Automatic image content retrieval - are we getting anywhere?

John P Eakins

Department of Computing, University of Northumbria at Newcastle, Newcastle upon Tyne NE1 8ST, United Kingdom

Introduction

Recent advances in computer-based storage, transmission and display of multimedia data have been truly impressive. The sight (and sound) of information emanating from all parts of the planet is becoming commonplace. Initially, users are often amazed at the technology that makes it all possible. As they gain familiarity with such systems, though, their expectations can rise rapidly. Difficulties in accessing desired sites, or slow response times when downloading images, are soon recognized as inadequacies in the level of service provided, and lead to demands for improvement.

Success in meeting users' needs is not just a matter of improving transmission speeds and image quality, however. Far more fundamental issues need to be considered. Users will want to put the next generation of multimedia systems to all kinds of new and unforeseen uses. In order to remain successful, system providers will have to devise new tools to meet these needs.

One requirement that can be predicted with some confidence is the need for far more sophisticated retrieval capabilities than any current system offers. The limitations of current technology are well illustrated in a recent editorial in IEEE Multimedia, which shows just how difficult and time-consuming it is to perform a simple task such as obtaining a few images of animals from the Web (Jain, 1995).

The capability required to meet this need is content-based image retrieval (CBIR) - the ability to retrieve images from a collection on the basis of features extracted or inferred from the images themselves. Computer-based CBIR, the application of image processing and related techniques to derive retrieval features without the need for human intervention, is now a highly active area for research and development (Cawkell (1993), Gudivada & Raghavan (1995)). While most of this work is still experimental, commercial-strength systems such as QBIC (Flickner et al, 1995) and Virage (Gupta et al, 1996) are now beginning to appear.

However, the range of facilities offered by present systems is often quite limited. In many cases, retrieval capabilities appear to be governed more by what current technology can offer than by user demand. This is perhaps not surprising, given the lack of maturity of much of the underlying technology and the difficulty of identifying what users really want. However, evidence on user requirements is beginning to accumulate (Enser, 1995), even if its conclusions are somewhat tentative at present. This paper attempts to assess how successful current developments in technology are likely to be in meeting these user needs.

The need for retrieval by image content

Does a system holding collections of electronic images need to provide retrieval by image content - in other words, some kind of image classification or indexing facility? The experience of generations of picture librarians, backed up by studies of user queries in such libraries (e.g. Enser, 1993), suggests powerfully that it does. Journalists requesting photographs from a newspaper library, designers
looking for materials with a particular colour or texture, and engineers looking for drawings of a particular type of part, all have a clear idea of what they want, even if they prefer to delegate the actual searching to an intermediary. The existence - and continuing use - of detailed classification schemes such as ICONCLASS (Gordon, 1990) for art images, and the Opitz code (Opitz et al, 1969) for machined parts, reinforces this message.

Is automatic retrieval by image content (i.e. computerized extraction of index features from images, and subsequent retrieval of images containing the desired features) also worthwhile? Again, the enthusiasm of art librarians using IBM's prototype QBIC system (Holt & Hartwick, 1994) suggests that it is, drastically cutting the time needed both to catalogue an image and to process a user query. Manual techniques of picture classification and indexing are inherently time-consuming, and there is evidence that their effectiveness is often limited. Studies of inter-indexer consistency (reviewed by Enser, 1995) have revealed wide variations in choice of keywords for any given image, and analysis of user queries suggest that a given image may satisfy such a wide range of queries that it is impossible to foresee them all. Enser's rather depressing conclusion is that "if the retrieval utility of an image suffers from low predictability, the subject indexing of that image must have low utility".

The case for improved methods of image retrieval is thus a strong one. To what extent can automated systems meet this need? To answer this question, we first need a framework within which we can distinguish between different types - or levels - of image content retrieval. Potentially, one could seek to retrieve an image by many types of attribute, including the following:

- low-level attributes such as its colour, texture or shape (e.g. an image containing green triangles);
- derived attributes such as the presence or arrangement of specific types of object (e.g. chairs around a table);
- inferred abstract attributes such as a particular type of event (e.g. a football match);
- identifying attributes such as the presence of named individuals, locations, or events (e.g. the Queen opening Parliament);
- subjective attributes such as the emotions one might associate with the image (e.g. happiness);
- non-derivable attributes such as who created the image, where and when.

