importance, it appears that the two browsers are optimized for different contexts of use. The Jasc browser appears best suited to individual users maintaining photograph collections for private use, where they can code photographs in uniquely meaningful ways. In line with our hypotheses, the TW3 browser appears to be better configured to support collaborative use of photographs. While first-attempt retrieval is perhaps disappointing with the TW3 browser, recovery is as strong as with the Jasc browser using ones own codes, and more importantly, it is much better when using someone else’s codes.

7. Conclusions

The design of the TW3 prototype was informed by converging results from three empirical methods that are often seen as diametrically opposed to each other. However, we argue that each can offer an essential and unique contribution to systems design. The experiments demonstrated the potential for categorization-based interventions to enhance collaborative retrieval. The brief sample from a longer ethnographic study highlights the point that photographs are routinely viewed as part of a collaborative set of activities and are used to support a broader set of social activities across the family. The case study showed how a four-dimensional scheme can offer a simple yet powerful approach to encoding and retrieving digital images. The case study also illustrates how methods used in experimental studies can be applied in more naturalistic and rich observational studies.

These ideas have come together within a set of image browsing tools that allow users to collaborate in encoding and retrieving images while supporting them in overcoming a major source of difficulty, namely errors of commission. The aim is to develop equivalents of social discourse around images for digital technologies. While researchers have explored the development of different presentation techniques for this purpose [16, 17], we are more interested in how digital photographs will be stored and retrieved as part of this process.

Experimental demonstrations of collaborative inhibition point to a phenomenon that must be addressed in all systems designed for collaborative use. The ethnographic studies provide support for an episodic approach to collaborative encoding and retrieval. The dominance of episodic discourse around photographs is consistent with results from the case study, notably the finding that recall of photographs by semantic keyword was very inefficient compared with recall by episodic category. This finding suggests that query-based approaches are of limited efficacy in managing large image sets, and do little to address problems of collaboration.

The importance of understanding contexts of use is emphasized by the results of the comparative evaluation, where it appears that the Jasc browser is optimized for individual use while the TW3 browser is better for shared use (albeit tested here in a context where users worked individually with codes produced by a nominal partner). As one encounters other contexts of use, this pattern might change. For example, it is possible that in professional contexts (e.g., commercial photo libraries), the advantages of detailed coding of individual photograph characteristics may outweigh the benefits of a restricted coding scheme.

The studies reported here show how different methods make valuable contributions to the design and evaluation of interactive systems. In planning empirical studies that inform design, there are competing pressures. The need for ecologically valid observation or real contexts of use must be balanced against the efforts required to collect data and the costs of early commitment to prototypes that can be evaluated in-situ. At the same time, there must be a recognition that no single method can provide everything a designer needs. Our mixed method approach allows both situated observation of contexts of use and also detailed assessment of the impacts of cognitive phenomena that are otherwise hard to observe and measure.

Acknowledgements

The TW3 project is supported by the ESRC/EPSRC ‘People at the Centre of Communications and IT’ initiative, L32830300198. We thank Fleur Finlay and Rachel Attfield for help with the experiments.

