PERGAMON

Pattern Recognition 35 (2002) 3–14

PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY

www.elsevier.com/locate/patcog

Towards intelligent image retrieval

John P. Eakins*

Institute for Image Data Research, University of Northumbria at Newcastle, Newcastle upon Tyne NE1 8ST, UK

Abstract

Research into techniques for the retrieval of images by semantic content is still in its infancy. This paper reviews recent trends in the field, distinguishing four separate lines of activity: automatic scene analysis, model-based and statistical approaches to object classification, and adaptive learning from user feedback. It compares the strengths and weaknesses of model-based and adaptive techniques, and argues that further advances in the field are likely to involve the increasing use of techniques from the field of artificial intelligence. © 2001 Pattern Recognition Society. Published by Elsevier Science Ltd. All rights reserved.

Keywords: CBIR; Semantic image retrieval; Image understanding; Machine learning; Review

1. Introduction

Content-based image retrieval (CBIR) — the retrieval of images on the basis of features automatically derived from the images themselves — is now a thriving field for research and development, with reports of new techniques appearing almost daily. As the field has matured, the nature of the problems faced by researchers and developers has inevitably changed. Much early research, exemplified by projects such as TRADEMARK [1], QBIC [2] and Photobook [3], was concerned primarily with establishing the feasibility of retrieving images from large collections using automatically-derived features. More recent research (see [4–6] for recent comprehensive reviews) has concentrated on identifying improved techniques for CBIR, including new types of feature, representation method and matching technique. Now the feasibility of the underlying technology has been demonstrated, effort can be devoted to the crucial question of how to design and build systems that successfully meet real user needs.

Most current CBIR techniques are geared towards retrieval by some aspect of image appearance, depending on the automatic extraction and comparison of image features judged most likely to convey that appearance. The features most often used include colour [7,8], texture [9,10], shape [11,12], spatial layout [13], and multi-resolution pixel intensity transformations such as wavelets [14] or multi-scale Gaussian filtering [15]. At least three CBIR packages making use of such techniques are now commercially available: QBIC from IBM (http://www.qbic.almaden.ibm.com/), the VIR Image Engine from Virage, Inc (http://www.virage.com/), and VisualRetrievalWare from Excalibur, Inc (http://www.excalib.com/).

While the technology behind current CBIR systems is undoubtedly impressive, user take-up of such systems has so far been minimal. This is not because the need for such systems is lacking–there is ample evidence of user demand for better image data management in fields as diverse as crime prevention, photo-journalism, fashion design, trademark registration, and medical diagnosis [16,17]. It is because there is a mismatch between the capabilities of the technology and the needs of users. The vast majority of users do not want to retrieve images simply on the basis of similarity of appearance. They need to be able to locate pictures of a particular type (or individual instance) of object, phenomenon, or event [18].

Gudivada and Raghavan [16] have drawn a useful distinction between retrieval by *primitive* image feature (such as colour, texture or shape) and *semantic* feature (such as the type of object or event depicted by the

* Tel.: + 44-191-227-4539; fax: + 44-191-227-4637.
E-mail address: john.eakins@unn.ac.uk (J.P. Eakins).

0031-3203/01/$20.00 © 2001 Pattern Recognition Society. Published by Elsevier Science Ltd. All rights reserved.
PII: S0031-3203(01)00038-3

image). Eakins [19] has taken this distinction further, identifying three distinct levels of image query, each of which can be further subdivided:

- Level 1, retrieval by *primitive* features such as colour, texture, shape or the spatial location of image elements (e.g. "find all pictures containing yellow or blue stars arranged in a ring").
- Level 2, retrieval by derived attribute or *logical* feature, involving some degree of inference about the identity of the objects depicted in the image (e.g. "find pictures of a passenger train crossing a bridge").
- Level 3, retrieval by *abstract* attribute, involving complex reasoning about the significance of the objects or scenes depicted (e.g. "find pictures illustrating pageantry").

