Search Techniques Within a Multiple Database Environment

C L Bird

Applied Science & Technology Group, IBM UK Laboratories,
Hursley Park, Winchester, Hampshire

P J Elliott

Applied Science & Technology Group, IBM UK Laboratories,
Hursley Park, Winchester, Hampshire

Abstract

For many information providers, over a range of applications, digital images are the medium of choice and the efficient and effective retrieval of those images is an active area of interest. Traditional keyword-based search techniques are being supplemented by image content-based methods and attention is being given to so-called meta-searchers, which enable multiple databases to be queried with a single request. In this paper, we describe work being conducted in a multiple database environment, using both conventional and content-based techniques, with particular emphasis on the need to cater for the naive user. The work is being carried out as part of the ELISE project (Electronic Library Image Server for Europe), which is partly funded by the European Community under its Fourth Framework Libraries Programme.

1 Introduction

Retrieval of images from a repository by specifying one or more key descriptive terms represents the traditional database approach. Evolution is extending this along two dimensions: firstly, content-based techniques are being developed not only to counter some of the perceived deficiencies of text-only methods but also to recognise the potentially high information content of digital images. (Note that in the context of this paper, we will use the expression “content-based” to mean “based on the contents of an image, and not on its associated text description”).

The second dimension is the extension to multiple databases, permitting simultaneous searching in a number of collections. This paper will describe some preliminary studies of text driven searching in multiple databases. It will then describe how content-based methods might be applied to those same databases on an individual basis. Finally, in looking at future directions, we shall speculate on some of the implications of combining the two to provide content-based access to multiple image databases.

It is worth mentioning at the outset that “the challenge of image retrieval” brings together the two distinct disciplines of *image analysis* and *information retrieval*. In such cases it is inevitable that differences of understanding, and the precise meaning of terminology, will cause confusion from time to time. Our background is in image analysis, so we trust that IR specialists will forgive us if we appear to be using the wrong word or saying the wrong thing. If we can raise the awareness of these problems and stimulate discussion, so that in due course the differences are resolved, we will have achieved one of our objectives in writing this paper.

The foundation for this work is the ELISE project (Electronic Library Image Server for Europe), a collaborative endeavour partly funded by the European Community under its Fourth Framework Libraries Programme. ELISE is currently in its second phase, running from October 1996 to September 1999. Phase I successfully demonstrated a technology for the text-based retrieval of images from cultural heritage collections. The second phase is transforming the early prototype into an operational infrastructure for networked image information in organisations as diverse as libraries, museums, universities and medical laboratories across Europe. For Phase II, both the number of participating groups and the range of collections has expanded, with the project having a focus on the search processes which will be described in this paper. The project also addresses the user and market issues involved in moving from a project to a product and will propose a legal and corporate framework within which an image service could operate.

Effective communication is vital to a project of this nature, with a variety of interests to be accommodated. We have consulted the consortium partners with an extensive questionnaire; their responses

provide a significant input to the discussion in this paper.

2 The ELISE project

The nine partners in ELISE are: De Montfort University (UK), Tilburg University (The Netherlands), the IBM UK Scientific Centre (UK), the Victoria and Albert Museum (UK), the University of Limerick (Ireland), Radio Telefis Eireann (Ireland), the Hunt Museum (Ireland), the Laboratory for Biomedical Informatics (The Netherlands), and the Universite Libre de Bruxelles (Belgium).

The project focusses primarily on the educational market which has a high demand for good quality images. The user groups include students in further and higher education, teachers and lecturers, suppliers (and rights holders) of images, libraries and librarians. End users in this area require easy access to a wide range of images with minimum bureaucracy at an affordable cost. Image suppliers require effective protection from unauthorised use, effective methods of payment and minimum administration. Librarians require ease of administration, effective licensing arrangements, high performance operational systems, good channels for inter-library cooperation and effective navigational and delivery facilities.

Access to the ELISE system is via a Web browser. The user interface client uses Java applets, served from a Java Web server which is part of the ELISE Broker. The Broker provides functions such as controlling access to the image collections, logging usage by different users, charging for use according to protocols agreed between the users and the image providers, and watermarking images to provide protection against unauthorised copying. The Broker forwards authorised search requests to the Database Image Server, which translates the requests into the Z39.50 protocol and routes them to the appropriate database(s).