References

1. Hammersley, M., Atkinson, P.: Ethnography: Principles in practice. Routledge, London (1983)
2. Hughes, J.A., King, V., Rodden, T., Andersen, H.: Moving out from the control room: Ethnography in system design. In Proc. CSCW ’94, Chapel Hill, North Carolina (1994)
3. Hutchins, E.: Cognition in the wild. Cambridge, MIT Press, MA (1995)
4. Suchman, L.: Plans and situated actions: The problem of human-machine communication. CUP, Cambridge (1987)
5. Ball, L. J., Ormerod, T. C.: Putting ethnography to work: The case for a cognitive ethnography of design. Int. J. Human-Computer Studies 53 (2000) 147-168
6. Nisbett, R., Wilson, T.D.: Telling more than we can know: Verbal reports as data. Psychological Review. 84 (1977) 231-259.
7. Murphy, G.C., Walker, R.J., Baniassad, E.L.A.: Evaluating emerging software technologies: Lessons learned from assessing Aspect-Oriented programming. IEEE Trans. on Software Engineering, 25 (1999) 438-455
8. Frolich, D., Kuchinsky, A., Pering, C., Don, A., Ariss, S.: Requirements for photoware. Proc. CSCW 2002, New Orleans, ACM Press (2002) 166-175
9. Kang, H., Shneiderman, B.: Visualization methods for personal photo collections, Proc. IICME 2000, New York: IEEE Computer Society (2000)
10. Shen, C., Lesh, F., Vernier, F., Forlines, C. Frost, J.: Sharing and building digital group histories. Proc. CSCW 2002, New Orleans, ACM Press (2002) 324-333
11. Wagenaar, W. A.: My Memory: A study of autobiographical memory over six years. Cognitive Psychology, 18 (1986) 225-252
12. Weldon, M.S., Bellinger, K.D.: Collective memory: Collaborative and individual processes in remembering. J.Exp Psych: Learning, Memory & Cognition, 23 (1997) 1160-1175
13. Basden, B. H. Basden, D.R., Bryner, S., Thomas, R.L.: A comparison of group and individual remembering: Does collaboration disrupt retrieval strategies? J.Exp Psych: Learning Memory & Cognition, 23, (1997) 1176-1191
14. Finlay, F. Hitch, G., Meudell, P.: Mutual Inhibition in collaborative recall: Evidence for a retrieval-based account. J.Exp Psych: Learning Memory & Cognition 26 (2000) 1556-1567
15. Furnas, G.W., Landauer, T., Gomez, L. Dumais, S.: Statistical semantics: analysis of the potential performance of keyword systems. Bell Systems Technical Journal, 62 (1983) 1753-1806
16. Balobanovic, M. Chu, L.L., Wolff, G.J.: Storytelling with digital photographs, Proc. CHI 2000, Amsterdam, ACM Press (2000) 564-571.
17. Vernier, F., Lesh, N. Shen, C.: Visualisation techniques for circular tabletop interfaces. AVI 2002, Trento, Italy, ACM Press (2002)

Discussion

[Michael Harrison] About titles and their semantics. What does it mean to fail to get the semantics right?

> [Tom Ormerod] Both recall and recognition of photo titles were very poor. Elements of the description didn't match more than 50% of the titles.

[Bonnie John] Are a lot of your results because of specific features of the photos you used? E.g., Relatively few (hundreds not thousands). Maybe the six categories is just because there are so few, which would be different if there were a lifetime of photos.

Not many that are actually photos of the same thing (e.g., the professional photographer did more of the exact same labeling, perhaps because professionals take many of the same thing, so why wouldn't there be the same label? -- and as people understand that digital cameras don't waste film, they'll take many of the same thing, too.).

> [Tom Ormerod] That's what I was trying to say on the last slide -- we don't know the exact locus of the effects we report. However, we have ongoing work with professional image colelctions where volumes are 20000 images plus. So far, results are promising.

[Hong-Mei Chen] Do you intend to generalize your research results beyond the family photo retrieval system to a general image retrieval system?

> [Tom Ormerod] Yes. We are currently exploring possibilities such as PDF file retrieval.

[Hong-Mei Chen] I think it may have some difficulties as family photos, as Bonnie pointed out, may have a lot of similar photos and the precision of retrieval may not be as critical as other applications such as document retrievals.

In addition, in your experiment, you used the British Royal family photos instead the subjects' own photos, that may affect your experimental results applicable to family photo retrievals as most people have intrinsic memories associated with their own photos.

> [Tom Ormerod] I don't really have answers to the first part of this question. However, we did an experiment looking at couples who handled their own photos, encoding either together or separately. To our surprise, we got similar effects with these personalised materials.

[Joaquim Jorge] Have you thought of methods for automatically capturing metadata ? People are not very adept at cataloguing photos and documents.

> [Tom Ormerod] Metadata can be re-used, e.g. when taking a series of photos on the same subjects. Also when temporal labels are very close the photos can "inherit" labels from others in the sequence.

[Joaquim Jorge] What about using "stories about photos" to create photo archetypes from those stories and extract content? Another possibility would be sketching descriptions for content-based retrieval?

> [Tom Ormerod] We have a different research agenda. We suspect that good browsers would do a little of both and minimize labeling problems.