Using this framework, the extent of the mismatch between user requirements and the capabilities of the technology becomes clear. Although the volume of research into user needs is not large, the results of those studies which have been conducted to date (e.g. [18]) suggest strongly that very few users need level 1 retrieval. The majority of image queries received by picture libraries are at level 2, though a significant number (particularly in specialist art libraries) are at level 3. The overwhelming majority of CBIR systems, both commercial and experimental, offer nothing but level 1 retrieval. A few experimental systems now operate at level 2, but none at all at level 3.

What are the prospects of bridging what has been referred to as the *semantic gap* [16], and delivering the image retrieval capabilities that users really want? This paper aims to answer this question by reviewing current research into semantic image retrieval, with particular emphasis on the contribution which techniques from related fields such as artificial intelligence (AI) are making to developments in this area. CBIR may have its roots in the field of classical image analysis; it relies on many standard image analysis techniques, such as convolution, edge detection, pixel intensity histogramming, and power spectrum analysis. But a successful solution to the problems of semantic image retrieval (if one exists at all) may well require a significant paradigm shift, involving techniques originally developed in other fields. CBIR has already benefited greatly from insights derived from related fields. A prime example of this process is the technique of relevance feedback [20], originally developed for text retrieval, where users indicate the relevance of each item of output received, and the system amends its search strategy accordingly. Relevance feedback is showing considerable promise in the image retrieval area, largely because users can rapidly judge the relevance of a retrieved image within seconds. It has now been successfully implemented in several experimental CBIR systems [21,22]. Other examples where CBIR has benefited from insights from related fields include relatively efficient direct access via multidimensional indexing, from the database management field [23], and retrieval by subjective appearance, drawing on Gestalt psychology [24].

AI, defined by Luger and Stubblefield [25] as "*the study of the mechanisms underlying intelligent behaviour though the construction and evaluation of artefacts that enact those mechanisms*", appears a particularly promising source of ideas for advancing the art of semantic image retrieval. It aims to develop techniques which allow a machine to:

- reason from available knowledge, even when incomplete or conflicting;
- generate solutions using heuristics where no algorithmic answer is feasible;
- interact with the environment and learn from past experience;
- use higher-level knowledge in problem-solving, and handle semantic issues;
- generate output matching that of a human expert;

in other words, to exhibit intelligent behaviour, defined by Newell and Simon [26] as "*behaviour appropriate to the ends of the system and adaptive to the demands of the environment*". Most observers would agree that assessing the contents of a set of images in order to decide their relevance to a query was indeed a task requiring intelligence in this sense. In the context of image retrieval, *the end of the system* is the identification of a set of images from a collection which meets a user's perhaps subjective and poorly-formulated need, and *adaptive to the demands of the environment* implies that the system should offer flexibility in allowing different modes of user interaction, and learn from user feedback.

2. The need for intelligent image retrieval

One crucial difference between primitive and semantic-level retrieval seems to lie in the extent of intelligent behaviour needed to decide whether a given image meets the specified search criteria. At the primitive level, images can normally be matched by algorithmic means purely on the basis of information contained within the images themselves. For example, colour similarity matching requires nothing more than the computation and comparison of two histograms representing the distribution of pixel colours across the two images. There is no requirement for what might be considered intelligent behaviour–reference to an external knowledge base, reasoning with conflicting or incomplete data, or learning from past experience.

Semantic retrieval requires the identification of images depicting desired types of object, scene, event, or abstract idea. According to the definition above, this is a process requiring intelligence, as it requires reasoning about the nature and significance of primitive visual cues from the image, and their relationships to each other and to the viewer's past experience. This latter aspect appears to be of crucial importance. Even at the simplest level (such as recognizing a curved yellow region in an image as a banana), extraction of an image's semantic content seems to require reference to some external store of knowledge. To identify a banana in an image requires experience of the range of colour, shape and texture combinations which have characterized previously-encountered examples, and the ability to use this knowledge to predict which yellow curved regions are in fact bananas, and which (say) parts of yellow rubber rings.