Seven partners in this second phase of the ELISE project are providing image collections ranging in size from 1,000 to 20,000 images, with a total of around 50,000. The subjects include paintings, furniture, museum objects, local history, newspapers, journals and biomedical images. In many cases images relating to a particular subject are contained in more than one collection, so that issues related to searching multiple databases are important, in addition to all the usual issues of efficiently searching individual image banks. Another important aspect of ELISE is the diverse user base that is available within the project for testing and evaluation of the system as it develops. It is intended to extend this user base by collaboration with other similar projects, and with individuals who are interested in providing constructive criticism and evaluation of the system.

Thus the ELISE project represents an ideal opportunity to test the effectiveness of image retrieval methods in an operational environment. Having set up a system employing conventional text-based searching facilities to deliver images to the users, the focus of the project will then be on the practical application of new ideas, and the problems involved with implementing them on a large scale. It will be a recurring theme of this paper that what we are trying to do is to provide help to non-expert users in finding the images they want, and to make the ELISE system attractive and easy to use.

It is an important part of the conduct of the ELISE project that the partners are fully involved in the evaluation, as already alluded to above. This is built into the workplan for both content-based retrieval and collection-level metadata. In each case, a proposal is written and sent to the users for comment. Their comments are incorporated into the specification of the system, which is then implemented. The users evaluate the system and provide feedback, which will result in modifications to the system. Further testing and evaluation is then carried out before the final report is written. By this means we hope to ensure that the system as finally implemented reflects what a wide range of users actually finds useful, rather than what the developers think they need.

3 Strategies for Searching

3.1 Background

The traditional approach to indexing multimedia databases involves authoring a set of key descriptive terms; the query mechanism then requires the searcher to specify one or more of those key terms. We consider two dimensions along which the retrieval process is evolving.

Firstly, content-based techniques are being developed, in part, to counter some of the perceived deficiencies of text-only methods such as keyword ambiguity and omission. Additionally, methods based on the actual contents of an image are held to be better able to recognise the potentially high information content of digital images. One of us has recently conducted a broad survey of the emerging technologies for content-based retrieval for the ELISE project [1] while, for a more detailed analysis of the principal techniques presently being used, see [2].

The second dimension is to extend the search to many different collections, possibly held in a form of

federation. The latter is a union of databases which serve a common interest. They may be competitive in nature, if they contain similar information but are provided by rival business interests. On the other hand, the federation may comprise databases with distinct contents but with common points of access, which is the case for ELISE.

The authors of the Stanford Protocol Proposal for Internet Retrieval and Search (STARTS) [3] list three main tasks for what they describe as “meta-searchers, which are services that provide unified query interfaces to multiple search engines”:

- Choosing the best sources to evaluate the query;
- Evaluating the query at these sources;
- Merging the query results from these sources.

To avoid misunderstanding, the verb to evaluate, used in this context, means to issue a query to a source and to obtain the results from that source.

The ELISE project brings together several disparate collections and a diverse user base. A query formulated for one collection will not necessarily be productive when directed to another. Dash and Hurson [4] argue that, for users who want to query heterogeneous, distributed databases without specifying their access, “it is necessary to develop an efficient and accurate means for detecting semantic homogeneity and similarity among the local information”. The differences between the collections in the ELISE federation and the use of different database schema both militate against any attempt to integrate those schema. Zhao [5] believes that the aim is better achieved by schema *coordination* than by *integration*. Against the schema integration approach, he argues that it “requires the resolution of both semantic and structural differences among the component databases”, and reasons that this presents continuing difficulties. He says: “The fundamental problem of metadata management in federated database systems is to provide a single database image so that users can conduct queries without having to know the schemas of individual component databases. The schema coordination approach solves this problem by mapping the component attributes with federated attributes”. Zhao’s solution is based on an “attribute correspondence matrix”, which maps all the attributes in the federation to a generic set, thereby achieving logical data independence.