This wide range of potential query types is of course one of the reasons why classification and indexing of images is so difficult. Not all types of query are applicable to all types of image - a

![Image of abstract images](https://example.com/image)

*Fig 1. Examples of abstract images, taken from the UK Trade Marks registry. Crown Copyright reserved*

database of abstract images of the type shown in Fig. 1 is unlikely to be worth querying at any but the lowest level, while a natural scene such as that shown in Fig. 2 would be more likely to answer queries at one of the higher levels.
Previous authors have proposed a number of possible frameworks within which one might discuss different types of image query, including Panofsky's classification of fine art image attributes into pre-iconography, iconography, and iconology (Panofsky, 1955), Enser's classification of queries to a picture library into unique, non-unique, unique refined and non-unique refined (Enser, 1993), and my own division of image content retrieval systems into spatial, image paradigm and shape retrieval systems (Eakins, 1992). However, none of these can encompass the entire range of query types listed above. Hence a new framework is proposed, as indicated below.

Three levels of content retrieval

The types of possible image query listed above range from the highly concrete (find all pictures containing red) to the very abstract (find pictures depicting "suffering"). As one moves towards the abstract end of this continuum, queries become harder to formulate and to execute, and opinions on what constitutes successful retrieval become more diverse. Hence a classification of query types in order of increasing abstraction seems appropriate.

Level 1, the lowest level, comprises retrieval by what Gudivada & Raghavan (1995) describe as primitive features such as colour, texture, shape or the spatial location of image elements. This level could perhaps be further divided into the following sub-categories:

- retrieval by colour feature (find all pictures containing yellow and blue regions);
- retrieval by texture (find images containing regions with texture similar to that of woven cloth);
- retrieval by shape (find drawings similar in shape to this sketch);
- retrieval by spatial location (find pictures with objects of interest in the top left-hand corner);
- retrieval by a combination of the above (find images containing yellow stars arranged in a ring);

though these distinctions can appear somewhat artificial to the user, since few experimental or commercial systems now limit their retrieval capabilities to a single type of feature. This kind of retrieval is mainly required for specialist applications such as location of drawings in a design archive (Eakins, 1993), registration of trade mark images (Eakins, 1994), or colour matching of fashion

---

1 corresponding to the shape retrieval category identified in my earlier review (Eakins, 1992).
2 corresponding to the spatial information retrieval category identified in my earlier review.
Level 2 comprises retrieval by derived attributes (Gudivada & Raghavan describe these as logical features), involving some degree of logical inference about the identity of the objects depicted in the image. It is also sometimes referred to as retrieval by semantic content, and corresponds both to the image paradigm retrieval distinguished in my earlier review, and to Panofsky's pre-iconographic level of picture description. This level of retrieval can usefully be further divided into:

- retrieval of objects of a given type (find pictures of a passenger train crossing a bridge);
- retrieval of individual objects or persons (find a picture of Nelson's column in Trafalgar Square).

The first of these sub-categories corresponds in part to Enser's non-unique and non-unique refined categories of pictorial query, the second to his unique and unique refined categories (Enser, 1993). This level of retrieval is probably of more general applicability than level 1 - particularly for representational images of the type shown in Fig. 2. Many of the queries analysed in Enser's study of newspaper picture libraries fall into this overall category, though often qualified by further search criteria which cannot be readily derived from the image - an example being a request for "North London street scenes during the 1950's" (Enser, 1993).

Level 3 comprises retrieval by abstract attributes, involving a high degree of abstract - and possibly subjective - reasoning about the meaning and purpose of the objects or scenes depicted. Again, this level of retrieval can usefully be subdivided, as follows:

- retrieval of named events or types of activity (find pictures of English folk dancing);
- retrieval of pictures with emotional or symbolic significance (find a picture depicting "atonement").

