Spatial querying for image retrieval: a user-oriented evaluation

Joemon M. Jose
School of Computer and Mathematical Sciences
The Robert Gordon University
Aberdeen, AB25 1HG, Scotland
www.scms.rgu.ac.uk/staff/jj/

Jonathan Furner
School of Information and Media
The Robert Gordon University
Aberdeen, AB24 4FP, Scotland
www.rgu.ac.uk/~sim/staff/jf/jf.htm

David J. Harper
School of Computer and Mathematical Sciences
The Robert Gordon University
Aberdeen, AB25 1HG, Scotland
www.scms.rgu.ac.uk/staff/djh/

Abstract Epic is an image retrieval system that implements a novel spatial-querying mechanism. A user-centred, task-oriented, comparative evaluation of Epic was undertaken in which two versions of the system—one set up to enable spatial queries only, the other allowing textual queries only—were compared. Use was made of the two systems by design professionals in simulated work task situations, and quantitative and qualitative data collected as indicators of the levels of users' satisfaction. Results demonstrated that users often had a 'mental image' of a potentially satisfying picture in mind, that they were happy to express this need in visual terms, and that in doing so they preferred to have access to Epic's spatial-querying facility. Success in obtaining statistically significant results appears to support validation of the novel methodological framework adopted.

1 Introduction

Awaiting the development of innovative techniques for image retrieval [9] [13] [15] are myriad applications in commercial environments. Spurred by the needs of this market, the quality and quantity of research undertaken in image retrieval is growing, drawing interest from research groups traditionally associated with a variety of related fields in the information sciences. Two matters of consensus in this community are:

- that effective retrieval of images is currently crucially dependent on the representative indexing of those images by content-based metadata (i.e., 'high-level', semantic features) [14]; but
- that the facilities provided for searchers of image databases to formulate queries should allow appropriate advantage to be taken of the necessarily visual (rather than textual) nature of the stored objects [1] [27].

A common assumption underlying the latter view is that some form of visual or spatial querying mechanism, inviting users to compose query-*images* rather than to specify query-*terms*, might reflect more accurately the operation of certain human cognitive processes. Few of the descriptions of experimental and operational systems that have appeared in the literature, however, have been accompanied by detailed evaluative studies of those systems.

We have been working on the design, development and implementation of a system called Epic, whose function is to assist users in the retrieval of photographic images from large collections, and which implements a novel spatial-querying mechanism [23] [24] [25]. We have latterly been concerned to test one or two assumptions that we had made in the design of the Epic interface:

- that searchers often have a reasonably well-defined 'mental image' of a picture that might satisfy their visual information need;
- that searchers are able to represent such a 'mental image' by drawing and labelling rectangles on a rudimentary electronic sketchpad; and
- that the provision of a facility allowing the specification of spatial queries in this way helps users to improve their searches, and thus to satisfy their information needs more effectively, efficiently and easily.

In pursuit of evidence to confirm or deny the validity of these assumptions, we have conducted an evaluative study [22] in which we compared two pared-down versions of the full Epic system: one stripped of its facilities for formulating textual queries (i.e., allowing spatial queries only), and another stripped of its facilities for formulating queries using spatial features (representative, therefore, of those systems that employ standard, keyword-based query-formulation mechanisms alone). Formally, our one-tailed experimental hypothesis was that the spatial-query system would be more acceptable or satisfying to the user than its more conventional counterpart, in any of several, varied respects. This primary hypothesis may be decomposed into a number of 'sub-hypotheses', each corresponding to one of the assumptions stated above, in accordance with the procedures suggested by 'Evaluation Light'—a set of loose guidelines developed by Harper and Hendry expressly to enable the efficient assessment of experimental interactive systems [18].

In this paper, we provide:

- in Section 2, an account of our motivations for an evaluative study, indicating the importance of evaluative studies to IR research, clarifying some of the terminology we will be using, and considering one aspect, highly relevant in the present context, of the 'paradigm shift' from a 'system orientation' to a 'user orientation' that is often identified as having recently occurred in the IR field [28];
- in Section 3, an overview of the salient features of the Epic system;
- in Section 4, a description of the methodology we used in our evaluation of Epic;
- in Section 5, a review of our experimental results; and
- in Section 6, a set of concluding remarks that we hope will be helpful not just for developers of image retrieval systems, but also for those concerned more broadly with issues relating to the evaluation of retrieval systems in general.

2 Evaluation in IR

The activity of evaluation has long been recognised as a crucially significant element of the process through which information retrieval systems reach implementation in a real-world, operational setting [16] [20] [30] [32]. Evaluative studies are concerned with assessment of the quality of a system's performance of its *function*, with respect (made explicit to a varying degree) to the needs of its users within a particular context or *situation*. The direction of such studies is commonly determined, and thus implicitly validated, by the adoption of some kind of structured methodology, or 'evaluative framework'. This serves to guide the researcher in decisions as to what *criteria*, *measures*, and *methods* of data collection and analysis to use, in what kind of experimental *setting*—and also to remind them of the importance of establishing a rigorous definition of the function of the system [11]. This definition will typically be couched in terms of the assistance or support the system provides to the human user in their performance of some *task*—some activity that a person engages in, in order to attain a particular desirable goal or change in some state of affairs.