Each of the ELISE collections has a set of fields that are capable of being queried, using the appropriate search engine. For a schema coordination approach, we need a general model which defines the kind of information that can be queried. The approach taken by the ELISE project was to adopt the Dublin Core model [6] and to invite each information provider to consider how their chosen catalogue fields could be mapped to a Dublin Core element (thus producing an attribute correspondence matrix). When formulating a query for a single database, the matrix can be used to find the nearest equivalent to the general term, and the appropriate results returned. For example, Resource Identifier might map to Registration Number in one instance, and to an image URL in another. Some fields, however, may well be common to several collections.

We note that the results returned will depend on the way the Dublin Core fields are interpreted by the different databases. We are also aware of the deficiencies of the Dublin Core approach from a strict Information Retrieval point of view. However, given the increasing popularity of the Dublin Core, and our objective to help *non-expert* users find the images they want, we believe this is a practical way forward.

3.2 Querying of Multiple Databases

Given that ELISE will be managing a number of inter-connected image databases with disparate contents, a fundamental question that any user may pose is:

“Which collection(s) should I search to find images about a particular subject?”

The simplest answer, of course, is to send the request to all of them! But with increasing numbers of databases and users, this could generate unacceptable levels of network traffic, and would require more powerful image servers to cope with the number of spurious queries. Another simple approach is to provide high-level descriptions of the contents of each collection, which the users can inspect to decide which ones look promising. An example of this in ELISE (for the Tilburg collection of images) is:

“Topographic pictures, historical prints, maps, portraits and pictures of social life in the province of Noord-Brabant (the Netherlands).”

However, this may not provide sufficient information. Clearly its usefulness depends on the descriptions being comprehensive, accurate and up to date. For a more effective search, we would like to provide the user with a means of directing a query to the collection or collections where it is most likely to succeed. To achieve this we need to generate a quantitative ranking of the query against the contents of each collection. Thus there are a number of problems to be solved:

- Generate a description of the set of ELISE image collections at an appropriate level of detail;
- Associate the number of images in each collection with that description;
- Be able to use the description to identify a topic of interest, and hence find which collection(s) have images

relevant to that topic.

3.2.1 Collection-Level Metadata

To produce the required description, ELISE will employ a system we call “collection-level metadata” – that is, metadata which describes the collection as a whole, rather than an individual item (which we refer to as “content-level metadata”).

We propose that it should take the form of a hierarchical classification scheme. The top level (or root node) will contain all the images in the complete set of ELISE collections. At the next level down, it would be divided into a number of categories, each sub-divided in turn, and so on. The number of levels needed has to be determined. If each level is sub-divided into ten, as in the well-known Dewey Decimal system, then we anticipate that about four levels will be sufficient for the 50,000 images in ELISE. Note that this classification scheme has to be constructed specifically for ELISE. It must contain *all* the images from *all* the collections, and each node at the lowest level (the leaves) must have at least *one* (or more) image from *one* (or more) of the collections.

How can we generate such a classification scheme? The image owners don’t understand their collections in enough detail to be able to tell us precisely what categories should be used, and which images would correspond to each category. Dash and Hurson [4] suggest that it can be done using a thesaurus. If all the collections have been catalogued using the same thesaurus (or other form of authority file), then it is a simple matter to run a batch process to establish how many images in each collection (if any) correspond to each category. The Art and Architecture Thesaurus (AAT) would be appropriate for several of the collections, but to date only one partner has used it to any significant extent, and then not completely. And there is no one thesaurus that covers all the different images that are available in ELISE¹. More importantly, if an authority file has not already been used to catalogue a collection of images, the amount of effort needed to do it in retrospect is substantial, and this effort is not available in the ELISE project.

The problem is similar to that of automatic *document classification*. Dolin et al [7] do this for WWW newsgroups by using a “training set” of electronic library catalogue records (which have already been classified) to build a rich set of terminology corresponding to each class. Newsgroups can then readily be associated (via this terminology) to the classification categories. User queries can similarly be associated to a category, so that a query can be matched to the appropriate newsgroup(s). However, in the case of our ELISE collections we don’t have such a training set with its associated classifications. We must rely on the text descriptions of the images produced by the cataloguers (the *content-level* metadata), along with a limited amount of intellectual effort from the image owners.