Identifying even a relatively simple artefact such as a chair is a rather more complex process. Since chairs come in a wide variety of colours, textures and shapes, primitive image features are unlikely to suffice on their own. The problem of recognizing a chair is not perceptually more difficult than that of recognizing a banana. The difference lies in the degree of interpretation necessary. Recognition of an object as a chair requires reference to some higher-level model, defining spatial, structural and perhaps other constraints. Such a model needs to be susceptible to modification, to include the possibility that new designs of chair may appear in the future (not a problem one would expect to encounter with bananas!). Humans build up and refine such a model automatically from past experience: for machines, the process is less straightforward. The need to gain such experience directly is one reason why Brooks [27] has advocated designing robots in humanoid form.

Identifying complex human artefacts is still more problematic. Experienced engineers can readily recognize a pressure-limiting valve in an engineering drawing, even though its actual shape may vary considerably — presumably because their training enables them to draw reasonable inferences from the appearance and layout of key components, as well as the nature of any larger structures in which they appear. But even a highly intelligent human would find such a task impossible without the requisite engineering training. The need to update one's mental model of a specialist device of this kind is likely to be even greater than for an everyday object such as a chair, since new designs are likely to appear at frequent intervals.

Yet another layer of complexity is encountered when trying to interpret scenes depicting specific types of event. To recognize a photograph as that of a child's birthday party demands not only the identification of objects which might be present in such scenes (young human figures, balloons, lighted candles), but a further level of reasoning about the relationship of these objects to each other and the extent to which these conform to prior expectations of what occurs at such events. Again, the ability to update such mental models in the light of changing circumstances is crucial.

The issues surrounding human recognition and classification of images have been extensively studied by Rosch et al. [28]. The most significant findings from these studies in the present context are as follows:

- Humans naturally categorize objects they encounter into basic categories such as *chair* or *banana*. Although visual appearance is of major importance in identifying these classes, other factors such as commonality of the motor movements needed to interact with such items (such as grasping with the fingers) also play a part in such characterization.
- The basic category appears to be a favoured level of abstraction for many purposes. Participants in experiments in free-naming of pictures, for example, overwhelmingly preferred to use basic category names rather than more specialized or generalized levels (*hammer* rather than *tool* or *claw-hammer*, for example). Developmental studies with young children show that basic category names are learnt earlier in life than those of other levels.
- Basic categories generally have a higher proportion of attributes common to all member of that class than subordinate or superordinate categories. In many (but not all) cases it is possible to construct an 'averaged' shape from typical members of the class which humans can readily recognize.

These findings give some indication of the likely success of semantic image retrieval techniques which rely on automatic derivation of object or scene labels from visual features of the image. Such techniques are most likely to succeed for objects within an image which correspond to basic classes (such as *banana* or *horse*) whose members share a strong visual similarity. For such objects it should be possible to construct or learn suitable object models permitting recognition of typical examples of each class. For other types of object (such as *bird* or *tree*), a similar approach based on visual similarity of subclasses (probably, though not necessarily, based on existing taxonomic divisions such as *sparrow*, *parrot* or *eagle*) may prove more effective. For object classes where many defining attributes are non-visual (such as *chair* or *pump*), however, this approach appears doomed to failure — though the fact that humans can recognize such objects from visual cues alone suggests that the problem is in principle soluble.

To develop a complete understanding of image contents at the semantic level is a formidable task, well beyond the capabilities of any current machine. Fortunately, such a complete level of understanding is not an essential prerequisite for successful semantic image retrieval, as several researchers in the field have pointed out [29,30]. Empirically, a retrieval system can be regarded as successful if it has the ability to classify a sufficiently high proportion of objects sought by users accurately enough for its retrieval output to satisfy a searcher's needs. In many contexts (including photo-journalism), this means that quite low classification accuracy may be acceptable, provided the searcher can in fact find a usable picture. An analogous situation holds in text retrieval, where effective retrieval systems have been around for years, despite continuing difficulties with automatic text understanding. Unfortunately it is not yet clear what level of image understanding is in fact required for successful classification and retrieval. The only way to resolve this question appears to lie in the development and evaluation of semantic image retrieval techniques.