The traditional framework for evaluative studies of information retrieval systems derives from the Cranfield projects in the early 1960s [6], and survives in the large-scale experiments undertaken annually under the auspices of TREC (see, for example, Harman & Voorhees [17]). Some essential characteristics of this framework are as follows:

- it assumes a low level of *interactivity* between the retrieval mechanism (or indeed the stored information) and the user;
- it controls environment variables through the use of experimental test collections and *pre-defined* queries;
- it establishes retrieval *effectiveness* as the primary criterion or dimension of performance; and
- it defines effectiveness with respect to the *relevance* of retrieved documents to requests, and establishes recall and precision as quantitative measures in this dimension.

On many occasions, and with increasing frequency over the past twenty years (see, for example, Hersh [19]), arguments in opposition to this framework have been rehearsed in terms similar to the following:

- in real-world retrieval systems, and especially in multimedia systems, the user maintains a high level of continuous control over the initiation, direction and termination of successive stages of the information-seeking process;
- real-world systems may be meaningfully evaluated only in real-world settings, where the researcher may observe the behaviour of real users, seeking to undertake real tasks through interaction with systems designed to provide appropriate support for those very undertakings;
- since recall and precision scores are quantitative indicators only of one dimension of system performance, reliance on these measures results in ignorance of other dimensions of system performance that are worthy, at the very least, of equal consideration—criteria such as *efficiency*, *usability* and *acceptability* (or level of *user satisfaction*); and
- once system effectiveness is defined more broadly, not merely as a simple function of the relevance of documents to requests, but as a criterion for assessing the whole outcome of a system's provision of assistance in the user's performance (or resolution) of some context-situated task (or problem), then it should be clear that recall and precision are themselves not even satisfactorily indicative of effectiveness.

The force of such arguments is reflected in the ongoing efforts of those involved in the 'interactive track' of the TREC experiments [2] or in the EC Working Group on the evaluation of Multimedia Information Retrieval Applications ('Mira') [8] [10], or those whose expertise ranges sufficiently widely to make explicit the applicability of related work in the cognate disciplines of natural language processing [12] and human-computer interaction [7]. Nevertheless, despite the frequency with which such arguments are stated, a standardised framework for the evaluation of *interactive*, *multimedia* retrieval systems comparable to that established by the Cranfield projects for *batch-mode*, *text* retrieval systems has yet to become widely accepted.

A review of published accounts of recent experiments (see, for example, [3] [4] [21] [33]), however, would seem to indicate that a consensus view of the essential elements of such a framework is currently evolving. Consideration of these studies demonstrates a common concern with

- meaningful evaluation of the whole range of a user's interaction with systems that allow them to maintain a high level of control over the direction of that interaction;
- the observation of the behaviour of 'real' users engaged in the performance of 'real-life' tasks (or, at least, accurate simulations of such tasks);
- performance criteria other than relevance-based effectiveness; and
- methods for the acquisition and analysis of data, often qualitative in nature, that may be used in the calculation of values for non-traditional performance measures.

Our study of the Epic image retrieval system shares these concerns.

3 The Epic system

The Epic image retrieval system [23] is just one of the applications that we have created using FLAIR (a Flexible Architecture for IR) [26], a development of ECLAIR (an Extensible Class Library for IR), which we in turn implemented using ObjectStore (an object-oriented DBMS) and C++. One of the benefits of using FLAIR for the development of new IR applications is the facility it provides for the seamless integration of DBMS and IR functionality in a single system, enabling both exact (attribute-specific) and inexact (free-text) retrieval.

The *interface* to Epic (see Figure 1) was created using Java. Notable elements of the Epic interface include

- a multi-modal *query-formulation* mechanism, allowing
  1. the composition of a spatial (or visual) query by the drawing and labelling of rectangles on a sketchpad, in order to represent the relative positions and names of the objects desired within any retrieved image,
  2. the construction of a free-text query made up of keywords,
  3. the specification of a field- or attribute-specific query involving the selection of the desired field-name, and
  4. an indication of the level of confidence (on a 0–1 scale) that the user has in each element (spatial, free-text, attribute-specific) of their composite query; and
- a bi-modal *visualisation* mechanism for the display of members of the retrieval set, allowing users both
  1. to view thumbnail images, ranked in order of potential relevance, of a set of retrieved documents, together with the contents of descriptive indexing fields, and
  2. to view a full-size image of a single selected document together with any brochure text associated with it.

The highlight of the Epic *retrieval mechanism* consists in its use of Dempster–Shafer theory [29] at the query–document matching stage, in modelling searchers' variable levels of uncertainty in the separate elements (spatial, free-text, attribute-specific) of composite queries, and in combining multiple sources of evidence, i.e. the multiple query–document similarity scores derived from analysis of those separate elements. The operation of this mechanism is the subject of a companion paper [25] and technical report [24], and is not considered further here.