3.2.2 Clustering

A more promising approach is that of *document clustering*. This is done by associating documents into clusters according to similarities in the text. Clusters can then be merged with other similar clusters to form nodes in a hierarchy which describes the complete set of documents at different levels of detail. We propose to use the clustering tool in the IBM Intelligent Miner for Text² system to investigate clustering of the text descriptions of the ELISE images. This will be done on each collection separately. Some experimentation will be needed to find the most appropriate similarity measures to use, and to control the number of levels that are generated. In addition we will investigate which parts of the catalogue information (in terms of the Dublin Core fields) are most useful for our purpose. For instance we expect that Subject/Keyword or Creator will be useful in generating clusters of similar items, whereas Resource Identifier (which, by definition, must be unique for each image) will not be relevant.

If this approach is successful, then we will have generated a set of nodes to classify each collection in ELISE. It then remains to name each node with an appropriate category, which for reasons of consistency should be derived from a relevant authority file (such as the AAT, or the Dewey scheme). At this stage we will need the assistance of the image owners to suggest names for each category in their collections. By way of a simple example, it could be that one collection contains images of *knives*, *forks* and *spoons*, which are identified as three clusters according to those words. It should be apparent to a human observer that a suitable name for a cluster at the next level up in the hierarchy (which contains clusters of knives, forks and spoons) would be *cutlery* – even if this word doesn’t actually appear in the descriptions of the individual images. So we need a combination of “bottom-up” data processing (to group the images into clusters), and “top-down” intellectual effort (to assign category names to the clusters). We are well aware that the process will, in practice, be more complicated than

¹ If a thesaurus *had* been used, it would also be useful in searching. The user’s query could be matched to the appropriate node and then a collection that has one or more corresponding images could be selected. If there is no match, the query could be moved up a level in the hierarchy (i.e. to a broader category) to see if there are matches there. Note that other work in ELISE, not reported here, is concerned with the use of a thesaurus as an Intelligent Search Intermediary for searching a single image collection.

² IBM and Intelligent Miner for Text are trade marks of International Business Machines Corporation.

this, and we may need several iterations to come up with a workable scheme. Nevertheless, it is our contention that the intellectual effort will be substantially less than that needed to re-catalogue the entire collection, and hence be achievable with the resources available in ELISE.

3.2.3 Generating the Metadata

The next step is to use the output from the clustering tool to count how many images from each collection correspond to each category. Finally, we will combine the individual classification hierarchies into a single classification scheme encompassing all the ELISE collections. Having used authority files to name the categories, it should be readily apparent where different collections overlap, and hence where a particular category can be associated with images from more than one collection. Using our simple example, it could be that a second collection has images of spoons, but no knives or forks. In that case, the higher level term “cutlery” will indicate that both collections have relevant images, as will the lower level term “spoons”, but “knives” and “forks” will indicate that only the first collection should be searched. Of course, if one particular collection is distinct from all the others there will be no overlap, and all of its categories will be populated only by that collection.

The collections that are analysed in this way will constitute our “training set”. If new images are added to existing collections, or new collections that cover the same subject areas are added to the ELISE federation, we can use the classification tool of Intelligent Miner for Text to automatically associate the new text descriptions to the categories already in use. However, if new collections covering *different* subject areas are added to ELISE, it will be necessary to carry out the clustering procedure on the new images. This will have to be followed by the assignment of category names to the new clusters as described above. In this way the collection-level metadata can be expanded as the ELISE system develops.

3.2.4 Using the Metadata System

Ideally we would like to take any search expression input by the user and match it to the appropriate classification category automatically (as in Dolin et al [7]). This would only be feasible if the training set covered all the required subject areas, which is not likely to be possible within the scope of the current ELISE project.

A practical alternative is to allow users to browse through the classification hierarchy in the same way that they browse through a file system on a PC. Even non-expert users should find such a system easy to understand and use. Once they have found a category that interests them, an indication will be given as to which collection(s) contain images in that category. Depending on the number of images (which in turn depends on the level in the hierarchy and the size of the collection), it may be feasible at that point to automatically retrieve all the images for that category from the appropriate collections. This can be done using the data from the clustering tool for the cluster associated with that category. Alternatively, if such an approach would return too many images, users could put in search expressions of their own with the option of (a) searching only the highest ranking collection, or (b) searching all those collections which contain relevant images.