3. Current trends in semantic image retrieval

Research into semantic image retrieval *per se* has a relatively short history; the vast majority of papers reviewed in this article date from 1996 or later. Many of the techniques now being applied to the problem have been adapted from related areas such as 'classical' object recognition or machine learning, and it is not always easy to distinguish between research into image understanding for its own sake and research motivated by a desire to develop better storage and retrieval systems. As yet, it is difficult to discern any body of techniques or hypotheses which belong solely to the field of semantic image retrieval. This is possibly an indication of the relative immaturity of the field. However, semantic image retrieval is a topic of growing research interest — at least at level 2 as defined above (retrieval by derived attribute such as the type of object or scene depicted). Several different areas of activity can be distinguished within the field, though many of the techniques used are common to more than one area, and the distinctions between different approaches are not always clear-cut:

- Automatic scene classification in whole images, typically using statistically-based techniques.
- Automatic object classification, using one of the following alternative approaches:
  - knowledge-based techniques based on detailed object models;
  - statistical techniques similar to those used for scene classification.
- Methods for learning and propagating labels assigned by human users.

Examples of each of these approaches are discussed in detail below.

By contrast, no significant research has yet been reported into CBIR at level 3 (retrieval by abstract attribute such as *freedom*). The issues involved are dauntingly complex. Little is known about the way in which humans interact with images at this level, making it almost impossible even to identify potentially fruitful lines of investigation.

4. Automatic scene classification

Automatic classification of scenes (into general types such as *indoors*, *city street* or *beach*) can be useful, both because this is an important filter which can be used when searching, and because this can help in identifying specific objects present. One of the earliest systems of this type was IRIS [31], which used a reasoning approach based on a combination of colour, texture, region and spatial information to derive the most likely interpretation of the scene, generating text descriptors such as *mountain*, *forest* or *lake* for input to a text retrieval system. Later researchers have identified much simpler techniques for scene analysis. For example, Oliva et al. [32] have used shape characteristics of whole-image power spectra sampled with Gabor filters to classify scenes by placing them on appropriate points on two semantic axes: artificial vs natural, and open vs closed. Szummer and Picard [33] use a combination of colour histograms, texture measures and discrete cosine transform (DCT) coefficients to train a nearest-neighbour classifier to distinguish between indoor and outdoor scenes. Empirical tests showed the method to have 90% accuracy in classifying a set of 1300 colour photographs. Lipson et al. [34] propose a different approach, based on qualitative reasoning from templates specifying expected combinations of colour layout for prototype scenes such as *mountain* or *field*. They report 75% accuracy in classifying photographs of mountains, with 12% false positives. Vailaya et al. [35] have developed a Bayesian classifier to group images into a number of semantically meaningful categories, including city vs landscape and forest vs mountain, using codebook vectors generated by vector quantization from feature vectors based on colour moments and Gabor coefficients. Reported accuracy is better than 90% for most classification tasks.

The difficulty of judging the accuracy of systems of this kind is illustrated by Paek et al. [36], who developed a prototype system for classification of news photographs into indoor and outdoor scenes based both on keywords in caption text and histograms of colour edge direction distributions. Their system achieved 86% classification accuracy—but easily outperformed Szummer and Picard's method [33], which achieved only 74% accuracy with this test data set. Another potential problem is the choice of category to which images are assigned. This often appears to have been subjectively chosen by the experimenters themselves, significantly limiting the validity of their results. One exception to this is the hierarchical classifier developed by Vailaya et al. [35] for vacation images, which used automatic clustering techniques based on subjective classification decisions made by an independent user panel to generate an objectively-defined set of categories and classification results to serve as ground truth.