One problem for prospective evaluative studies of multimedia IR systems is the lack of a standard multimedia test collection. In the absence of an appropriate alternative, we created our own *database* of 800 photographs drawn from the archive of the National Trust for Scotland. The main themes of the photographs in this collection are: exteriors of castles and other grand buildings; the grounds of such buildings (with gardens, lakes, ornaments, etc.); and skyscapes, seascapes and landscapes. Each field of each record in the database contains either:

- a digitised photographic image;
- descriptive indexing data, such as the name of the photographer, the name of the property depicted or the date on which the photograph was taken—all recorded manually by the archivist at the time of the photograph's original storage;
- topical indexing data, in the form of keywords or captions describing the content of the image—assigned manually by a qualified information scientist at the time of the creation of the test collection;
- the text of a brochure about the property depicted in the image; or
- *spatial features.* Each spatial feature takes the form of an ordered pair, consisting of (i) a label representing an object depicted in the image, and (ii) positional coordinates specifying the location within the image of that object. These spatial features were derived semi-automatically by computer-science research students making use of dedicated spatial-indexing software.

4 Experimental methodology

In our evaluative study, we made use of a within-subjects (repeated-measures) experimental design. The independent variable was system type; each member of a single set of subjects was required to interact on separate occasions with systems of two types; two separate sets of values of a variety of dependent variables indicative of acceptability or user satisfaction were to be determined through the administration of questionnaires to each subject. One of the two systems that each subject made use of was a version of Epic allowing spatial queries only, henceforth referred to as 'System A'. 'System B' was a version of Epic allowing textual queries only. Our experimental hypothesis was that System A would prove to be more acceptable or satisfying to the user than System B.

We recruited 8 people who agreed to participate in our study as system users, i.e. as the subjects of the experiment. Since we had in mind that the members of this population would be making use of such systems in pursuit of tasks that would be germane to certain types of professional work, selection was conducted on the basis of our knowledge of the characteristics of members of the local professional community; individuals were approached, and recruited if agreeable. 5 subjects were employed in the art school of our University, 3 in that University's central printing service: all were graphic design professionals. Responses to a pre-search questionnaire indicated that our subjects could be assumed to have a good understanding of the design task we were to set them, but a more limited knowledge or experience of the search process. We could also safely assume that our subjects had no prior knowledge of the experimental systems that they would be asked to use.

We met one subject at a time, each on a separate occasion. For each subject, our procedure was as follows:

- an introductory orientation session;
- a pre-search questionnaire;
- a training session on the first system with which the subject was to interact;
- a hand-out of written instructions for the first task;

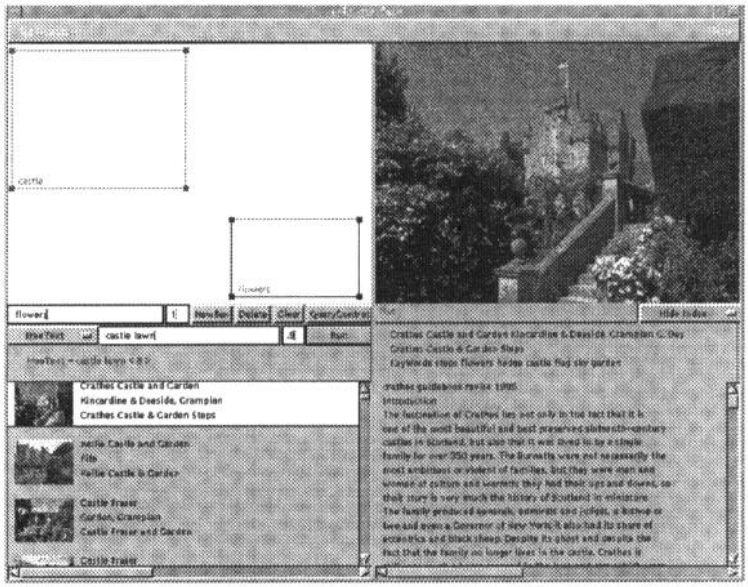

Figure 1: EPIC interface

- a search session in which the subject interacted with the first system in pursuit of the first task;
- a post-search questionnaire;
- a training session on the second system;
- a hand-out of instructions for the second task;
- a search session on the second system in pursuit of the second task;
- a post-search questionnaire; and
- a final questionnaire.

We were concerned that our subjects should be placed in a *simulated work task situation* [3] in which their information needs would evolve, in just the same dynamic manner as such needs might be observed to do so in subjects' real working lives. To this end, we had already carried out interviews with a separate sample of graphic design professionals, and (given the subject matter of our test database) reviewed the form and content of typical Tourist Board publications, with a view to establishing an appropriate task specification. The result of these preliminary investigations was the simulation, in our experiments, of a work task situation that defined for our subjects the context in which their tasks were to be performed, the source of their information needs, and the use to which the outcomes of their tasks were to be put.

In the written instructions that we gave to each subject, we asked them to imagine that they were a freelance designer, with responsibility for the design of leaflets on various subjects for the Scottish Tourist Board. We supplied them with a template for one of these leaflets, which identified the locations and the varying characteristics of 3 'slots' where photographs were to be inserted. (One slot on the front page, for instance, was to be superimposed with some title text; another was to be inlaid with smaller images.) We told each subject to assume that their task was to make a selection, from a large collection of images, of those 3 images that in their opinion would be most appropriate for filling the slots in a leaflet covering a particular theme. Subjects were asked to carry out this task twice, given a different theme and using a different system in each case; the 2 themes were 'The scenic splendour of the Scottish countryside in Spring and Summer' and 'The scenic splendour of the Scottish countryside in Autumn and Winter'. In the course of each of their 2 executions of their task, then, each subject was to carry out 3 searches, one search for each image. In the course of any individual search, a subject was able to issue as many separate queries as they wished.