In summary, we have proposed a semi-automatic method for generating quantitative collection-level metadata. Once a suitably large training set has been accumulated, it could be applied to other collections in similar fields fully automatically. It remains to be seen how much intellectual effort is actually needed by the ELISE image providers in associating categories with the clusters to generate the classification hierarchy. Clearly this will depend on the diversity of images within any one collection, since a large number of small clusters would need too much effort at the association stage. Even if we are only partially successful, we hope to go some way towards achieving our aim of making the ELISE system easy to use.

3.3 Content-Based Querying

We have already mentioned the limitations of searching on the text associated with images. If the user has a description of the *contents* of an image (either in conceptual terms such as “a red circle on a green background”, or by means of an example image which is similar to what they are looking for), content-based methods enable a query to be expressed in the form:

“Show me some images that look like this”.

Our work on the application of content-based retrieval methods to the ELISE project began with a broad survey and assessment of current content-based retrieval techniques for image and other multimedia data types [1]. It is clear that, while much effort has been aimed at developing *techniques* in recent years, there has been limited evaluation by real users of systems *using* these techniques. This may well be connected with the fact that content-based image retrieval is still regarded as a novelty, with little serious usage to date. We therefore propose to design and evaluate a content-based search interface for the ELISE image collections, with particular emphasis on:

- User interface issues: for example, how to provide a query interface that a non-expert user can easily use,

whilst allowing for more complex queries from experienced users;

- Query effectiveness in different contexts - essentially user perceptions of “goodness” in terms of precision, recall and relative ranking;
- How query performance might be assessed;
- How we might attempt to “tune” the classification algorithms to suit a particular class of images.

The implementation of content-based search methods for ELISE will be based on the QBIC technology (Query By Image Content), as developed at the IBM Almaden Research Centre.

3.3.1 QBIC (Query By Image Content)

QBIC employs image analysis techniques to extract feature characteristics from the images in a collection, storing those features as vectors in a database, to which queries are addressed. Queries are formulated visually, either by specification or by example, and issued to a matching engine that finds images from the database with similar features [8].

The feature extraction algorithms used by QBIC are pixel-based, in that the computations operate upon aggregates of the individual pixels in the image, whether considering the entire scene or objects within it (subparts of the image). The computations are essentially based on the properties of colour, texture and shape. Pixel-based characteristics are often particularly useful for specimen images, that is those images in which the significant object or objects is or are depicted against a neutral background.

3.3.2 Evaluation of QBIC in ELISE

The ELISE programme envisages the potential end user of a network of image libraries to be the European public at large. Our overall goal is to investigate where content-based retrieval does have realistic and practical benefits, particularly for users who, while familiar with the subject of their enquiry, are not conversant with sophisticated database retrieval techniques.

To make this a tractable experiment, we intend to concentrate on specimen images as described above. This by no means covers all the images in the ELISE federation, but does provide us with a useful subset on which we can base our first analysis of the results. Although a variety of features is available for characterising such specimen images, we shall initially concentrate on colour layout. With this mode of query specification, the user paints - or draws - a colour sketch of the sort of image they are looking for. Alternatively, the query can be launched from an example image, in which case the resulting images will be returned according to their similarity to the example in terms of colour arrangement. We chose colour layout for the first experiment, as it seems to be the feature that is most applicable to the images in the ELISE collections.

The user interface will be made as simple and straightforward as possible, building on our previous experience of applying content-based techniques to a variety of image applications [9, 10]. Users will be restricted to searching just one collection at a time, to avoid experimental complications that would not assist in the achievement of the overall goal. They will be asked to provide a limited form of relevance feedback on the results, by rating each of the returned images on a qualitative scale, according to how relevant the image was to their original query. They will also have the opportunity to give free-form comments in support of their verdict, primarily based on visual factors such as the size, location and orientation of structures within the target and each result image. In addition we will be posing questions as to the preferred means of starting a search i.e. (a) painted specifications or (b) example images, and the users’ overall satisfaction with the results (as opposed to feedback on specific results).

Depending on the outcome of this first experiment, and our experiences in evaluating the results, we will then move on to a second stage. This could involve the use of another image feature such as texture, instead of colour layout. In summary, we hope to gain valuable insight into the practical benefits of content-based querying in a real-life situation.