The difference between these two sub-categories is in some respects similar to Panofsky's distinction between iconography (describing a picture's actual subject matter) and iconology (its deeper artistic or religious significance). Although somewhat specialized, queries at this level can be encountered in both newspaper and art libraries, and pose a major challenge to the ingenuity of the searcher.

The current state of the art

At level 1, significant progress has been made by exploiting the results of computer vision research. Experimental and even a few commercial systems now exist which can provide content-based retrieval on the basis of primitive features such as colour or shape. Systems are typically example-based, allowing users to submit a query image and retrieve the most similar stored images from a database. Some types of feature - particularly colour - have proved more effective as retrieval cues than others.

Retrieval by colour, in which colour histograms are generated for each stored image (or selected regions within the image) and matched with a histogram extracted from the query image, was first introduced by Swain & Ballard (1991). It has since been significantly improved by researchers such as Stricker & Orengo (1995), and can now generate some quite impressive-looking results, particularly when combined with some element of spatial matching (Stricker & Dimai, 1996). Retrieval by texture, comparing values of automatically-extracted features such as contrast, coarseness, directionality and regularity as defined by Tamura et al (1978), has also been achieved with some degree of success, particularly when used in conjunction with colour matching (Niblack et al, 1993).

Shape retrieval has proved a considerably greater challenge, despite considerable research into the topic (Niblack et al, 1993). At present, most shape retrieval systems operate by using well-established...
image-processing techniques to compute a set of shape features for identified regions of interest within each stored image. Queries are then answered by computing the same set of features for the query image, and retrieving those stored images whose features most closely match those of the query. Two main types of shape feature have been used (see Fig. 3) - *global* features such as aspect ratio, circularity and moment invariants (Niblack et al, 1993) and *local* features such as points of high curvature (Ramesh & Sethi, 1995) or sets of consecutive boundary segments (Mehrotra & Gary, 1995). However, the effectiveness of current shape retrieval systems still leaves a lot to be desired. Evaluation results reported with IBM's QBIC system (Faloutsos et al, 1994) show clearly that its retrieval effectiveness is markedly better for colour matching than for shape matching. And there is worrying evidence that few, if any, of the shape feature measures in current use are accurate predictors of human judgements of shape similarity (Mumford, 1991; Scassellati et al, 1994).

Current commercial systems such as QBIC and Virage offer retrieval by colour, texture or shape, either singly or in combination. They provide the user with a range of interfaces, to allow query by image example, by menu selection of colour or texture, or by hand-drawn sketch. In order to improve search efficiency, they use multidimensional index structures such as the R*-tree (Beckmann et al, 1990), which avoid the need to match the query with every image in the database. Despite the limitations of the underlying technology, such systems can provide useful results, particularly in the hands of a skilled user. Of all current systems, Virage is perhaps the most interesting. Unlike its competitors, it is designed as a series of independent modules, which can be combined in a variety of ways. This makes it easy to extend the system by building in new types of query interface or additional customized modules to process specialized collections of images such as trademarks. Alternatively, Virage modules can be incorporated within existing database management products to extend their image-handling capabilities. An on-line demonstration of the Virage system is available on the World-Wide Web at [http://www.virage.com/online/](http://www.virage.com/online/).

The common thread running through all systems supporting level 1 retrieval is their reliance on automatic extraction and matching of primitive features such as boundary shape characteristics or colour histograms. Feature extraction may be performed once and for all when the images are first added to the database. Alternatively, as with MIT's Photobook system (Pentland et al, 1994), features may be computed at search time as required, giving greater flexibility at the expense of speed. Either way, the emphasis is on concrete features which are relatively easy to extract from raw image data - and, more importantly, can be derived purely from information present in the images themselves. The underlying assumption of all current level 1 systems is that there is no need to refer to any external knowledge base when indexing a new image or processing a query.