The whole procedure was piloted on a 9th subject, also a design professional. In deciding on the order in which each of the 8 subjects in the study proper was to be (i) introduced to the 2 systems, and (ii) assigned the 2 themes, we employed a Greco-Latin square design [31], attempting to control for the sequence effects that might arise as a result of the learning that subjects would acquire from one search session to the next.

5 Experimental results

5.1 Pre-search questionnaire

Before being introduced to the experimental tasks or systems, subjects were asked in a pre-search questionnaire to choose from a list of facilities the one they would prefer for selecting photographs from a collection. 2 respondents specified 'a keyword-based search system for specifying queries made up of search terms', 3 specified 'an unordered sequence of small thumbnail images for scanning or browsing through', and 3 selected both of these. No respondent chose 'a classified index or catalogue for looking up specific photographs', or chose to specify any other kind of mechanism in their own words.

Subjects were also asked 'What sort of criteria do you use in measuring how successfully you complete a task such as the selection of photographs for the design of a leaflet?'. Most responses included references to features of the content or composition of the selected images—e.g., their relevance to the subject area. Only one respondent referred to the approval of the client on whose

Was the *task* ...?	
clear	unclear
simple	complex
familiar	unfamiliar
Was the *search process* ...?	
relaxing	stressful
interesting	boring
restful	tiring
easy	difficult
simple	complex
pleasant	unpleasant
Was the *retrieval set* ...?	
relevant	irrelevant
important	unimportant
useful	useless
appropriate	inappropriate
complete	incomplete
Was the *system* ...?	
efficient	inefficient
satisfying	frustrating
reliable	unreliable
flexible	rigid
useful	useless
easy	difficult
novel	standard
fast	slow
simple	complex
stimulating	dull
effective	ineffective

Table 1: Semantic differentials

behalf the leaflet was being produced.

5.2 Post-search questionnaires

Semantic differentials Each respondent was asked to describe various aspects of their experience of using each system, by scoring each system on the same set of 25 7-point semantic differentials [5]. 3 of these differentials focused on the *task* that had been set; 6 focused on the *search process* that the respondent had just carried out; 5 focused on the set of photographs *retrieved*; and 11 focused on the *system* itself (see Table 1). The result was a set of 400 scores on a scale of 1 to 7: 8 respondents scoring each of 2 systems on each of 25 differentials. On the questionnaire form, the arrangement of positive (e.g., 'fast', 'stimulating') and negative (e.g., 'slow', 'dull') descriptors was randomised so that a positive assessment would be represented sometimes by a high score (i.e., approaching 7) and sometimes by a low one (i.e., approaching 1). At the analysis stage, scores of the former type were reversed so that in all cases positive assessments were represented by low scores.

In our within-subjects design, the set of 8 scores on each differential for System A was compared with the corresponding set of 8 scores on each differential for System B. Our one-tailed experimental hypothesis was that, in any individual case, the set of scores for System A (enabling spatial queries) was drawn from a population of lower (better) scores than that for System B. Given the ordinal scale of the data, we calculated values of the non-parametric Wilcoxon signed-ranks statistic in order to test this hypothesis. At a significance level of $p \leq 0.050$, the results of the analysis were non-significant in every case, and the null hypothesis could not be rejected. When, however, the two sets of 88 scores on those differentials focusing on *system* were compared, the value of the Wilcoxon statistic was found to be significant at a level of $p = 0.045$, and the null hypothesis could be rejected. In other words, in one particular set of respects, the ratings given by users to System A were found to be significantly better than those given to System B.

The results obtained from respondents' interpretations of the semantic differentials were also analysed with a view to proposing a reduced set of differentials for future use in evaluative studies of this kind. The full set of 16 scores on every differential was compared with that on every other differential, and values of Spearman's rank correlation coefficient were calculated in order to test the one-tailed experimental hypothesis that in any of 300 $(= (25 \times 24)/2)$ cases the two sets of scores were positively correlated. In 52 cases, the correlation observed was found to be significant at a level of $p \leq 0.010$. The 'easy'/'difficult' (search process) differential alone, for instance, was significantly correlated with 10 other differentials. A minimal set of 7 differentials could be defined so that each of the original set of 25 was significantly correlated with at least one of the members of the new set. Such a set would include 'clear'/'unclear' (task), 'interesting'/'boring' and 'easy'/'difficult' (search process), 'appropriate'/'inappropriate' (retrieval set), and 'stimulating'/'dull', 'reliable'/'unreliable' and 'useful'/'useless' (system).

Likert scales Each user was invited to indicate, by making a selection from a 5-point Likert scale [5], the degree to which they agreed or disagreed with each of 5 statements about various aspects of their interaction with the system. These statements were phrased in such a way that responses would indicate the extent to which:

- the user's original information *need* was well-defined (statement: 'I had a mental image of a photograph that would satisfy my requirements');
- the user was able to formulate a *query* precisely representative of or coextensive with that need ('My query was an accurate representation of the type of image(s) I had in mind');
- the user was satisfied with the outcome of their *search* ('I am very happy with the image(s) I chose');
- the user was satisfied with the level of *recall* attained ('I believe that I have seen all the possible photographs that satisfy my requirement'); and
- the user was satisfied with the *overall* outcome of their interaction with the system ('I believe that I have succeeded in my performance of the design task').