4 Future Directions

Consider, in the ELISE context, combining the functions of sections 3.2 and 3.3, and posing a question such as:

“Which collection(s) should I search to find images that look like this?”

Recalling the model outlined in the Introduction of a two-dimensional expansion of the options for image retrieval, we are now looking into the fourth cell of the matrix. How, then, might the collection-level metadata concept be extended to high-level visual descriptions of the types, styles and classes of images held in each collection?

Chang and her co-workers [11] have considered this problem for a web-based multimedia information retrieval environment, and suggest the use of “image templates as mediums of visual abstractions”. We consider this to be a useful approach to explore in seeking visual forms of metadata. To illustrate how the model could be applied in one situation, we might be looking for images with autumnal hues. If our metadata includes a colour-

range indicator, we could indicate that a certain collection is more likely to generate appropriate results than another which contains images comprising predominantly primary colours.

The methodology outlined in section 3.2 is applicable to this situation - in fact, it will be simpler since the clustering and naming stage, which connects the categories in the hierarchy to the individual image descriptions, is not needed. The appropriate categories or visual templates will merely consist of specified ranges of colour features. These would be arranged in a hierarchy with all colours at the top level, and progressively narrower ranges of colour down to the lowest level. By comparing these ranges with the colour feature vectors produced by QBIC for each image, we can count how many images correspond to each category in the classification hierarchy. The collection-level metadata thus produced could then be used to direct a user to the appropriate collection(s) depending on the colour(s) that are required.

This concept provides an interesting direction for future research. Whether or not it can be carried out under the auspices of the ELISE project remains to be seen.

5 Conclusion

We have described several key aspects of the ELISE project with particular reference to the current areas of research, namely the use of quantitative collection-level metadata for guided retrieval from multiple databases in a federation, and the evaluation of image content-based retrieval methods in the ELISE context. The results of our studies will be presented in due course.

We have discussed our proposals for progressing this research, in the hope that they will provoke constructive debate within and between the image analysis and information retrieval communities.

6 Acknowledgements

We gratefully acknowledge stimulating discussions with members of the ELISE consortium, particularly Douglas Dodds of the National Art Library at the Victoria and Albert Museum, London. And we acknowledge the financial support given by the European Commission to the ELISE project.

7 References

1. Bird C. An Assessment of Current Content-Based Retrieval Techniques for Image and other Multimedia data Types, ELISE Deliverable D4.3/1, July 1997 (copies available from the author)
2. De Marsicoi M, Cinque L, Levialdi S. Indexing pictorial documents by their content: a survey of current techniques, *Image and Vision Computing* 1997;15:119-141
3. Gravano L, Chang C-C, Garcia-Molina H, Paepcke A. STARTS: Stanford proposal for Internet Meta-Searching, *SIGMOD Rec. (USA)* 1997; 26:207-218
4. Dash K, Hurson A. The semantic matrix model (SMM): a knowledge based solution to semantic homogeneity in multidatabases, *Proceedings of the 1995 ACM CIKM International Conference on Information and Knowledge Management*, 1995; 122-128.
5. Zhao J. Schema coordination in federated database management: a comparison with schema integration, *Decision Support Systems* 1997, 20:243-257
6. UKOLN Metadata Resources: Dublin Core, <http://ukoln.ac.uk/metadata/resources/dc.html>
7. Dolin R, Agrawal D, El Abbadi A, Pearlman J. Using automated classification for summarising and selecting heterogeneous information sources, *D-Lib Magazine*, 1998, <http://www.dlib.org/>
8. Flickner M, Sawney H, Niblack W et al. Query by Image and Video Content: The QBIC System, *IEEE Computer* 1995; 28:23-30
9. Bird C, Elliott P, Griffiths E. User Interfaces for Content-Based Image Retrieval, *IEE Colloquium on Intelligent Image Databases*, London 1996, 8/1-4
10. Bird C, Chapman S, Ibbotson J. Content-Driven Navigation of Large Databases, *IEE Colloquium on Intelligent Image Databases*, London 1996, 13/1-5
11. Chang W, Murthy D, Zhang A, Syeda-Mahmood T. Metadatabase and search agent for multimedia access over Internet, *Proceedings IEEE International Conference on Multimedia Computing and Systems*, 1997, 626-627