At level 2, progress has been more limited. Retrieval at this level involves matching derived or abstract attributes of the image such as the presence of specified types of building or animal in a scene. This kind of feature matching is qualitatively different, as it requires not just feature analysis, but a complex inferential process, making reference to some stored knowledge base of object paradigms. No amount of colour, texture or shape analysis can tell us on its own whether a particular image represents a dog, since dogs come in many shapes, sizes and colours. The ease with which humans can recognize an image as a dog conceals a complex and as yet incompletely-understood process. For this reason, there are many who consider that the only way to provide level 2 retrieval is to employ human indexers to add suitable keywords (Gudivada & Raghavan, 1995).

However, some researchers are attempting to bridge the gap between primitive-level and semantic-level retrieval. Although elements of this approach were present in the early IMAID system (Chang & Fu, 1981), the first systematic attempt to tackle this problem appears to be that of Rabitti & Stanchev (1989). Their GRIM_DBMS was capable of retrieved line drawings of objects of a given type within a narrow predefined domain, using a schema constructed by the systems designer as its knowledge base.
The system analysed object drawings, labelling each with a set of possible interpretations and their probabilities. These were then used to derive likely interpretations of the scene within which they appeared. Object and scene interpretations could be queried using an SQL-like command language. Despite the fact that the domain of objects chosen to illustrate the system was severely restricted, and that no retrieval effectiveness results were reported, GRIM_DBMS remains important as an indicator of how automatic semantic image retrieval might be achieved.

More recently, Hermes et al (1995) have used similar techniques to derive natural-language descriptors directly from images of outdoor scenes for their IRIS system. Colour, texture, region and spatial information from the image are input to a graph parser which derives the most likely interpretation of each region within the image, and hence of the overall scene, generating text descriptors which can be input to any text retrieval system (see Fig. 3). Like GRIM_DBMS, IRIS uses a complex reasoning process, linked to an external knowledge base, to derive image interpretations; also like GRIM_DBMS, the process of building up the knowledge base requires considerable human effort. However, its image analysis and interpretation capacity is much more highly-developed than GRIM_DBMS. Its success, albeit within a highly restricted domain, reinforces the hypothesis that level 2 retrieval is basically a deductive reasoning process which can in principle be automated.

Fig 3. Combining different aspects of an image to generate automatic labelling of a natural scene in IRIS.

At level 3, the problems have so far proved insurmountable. Identification of the objects or individuals in a scene is a formidable enough task; attempting to understand the significance of their presence or location in the scene well-nigh impossible without considerable intellectual effort. It might be possible to teach an image understanding system that a combination of small human beings around a table bearing a cake topped with candles can be interpreted as a child's birthday party. But identifying the connotations of a picture of, say, the Jarrow march is clearly not a logical reasoning process, and it is difficult to see how retrieval at this level could ever be automated. Even the former task requires access to a knowledge base of formidable complexity. Prospects for viable image retrieval at this level seem vanishingly small - though this has not deterred Kato (1992) from attempting to index a
collection of paintings by generating subjective descriptors such as "warm" and "romantic" directly from their colour histograms. From the details reported, his approach seems somewhat naive.

In summary, significant progress has been made in delivering usable image retrieval at level 1 - though even here there is still considerable scope for improvement in both search efficiency and effectiveness. By contrast, progress at levels 2 and 3 has been extremely slow, though some potentially promising lines of investigation have been identified. Evidence both from informal discussions with picture librarians and from studies such as Enser (1993) makes it quite clear that while retrieval at level 1 might be nice to have, most users' real needs can be met only by providing higher levels of retrieval. There is thus a significant gap between users' needs and the capabilities of even the most advanced of today's experimental systems. What are the prospects of bridging this gap in the foreseeable future?

Research issues

At level 1, the retrieval process itself appears to be relatively well understood, and the main task is to improve both its efficiency and effectiveness. Better ways need to be found of identifying salient image features for a given range of applications, particularly those characterizing object shape. Improved methods of recognizing key features in noisy or distorted images are also required. Further advances in indexing and matching efficiency are needed to handle ever larger collections of images. All these areas are actively being researched, and significant improvements in the performance of image retrieval systems operating at this level can reasonably be expected.