Each user was asked to respond to each of the first 4 of these statements 6 times (after each of the 3 searches they carried out on each of the 2 systems), but to respond to the final statement only twice (after each of their 2 executions of the complete task). The result was a set of 208 scores on a scale of 1 to 5 (with 1 representing the response 'I agree completely' and 5 representing 'I disagree completely'): 8 respondents scoring each of 2 systems with respect to each of 13 statements. System A scored best when respondents recorded their reactions

to the first statement, about their pre-query 'mental image' (24-score mean: 1.21). System B scored best when respondents reacted to the third statement, about search outcome (24-score mean: 1.46), and scored almost as well on the first statement (24-score mean: 1.50).

In our within-subjects design, the set of 24 scores for each of the first 4 statements about System A was compared with the corresponding set of 24 scores for each statement about System B. Values of the Wilcoxon signed-ranks statistic were again calculated in order to test the one-tailed experimental hypothesis that, in any individual case, the scores for System A (enabling spatial queries) were lower (better). In the case of the first statement (about 'mental image'), the value of the Wilcoxon statistic was found to be significant at a level of $p = 0.030$; and in the case of the second statement (about query formulation), the value of the Wilcoxon statistic was found to be significant at a level of $p = 0.009$. Additionally, when the full set of 96 scores for all of the first 4 statements about System A was compared directly with the corresponding full set of 96 scores for System B, the value of the Wilcoxon statistic was found to be significant at a level of $p = 0.003$. In these three cases, therefore, our conclusion was that System A indeed scored significantly better than System B.

Qualitative data After completing the search session in which they made use of each system, subjects were asked to specify which features of the system that they liked, which that they disliked, and what features that they would like to have seen added. General comments about System B made by respondents at this stage included those that it was 'generally helpful', 'clear, concise and very easy and adaptable to use'. For two of those who used System B in their second search, the lack of a spatial query mechanism was perceived as a bonus (one, for example, referred to keyword specification being 'quicker than drawing boxes'): no other general pattern was revealed.

In contrast, respondents had uniform praise for the query-formulation features of System A (allowing spatial queries), extolling the virtues of 'being able to position objects in a graphic way, visually setting out the image I had in my mind's eye'. And: 'the ability to designate particular areas with objects', 'the positioning of features (boxes)', 'the spatial location and name selection system', 'the allocation of visualisations to positions ... and its further potential', 'the way the composition can be made ... a quicker way of searching for the correct image'. Interesting features that respondents would have liked to have seen added to System B included: 'an ability to change the squares' shapes to more irregular (and specific) shapes if necessary', 'more specific relationships between scale and positioning in the requested image and the images presented', 'a cut-and-paste to build a new image', 'a facility to click-and-drag selections to an area'. When invited to add 'any other comments' about System A, all 8 respondents made highly positive remarks, describing the system variously as: 'pleasant' and 'very easy' to use; 'a good idea', 'interesting', 'stimulating' and 'inspirational'; 'attractive', 'accurate' and 'flexible'; and 'more useful' than keyword-based systems.

5.3 Final questionnaire

Once each user had completed the tasks they had been set on both System A and System B, and the corresponding questionnaires, a final questionnaire was administered in which respondents were asked to compare the two systems and to specify (i) the one that 'helped' more in the execution of their tasks, and (ii) the one they 'liked' better.

7 of the 8 respondents decided that System A (allowing spatial queries) 'helped' more than System B; 6 of the 8 decided that they 'liked' System A better. Given the nominal nature of this data, the sign test was used to test the one-tailed experimental hypothesis that the stated preferences for System A in either case was significant. Indeed, the former result was found to be significant at a level of $p = 0.035$; the latter result, however, was found to be significant only at a level of $p = 0.145$.

When those respondents who said System A 'helped' more were asked 'Why?', answers were given that referred occasionally to basic retrieval features—e.g., 'I got better photographs from it'—but primarily to query-formulation features—e.g., 'I could create a retrieval query based on a picture in my mind's eye', 'visualisation + descriptor = more powerful', 'greater specificity', 'you can choose the composition you want'. One respondent commented that 'It is more intuitive, more fluid, in tune with the way that one works with images; it becomes more possible to interact with one's imagination and the system.' The single respondent who believed System B 'helped' more answered the question 'Why?' by stating 'I found it easier, i.e. being able to type in words'.

When those respondents who 'liked' System A better were asked 'Why?', they confirmed a shared enthusiasm for the query-formulation mechanism—e.g., 'it reflects more accurately the way that I think about images', 'matched current working method', 'visual memory and imagination related directly to visual requirement'. Comments in favour of System B were that it involved 'less fiddling with boxes', and was 'more direct'.