Systems of this kind provide a degree of classification for semantic image retrieval; Permitting automatic assignment of keywords such as *beach,* m*ountain* or *city scene* to appropriate images. At least at present, they tend to represent a fairly basic approach to semantic retrieval, in some cases just classifying images as *X* or *not-X*, where *X* might be *beach* or *city scene*. Inevitably, this raises questions over scalability. Most of the really high success rates have been reported with *X*/*not-X* classifiers. The extent to which such systems can be said to exhibit intelligent behaviour as defined in Section 1 is debatable. They make little use of high-level knowledge, and have little ability to reason from incomplete or conflicting information. In particular, few have any ability to continue learning when in operational use — an exception being the incremental learning system described by Vailaya and Jain [37]. This solves the problem of assigning appropriate weights to new and old training data by recreating an approximation to the original training set from the codebook vectors generated by vector quantization, and then repeating the classification process with the enlarged training set.

5. Automatic object recognition

Automatic object recognition is clearly important for semantic image retrieval. The ability to identify a given type of object in a scene is useful both as an end in itself, and as an intermediate step in the interpretation of more complex scenes. Vailaya and Jain [38], for example, have proposed a scheme for semantic indexing of an image which combines scene identification using global features and object detection using local features. The two main approaches used to date for automatic object recognition can be described respectively as knowledge-based and statistical, though the distinction is not entirely clear-cut.

5.1. Knowledge based object recognition

One of the most effective methods for object recognition in an image can be to specify a model for each type of object of interest, and then examine the image for regions conforming to that model. In this way, the past experience needed to understand the image is embedded into the object models used by the software developed to process the image. One of the earliest implementations of these principles was Brooks' ACRONYM system [39], which used generalized shape modelling to identify and locate instances of desired objects in aerial photographs. After an initial edge detection step, descriptions of possible objects of interest were derived in the form of generalized cones, shapes generated by sweeping a given cross-section along a defined trajectory. A set of production rules was then used to infer the presence of specified types of aircraft from the pattern of cones derived from the image. Similar strategies were used in Matsuyama and Hwang's SIGMA aerial image interpretation system [40], which used a frame-based approach in a so-called *interpretation cycle* of progressive object recognition, and Draper's SCHEMA [41], which used a blackboard architecture to combine low-level image tokens into plausible object interpretations.

Systems such as ACRONYM were not developed with any specific application in mind — their designers' main aim was to demonstrate the feasibility of knowledge-based techniques for scene interpretation. Later researchers have adapted these techniques for use in the specific context of CBIR. One of the best-established systems of this type is PICTION [42], which identifies human faces in natural scenes by matching candidate face shapes generated by multi-resolution edge detection techniques with a simple three-contour model of hairline and left and right face contours. A more sophisticated technique of this kind is based on the extraction of so-called Composite Visual Objects [43]. Effectively, this technique is an expert system for recognizing and characterising visual objects made up of simple connected regions. Each composite object has to be defined as a model consisting of one or more components. Images of (clothed) humans, for example, can be modelled as a specific arrangement of primitives such as face, hair, jacket, or hat. Their human detector separates out image regions using conventional means, and then attempts to reason about their likely identity by comparing them with the model.

Perhaps the most highly-developed technique in this area is that of Forsyth et al. [29]. Their approach is based on developing a model of each class of object to be recognized, and using this to build up evidence for or against the object's presence in the image. Evidence could include features of the candidate region itself (colour, shape or texture), or contextual information such as its position and the type of background in the image. Object classification is a three-stage process: (a) segmenting images into coherent regions using a combination of edge, colour and texture information; (b) fusing colour, texture and shape information to identify possible descriptions of each region (for example, as a human arm); and (c) classifying objects from their constituents in terms of component descriptions. The method has been applied with some success to the identification of a range of object types, including unclothed human bodies, horses and trees, though the retrieval effectiveness of the system is fairly modest at present (a typical score in experiments with the horse classifier was 15% recall at 66% precision[1]).