At level 2 and beyond, the main problem is that the retrieval process is not well understood. The main block to progress is the process of interpreting an image in order to generate a sufficiently rich set of logical features for retrieval. As indicated above, many consider the task of automatically generating such interpretations too far beyond the limits of current technology to be worth pursuing (Gudivada & Raghavan, 1995). However, the experiments of Rabbitti & Stanchev (1989) and Hermes et al (1995) suggest otherwise. Given improved mechanisms for building up a knowledge base of image feature interpretations, it should be possible to construct systems that offer level 2 image retrieval within restricted but non-trivial domains, such as photographs of sporting events.

A number of potentially fruitful lines of approach are available. One is the use of neural networks for object recognition - a technique that has already been used with some success in the image retrieval field (Rickman & Stonham, 1993). Neural networks certainly have the flexibility to distinguish between images of, say, dogs and sheep - though since their reasoning process cannot be examined, one can never have complete confidence in their conclusions. Another is the use of ideas arising from studies into human visual cognition, such as those of Marr (1982) and Biederman (1987), who attempt to explain the mechanisms by which humans recognize and identify concrete objects such as a chair or a bicycle. Biederman, for example, has outlined a possible process involving edge detection, primitive feature extraction and component identification as intermediate steps. Whether or not this eventually proves valid for human vision, it can still provide a useful starting-point for designers of automatic object recognition systems.

This approach offers a real possibility of worthwhile progress at level 2 - though it clearly still leaves a lot of questions unanswered, particularly the structure of the knowledge base needed to support the recognition process, and how it would be populated. Even at level 1, such an approach can prove useful. My colleagues and I (Shields et al, 1995; Eakins et al, 1996) are currently developing a shape retrieval system for trade mark images known as ARTISAN, which incorporates elements of

3 Automatic Retrieval of Trademark Images by Shape ANalysis
Biederman's hypothesis, grouping image elements into patterns by recognizing non-accidental properties such as shape similarity, symmetry and collinearity. These are used both to "clean up" noise-affected images and to group image components into families from which additional shape features can be extracted (Fig. 4). Early results from this project (Fig. 5) are encouraging.

A further research area we cannot afford to neglect concerns the user. We need to know much more about why people request images, what they do with them, and how they judge the relevance of retrieved images to their needs. More comprehensive studies of the type conducted by Enser (1993) are urgently needed. Without this knowledge, we cannot adequately evaluate image retrieval systems; without systematic evaluation, we cannot expect systems to deliver results which users find acceptable. Even at level 1, where the issue reduces to comparing human and machine judgements of shape or texture similarity, very little work has been reported (a notable exception being Scassellati et al (1994)). While such studies will not in themselves solve the problems of image retrieval, they are a necessary prerequisite to success.
Conclusions

There seems little doubt that automatic image content retrieval will be seen as a desirable, if not essential, facility in the next generation of multimedia systems. Given current research trends, it should be feasible to equip such systems with effective low-level automatic image content retrieval facilities. The development of systems capable of matching users' real needs is a much more formidable task. However, the problems involved in meeting these needs, at least up to level 2, do not appear insoluble, and a number of possible ways of tackling these problems have already been identified. They should provide a worthwhile challenge to the research community for some time to come.

Acknowledgements

Thanks are due to J M Boardman for his development work on the ARTISAN project, to T Hermes for providing images from IRIS, and the British Library and the UK Patent Office for their financial support.

References

Beckmann, N et al (1990) "The R*-tree: an efficient and robust access method for points and rectangles" ACM SIGMOD Record 19(2), 322-331


Flickner, M et al (1995) "Query by image and video content: the QBIC system" IEEE Computer, 28(9), 23-32


Holt, B and Hartwick, L (1994) "Retrieving art images by image content: the UC Davis QBIC project" Proc First International Conference on Electronic Library and Visual Information System Research, de Montfort University, Milton Keynes, 93-100


Marr, D (1982) "Vision" Freeman, San Francisco


Petkovic, D (1996) "Query by Image Content" oral presentation to Storage and Retrieval for Image and Video Databases IV, San Jose, California, February 1996


Tamura, H et al (1978) "Textural features corresponding to visual perception" IEEE Transactions on Systems, Man and Cybernetics 8(6), 460-472