Respondents were also asked to rank in order of preference, given the hypothetical opportunity to carry out further searches, four types of system with facilities corresponding directly to those listed for approval in the earlier, pre-test questionnaire: an unordered sequence of small thumbnail images; a classified index or catalogue; a keyword-based search system; and a spatial-feature-based search system. 6 of the 8 respondents ranked spatial-feature-based systems more highly than any other: 1 ranked keyword-based systems most highly, and 1 ranked thumbnail sequences most highly. When responses were scored on a scale of 1 (high rank) to 4 (low rank), spatial-feature-based systems attained a mean score of 1.62, keyword-based systems 2.25, thumbnail sequences 2.38, indexes/catalogues 3.75. These responses stand in contrast to those obtained from the corresponding question in the pre-search questionnaire (see Section 5.1), where 5 of the 8 subjects selected a keyword-based system as their preference. In order to test the experimental hypothesis that the sets of 8 post-search scores for each system type were sampled from different populations, we calculated the value of the non-parametric Friedman statistic, which was found to be significant at a level of $p = 0.018$. Our conclusion, therefore, was that spatial-feature-based systems scored significantly better than any other.

Those respondents who rated systems based on spatial features most highly gave reasons which included the following: 'enrichment of the design process', 'it would facilitate the process of designing from image first, followed by text: more natural', 'focus on elements of required imagery expands imagined selective range and capacity', 'greater opportunity of matching initial ideas with com-

position', 'I can find an appropriate image ... in a more visually-oriented way'. The single respondent who rated the keyword-based system most highly stated that 'it would be the quickest'.

6 Conclusions

Conclusions may be drawn from this study not only about the value of the Epic system itself, but also about the value of the methodology we used to evaluate it.

6.1 The Epic system

Chief among our findings about the Epic system are the following:

- The statement with which users of the spatial-query system agreed most completely was 'I had a mental image of a photograph that would satisfy my requirements'. Users of the textual-query system were also in general agreement with this statement. These results appear to support our initial sub-hypothesis: that searchers often have a reasonably well-defined 'mental image' of a picture that might satisfy their visual information need.
- Users of the spatial-query system agreed, to a *significantly* greater degree than users of the textual-query system, with the statement 'My query was an accurate representation of the type of image(s) I had in mind'. This result appears to support the second of our sub-hypotheses: that searchers are able to represent their 'mental image' by drawing and labelling rectangles on the rudimentary electronic sketchpad that the spatial-query system alone provides.
- Users rated the spatial-query system *significantly* better than its more conventional counterpart in a variety of respects:
  1. When users were asked to rate each system on each of a set of 11 semantic differential scales indicating, for example, how 'useful' or 'useless', how 'effective' or 'ineffective', how 'satisfying' or 'frustrating' they found it, System A scored significantly better than System B.
  2. Similarly, when users were asked to rate each system on each of a set of 5 Likert scales indicating how satisfied they were with different aspects of their interaction with the system, System A scored significantly better overall than System B.
  3. Again, when users were asked to specify which of the two systems 'helped' them more in the execution of their tasks, a significantly larger number specified System A.
  4. Finally, when users were asked to rank in order of preference 4 types of system, spatial-query-based systems scored significantly better than any other type.
- Moreover, users were particularly enamoured with the query-formulation features of the spatial-query system, above any other feature of either system. At different stages of the exercise, users were asked:
  1. which features of each system they liked most;
  2. why they specified one system as 'helping' more than the other; and
  3. why they ranked a particular type of system more highly than another.

  In each case, the facility mentioned most frequently in positive terms was Epic's spatial querying mechanism. In conjunction with those presented above, these results appear to support the third of our sub-hypotheses: that the provision of a facility allowing the specification of spatial queries helps users to improve their searches, and thus to satisfy their information needs more effectively, efficiently and easily.

We believe that these results successfully demonstrate the clear interest and value of further investigation into the ways in which spatial querying may be developed.

6.2 The evaluative framework

In Section 2, we identified the gradual emergence of a standardised framework for the evaluation of interactive, multimedia retrieval systems, and highlighted four methodological elements that are consistently shared by contemporary studies of such systems. In our own practice, we have taken care both to follow the individual guidelines established in this previous work, and to attempt to embed such practice in an overall framework that exhibits internal coherence and that is capable of general application. In particular:

- In accordance with procedures recommended by 'Evaluation Light' [18], we decomposed our general hypothesis into a set of sub-hypotheses, each rather more limited in scope but consequently more manageable. Necessarily, any test of a primary hypothesis such as ours requires study of the system as a whole. In the evaluative tradition that emerged in the heyday of batch-mode retrieval, it was assumed that a suitably holistic assessment could be arrived at through successive consideration of isolated episodes of interaction, such as those bounded by the communication from user to system of a query and the communication from system to user of a response. With highly interactive systems such as Epic, which allow the user to maintain a high level of continuous control over the information-seeking process, it is more difficult to break down search sessions into neat sequences of episodes that have equal significance for the user (or, therefore, for the evaluator). It is still possible to evaluate the whole—but the process depends more on the isolation of individual characteristics of users and system, and the determination of the degree to which particular system features reflect users' abilities, cater for their preferences, or satisfy their requirements.
- We did not test out the Epic system on volunteer computer-science undergraduates: we recruited subjects from that particular professional user group for whom our system is intended. Moreover, in line with principles developed by Borlund and Ingwersen [3], we placed our subjects in a simulated work task situation whose design was based on the results of an extensive, preliminary investigation of the real-life work patterns of members of that user group. We did not, for instance, merely specify