Model-based systems of this kind are perhaps more obviously 'intelligent' than the scene classifiers described in Section 4 above. They certainly make greater use of AI techniques of the type listed in Section 1. Their underlying object models are often implemented as large and carefully-structured knowledge. They are clearly capable of reasoning about the nature of the objects in an image, using base data that may often be sparse or inconsistent. Reasoning often uses heuristic methods, guided by higher-level knowledge – the models themselves. The problems they face — shared by many expert systems — are firstly that their knowledge is often very domain-specific, and that they can therefore handle only a very restricted set of object types, and secondly that their knowledge is embedded by their designers. Hence they have no mechanism for improving their performance by learning. As with the scene classifiers, it is not clear how well they can scale up from situations where image objects are classified as *X* or *not-X*, to a real-life situation where there may be hundreds, if not thousands, of different types of objects to distinguish. And rich though their underlying models may be, little success has yet been achieved in incorporating the kinds of non-visual knowledge which are clearly needed to recognize examples of some types of object [28].

5.2. Statistical techniques for object recognition

A conceptually simpler approach to image interpretation, which does not require the construction of any high-level object model, is the use of statistical techniques (often very similar to those used in scene classification) to assign appropriate semantic labels to individual regions within an image. A good example of this approach is provided by Campbell et al. [44], who use a combination of colour and texture features to train an radical basis function (RBF) network to distinguish between 11 different types of region in a scene, including sky, vegetation, road, building, fence and 'mobile object' — typically a car. They report over 80% accuracy in classifying over 3700 regions from 350 images. Vailaya and Jain [38] have adapted their whole-image classification technique [35], based largely on local colour and texture measures, to the recognition of sky and vegetation in outdoor images, with encouraging preliminary results. They plan to extend the technique to classify a larger range of region types.

Similar work is reported from a number of other laboratories. Martinez and Serra [45] have used discriminant analysis based on feature vectors derived by principal component analysis (PCA) from images convolved with Gaussian derivatives to classify images into a variety of categories, including animals, humans, cars and houses. Little information is provided on the effectiveness of their approach. Belongie et al. [46] have developed a so-called *blobworld* representation of image regions, based on segmentation by colour and texture features using the expectation-maximization algorithm. Although they do not claim that their technique offers semantic retrieval, they show that it can be used to retrieve images of objects such as tigers and aircraft from a database. Leung and Malik [47] have developed a method for identifying material within textured regions of an image (as leather, cork, plaster, etc) using microstructures known as *3-D textons* derived from primitive texture measures. At a more specialist level, Bregler and Malik [48] have used some novel texture measures to train a hierarchical mixture of experts (HME) classifier capable of distinguishing between five different types of vehicle from surveillance videos. And Schneiderman and Kanade [49] have shown that a Bayesian classifier based on vectors derived by PCA from pixel intensities in image subregions sampled at three levels of resolution can correctly detect over 90% of human faces in an image collection with a false positive rate of less than 25%, outperforming earlier face classifiers based on back-propagation networks [50].

An idea of potentially wide applicability is the use of statistically-generated *visual classes* for object recognition, proposed by Schiele and Crowley [51]. This aims to get round the problem of variability in appearance of objects such as chairs by identifying a number of specific visual classes of chair, each of which is sufficiently homogeneous to be identified purely by visual appearance. These could be detected using the authors' earlier technique of object recognition using multidimensional receptive field histograms [52]. Unfortunately the authors provide no convincing evidence that their concepts model reality sufficiently well to be of any practical use for generic object recognition. It is not clear how (if at all) visually homogeneous subclasses of objects such as chairs can be identified; nor is it clear that multidimensional receptive field histograms are useful for identifying different instances of a given class of visually similar objects, as opposed to the same object at different 3-D orientations. A rather more well-developed technique for recognising objects (such as oranges) or types of material (such as sand) in an image is the method of Buijs and Lew [30] for inducing 'simple semantics' from primitive image features. They do this by identifying both positive and negative example images, identifying a subset of primitive features with high discriminating power, and using these to train a minimum distance classifier.