the topic of an information need on which each of our subjects would be required to search; neither did we require each subject to imagine an information need of their own; instead, we defined for our subjects a work task *situation* made up of specifications of: a *context*—i.e., the day-to-day work of a graphic designer; a *task*—i.e., that of illustrating a promotional leaflet; and two related *themes*—i.e., general topics that might be covered by such leaflets. The benefit of using this approach was that we were able to observe (at a 'macro' level) subjects' use of system features in their performance of context-situated tasks, rather than (at a 'micro' level) systems' responses to pre-ordained queries. Some evidence that our simulation was successful is provided by our subjects' post-search rating, on a semantic differential scale, of the clarity of the task set them. With both systems, the best (i.e., lowest) average score on any of the 25 differentials was obtained when subjects were asked how 'clear' or 'unclear' the task was (8-score mean for System A: 1.125; for System B: 1.375).

- We focused on a dimension of system performance—acceptability, or level of user satisfaction—that is casually ignored in some equivalent studies. Rather than use pre-search relevance judgments to determine the *effectiveness* of the system's response to each query, we determined the general *acceptability* of the system by analysing the opinions of users—opinions that were not simply concerned with the system's responses to individual queries, but that illuminated various aspects of the users' wider interaction with the system in the course of performing certain tasks.

- Our focus on user satisfaction as a performance criterion led us to make extensive use of methods of data collection and analysis that have more honourable histories in the fields of information science and cognitive science, amongst others. Given our success in obtaining results that relate clearly to each of our sub-hypotheses, we are content to echo the conclusion drawn by Brajnik et al. [4]—one that admittedly would not startle many outside the IR community—that 'the use of semantic differentials [and] Likert scales ... has proven to be an effective and accurate method for acquiring, validating and analysing [subjective] data'.

The size of the sample in studies of this nature is a detail that inevitably attracts attention. Given the familiar constraints of time and money, a small sample is a common feature of studies that make every attempt to observe interaction with information in real-world settings. We would not deny that we would very much like to have had the wherewithal to involve a larger number of subjects in our experiments. Nevertheless, we would argue that, in the present context, a small sample should not necessarily be viewed as a fatal shortcoming. Indeed, the very fact that we obtained results of statistical significance *despite* our sample's small size would seem only to validate our methodological design.

In summary: we believe that, in obtaining results both of statistical significance and of wide interest and potential application, we have made a substantial contribution to a demonstration of the validity and reliability of our adopted methodological framework. It is clear, nevertheless, that we have emphasised our concern with levels of user satisfaction to the almost total exclusion of consideration of more traditional measures of retrieval effectiveness or efficiency. It is hoped that in our next round of experiments we will be able to draw meaningfully, not only on a larger sample of subjects, but on the contents of detailed transaction logs and video recordings, and to develop novel measures of effectiveness that are appropriate to interactive multimedia retrieval.

Acknowledgements Joemon Jose's work was supported by a grant from the Principal's Research Fund at the Robert Gordon University (RGU), Aberdeen; he is currently working in the Department of Computing Science at the University of Glasgow. Jonathan Furner is now a member of the Department of Library and Information Science at the University of California, Los Angeles. The authors would like to thank our experimental subjects, especially Ian Burt for his valuable insights into the work of graphic designers; Micheline Beaulieu and Nick Belkin for their encouragement and advice; and the SIGIR referees for their helpful comments on the first draft of this paper.

References

[1] L. Armitage. The importance of contextualising image retrieval user behaviour. Paper presented at the Workshop on Content-Based Image Retrieval, Institute of Image Data Research, University of Northumbria at Newcastle, Newcastle-upon-Tyne, February 1998.

[2] M. Beaulieu, S. Robertson, and E. Rasmussen. Evaluating interactive systems in TREC. *Journal of the American Society for Information Science*, 47(1):85–94, 1996.

[3] 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.

[4] G. Brajnik, S. Mizzaro, and C. Tasso. Evaluating user interfaces to information retrieval systems: a case study on user support. In H.-P. Frei, D. Harman, P. Schäuble, and R. Wilkinson, editors, *SIGIR '96: Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 128–136. Association for Computing Machinery, New York, 1996.

[5] C. H. Busha and S. P. Harter. *Research methods in librarianship: techniques and interpretation.* Academic Press, San Diego, CA, 1980.

[6] C. W. Cleverdon, J. Mills, and E. M. Keen. *Factors determining the performance of indexing systems.* College of Aeronautics, Cranfield, 1966.

[7] S. Draper. *Brief ideas about IR evaluation.* Department of Psychology, University of Glasgow, Glasgow, July 1995. Available online at URL: `http://www.psy.gla.ac.uk/~steve/IReval.html`.

[8] M. Dunlop, editor. *Proceedings of the Second Mira Workshop*, Technical Report TR-1997-2. Department of Computing Science, University of Glasgow, Glasgow, January 1996. Available online at URL: `http://www.dcs.gla.ac.uk/mira/workshops/padua_procs/`.

[9] P. G. B. Enser. Pictorial information retrieval. *Journal of Documentation*, 51(2):126–170, 1995.

[10] N. Fuhr, C. J. van Rijsbergen, and A. F. Smeaton, editors. *Evaluation of multimedia information retrieval*, Schloss Dagstuhl Seminar No. 9716, Report No. 175. Universität des Saarlandes, Saarbrücken, 1997. Available online at URL: `ftp://ftp.cs.uni-sb.de/pub/dagstuhl/reporte/97/9716.ps.gz`.

[11] J. Furner. The evaluation of hypermedia IR systems: a statement of the problems. In M. Dunlop, editor, *Proceedings of the Second Mira Workshop*, Technical Report TR-1997-2, pages 27–36. Department of Computing Science, University of Glasgow, Glasgow, January 1997. Available online at URL: `http://www.dcs.gla.ac.uk/mira/workshops/padua_procs/`.

[12] J. R. Galliers and K. Sparck Jones. *Evaluating natural language processing systems.* Technical Report TR291. Computing Laboratory, University of Cambridge, Cambridge, March 1993. Available online at URL: `http://www.cl.cam.ac.uk/ftp/papers/reports/TR291-ksj-jrg-evaluating-nl-%systems.ps.gz`.

[13] V. N. Gudivada and V. V. Raghavan. Special issue: Content-based image retrieval systems. *IEEE Computer*, 28(9), 1995.

[14] V. N. Gudivada and V. V. Raghavan. Modeling and retrieving images by content. *Information Processing & Management*, 33(4):427–452, 1997.

[15] A. Gupta and R. Jain. Visual information retrieval. *Communications of the ACM*, 40(5):71–79, 1997.

[16] D. Harman. Special issue: Evaluation issues in information retrieval. *Information Processing & Management*, 28(4):439–528, 1992.

[17] D. K. Harman and E. M. Voorhees, editors. *The Fifth Text REtrieval Conference*, NIST SP 500-238. National Institute of Standards and Technology, Gaithersburg, MD, 1997.

[18] D. J. Harper and D. G. Hendry. Evaluation light. In M. Dunlop, editor, *Proceedings of the Second Mira Workshop*, Technical Report TR-1997-2, pages 53–56. Department of Computing Science, University of Glasgow, Glasgow, January 1997. Available online at URL: `http://www.dcs.gla.ac.uk/mira/workshops/padua_procs/`.

[19] W. Hersh. Relevance and retrieval evaluation: perspectives from medicine. *Journal of the American Society for Information Science*, 45:201–206, 1994.

[20] W. Hersh and M. Hancock-Beaulieu. *Evaluation of information retrieval systems.* SIGIR '95 tutorial. Biomedical Information Communication Center, Oregon Health Sciences University, Portland, OR, 1995.

[21] W. Hersh, J. Pentecost, and D. Hickam. A task-oriented approach to information retrieval evaluation. *Journal of the American Society for Information Science*, 47(1):50–56, 1996.

[22] J. M. Jose. Evaluation of the Epic photograph retrieval system. In N. Fuhr, C. J. van Rijsbergen, and A. F. Smeaton, editors, *Evaluation of multimedia information retrieval*, Schloss Dagstuhl Seminar No. 9716, Report No. 175. Universität des Saarlandes, Saarbrücken. Available online at URL: `ftp://ftp.cs.uni-sb.de/pub/dagstuhl/reporte/97/9716.ps.gz`.

[23] J. M. Jose. *An integrated approach for multimedia retrieval.* PhD thesis, The Robert Gordon University, Aberdeen, 1998.

[24] J. M. Jose and D. J. Harper. *Epic: a photograph retrieval system based on evidence combination approach.* Technical Report TR-97/2. School of Computer and Mathematical Sciences, The Robert Gordon University, Aberdeen, 1997.

[25] J. M. Jose and D. J. Harper. A retrieval mechanism for semi-structured photographic collections. In A. Hameurlain and A. Min Tjoa, editors, *Database and Expert Systems Applications: 8th International Conference: DEXA '97: proceedings*, Lecture Notes in Computer Science, Volume 1308, pages 276–292. Springer, Berlin, 1997.

[26] J. M. Jose, D. G. Hendry, and D. J. Harper. *FLAIR: a flexible architecture for information retrieval.* Technical Report TR-97/3. School of Computer and Mathematical Sciences, The Robert Gordon University, Aberdeen, 1997.

[27] M. 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*, 44:777–799, 1996.

[28] T. Saracevic. Evaluation of evaluation in information retrieval. In E. A. Fox, P. Ingwersen, and R. Fidel, editors, *SIGIR '95: Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 138–146. Association for Computing Machinery, New York, 1995.

[29] G. Shafer. *A mathematical theory of evidence.* Princeton University Press, Princeton, NJ, 1976.

[30] K. Sparck Jones. *Information retrieval experiment.* Butterworths, London, 1981.

[31] J. Tague-Sutcliffe. The pragmatics of information retrieval experimentation, revisited. *Information Processing & Management*, 28(4):467–490, 1992.

[32] J. Tague-Sutcliffe. Special topic issue: Evaluation of information retrieval systems. *Journal of the American Society for Information Science*, 47(1):1–105, 1996.

[33] A. Veerasamy and N. J. Belkin. Evaluation of a tool for visualization of information retrieval results. In H.-P. Frei, P. Harman, P. Schäuble, and R. Wilkinson, editors, *SIGIR '96: Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 85–92. Association for Computing Machinery, New York, 1996.