Statistical approaches have the advantage of not requiring the construction of complex and possibly domain-specific models of each type of object to be

[1] Precision is defined as the percentage of retrieved objects relevant to the query; recall as the percentage of objects in the entire database relevant to the query.

recognized, though they obviously suffer from the lack of any high-level knowledge about the domain, relying totally on statistical associations between image semantics and quantifiable low-level properties, learnt in most cases from a training set of a few hundred examples at best. When judged by the criteria of Section 1, these techniques may appear less 'intelligent' than the model-based approaches described above, because of the lack of high-level reasoning capabilities — or even, in many cases, the ability to cope with missing or uncertain information. This does not necessarily make them less useful.

6. User-assisted retrieval techniques

The problems of achieving effective semantic image retrieval by purely automatic means have led many researchers to investigate methods which retain some degree of human intervention during the actual operation of the system, either at input or search time. Many of the techniques described here are very similar to the statistical object or region classifiers described in Section 5.2. However, there is one crucial difference: the techniques described here are capable of continuous learning through run-time interaction with end-users, while those described in Section 5.2 learn only during their initial construction.

Semantic retrieval techniques depending on user interaction cover a broad spectrum of intelligence. A relatively low-level (though again potentially very useful) technique for bridging the semantic gap is *content-based navigation*, the use of generic links specified in terms of text or image features. This can be used to construct a *multimedia thesaurus* [53] specifying semantic relationships between source items in the link database, whether text, image or sound. This allows system users to build up a database of semantic relationships between text terms and their corresponding images. However, human intelligence is still required to establish linkages between image types and their semantic meanings. The system itself acts purely as a repository for this knowledge; it provides no mechanism for automated reasoning or learning.

More obviously 'intelligent' is a family of techniques based on extensions of the relevance feedback principle. One of the earliest systems to provide this kind of interaction was FourEyes [54], which allowed a user to group arbitrary regions of images (such as particular types of building, or species of plant), and optionally give these regions semantic labels such as *grass* or *sky*. Once the user has assigned labels to several examples of the same type from one or more images, the system attempts to induce grouping rules from the positive and negative examples at its disposal. A number of learning techniques were compared for this purpose, with set coverage [55] proving the most successful with the training examples used. FourEyes then uses these rules to assign labels to new examples sharing the same range of feature values. Feedback from the user can be used to refine the selection rules. Effectively, then the system can learn what areas of *grass* and *sky* look like, and can then search for images containing such areas. Another method for propagating human annotation of image objects is described by Frederix and Pauwels [56]. This segments images into regions of interest using unsupervised clustering techniques, extracts suitable shape features from region boundaries, and searches the database to identify annotated examples of similar shapes. Appropriate annotations can then be propagated to the new image. As yet few technical details appear to be available.

The Semantic Visual Template approach proposed by Chang et al. [57] is based on similar principles, though here labelling is attached to queries rather than image regions. Users are requested to identify a set of possible low-level feature combinations which might meet their semantic query. A sunset, for example, would contain large areas of colours such as orange or purple, and possibly a bright circular object. The system then identifies regions of primitive feature space enclosing all the examples given, and generates an initial set of query icons. This query set can be refined using relevance feedback techniques, and stored in a query database for later use and possible modification. Similar considerations have motivated the development of Wood et al.'s Image Database Query System [58]. This uses a two-stage training procedure to derive a semantic classifier capable of identifying objects such as sailboats, mountains and faces. In the interactive phase, a searcher uses relevance feedback to train an learning vector quantization (LVQ) algorithm to recognize positive and negative examples of the desired concept. This interaction can be stored for re-use and possible modification. This is followed by an off-line phase in which output vectors from the LVQ step are used to train an RBF classifier which can then be used to search for regions corresponding to the desired object class.

A novel framework for capturing semantic information about an image collection through relevance feedback is reported by Lee et al. [59]. This incorporates a second feedback loop so that users' input is remembered permanently, and used to store semantic links between images as well as similarity of appearance. Initially, images are clustered purely on the basis of primitive feature similarity. Users who search the system are asked to indicate which retrieved images are relevant and which irrelevant. This information is then used to split and merge clusters of similar images, gradually introducing an element of semantic similarity in the process. The authors refer to this as 'warping feature space'. Evaluation results on a collection of 1000 images pre-clustered into 50 groups suggest that repeated use of the system does indeed yield a steady increase in search precision.

Jaimes and Chang's *Visual Apprentice* [60] aims to provide users with a general framework for building up visual classes which can represent specified types of object or scene. Users can define a visual class by specifying labels for objects and their key constituent parts, together with a set of training examples in which image regions are labelled according to the class definition. The system then uses a combination of lazy learning, decision trees and genetic algorithms to build up a hierarchical object definition in which image regions generated by primitive-level segmentation routines are grouped progressively into *perceptual areas* (groups of regions likely to be perceived as a whole), object parts, whole objects and scenes. The system is still at a relatively early stage of development, though is already capable of quite impressive results, retrieving images containing visually distinctive objects such as ships and elephants from small image collections with over 90% recall and 70% precision. The authors are quick to point out that their approach is not suitable for objects whose visual appearance is more variable.

These techniques exhibit at least one aspect of intelligent behaviour as described in Section 1, in that they share the ability to build up a knowledge base of past experience through user interaction. The nature and location of this knowledge base differs between systems. In FourEyes, for example, linkages between primitive and semantic features are regarded as a property of the image region (type, not instance); in the Semantic Visual Templates method, such information is a property of the query (instance, not type). All such techniques possess the ability to reason with incomplete information, since they inevitably begin with an empty knowledge base. They cannot however benefit from the kind of higher-level knowledge built into the model-based systems described above. Hence their reasoning is inevitably *ad hoc*, and the quality of their knowledge base depends crucially on the quality of their past interaction with users.

7. Comparison of techniques

Even though one can argue that there is a linkage between the degree of 'intelligence' exhibited by the techniques described above and their retrieval effectiveness, the latter is clearly more important than the former. Unfortunately this is not easy to establish. Researchers in this area are now beginning to publish their evaluation results, normally in the form of precision/recall figures, but it is seldom possible to draw valid comparisons between systems. Different sets of researchers have tested out their systems on different collections of images, using different measures of effectiveness, different sources of query images, and different ways of judging the correctness of system output. Since in the last analysis, image retrieval systems have to model subjective human judgements, reliable ground truth is hard to come by. There are few widely-available collections of images available for comparative studies, let alone sets of standard queries and relevance judgements to provide essential ground truth for comparative studies. And there is still remarkably little awareness of the need to obtain independent relevance judgements for evaluating system effectiveness. Judgements made by members of the development team are inherently flawed, as shown clearly by the experiments of Squire and Pun [61] comparing human and machine performance in partitioning images into similar groups.

Two main current approaches to semantic retrieval can be distinguished, though each has been applied in a variety of ways — model-based scene and object recognition, and statistical techniques for semantic labelling of whole images or regions via human interaction. The latter class of approaches can be divided further, into those where system training is carried out at the design stage (e.g. [51]), and those where user interaction is an integral part of the learning process (e.g. [54]). Each of these approaches has its strengths and weaknesses. It was argued in Section 2 that some store of prior experience was essential for interpreting image data at the semantic level. Both types of approach have mechanisms for constructing a set of semantic interpretation mechanisms for primitive-level data which could constitute such a store. The key difference is that model-based techniques tend to have a richer—but less flexible–set of interpretation mechanisms than the statistical methods. So is one approach superior to the other? At present, too little is known about the knowledge structuring and reasoning processes involved in image interpretation to draw any firm conclusions. In the short term, the potential depth and richness of high-level object models of the type developed by Forsyth et al. [29] appear more likely to be able to provide a successful basis for high-level image interpretation than the often *ad hoc* associations on which statistical methods depend. A further problem with both types of statistical technique is that they assume that, given enough instances, the mapping between primitive image features and previous human semantic judgements can always be learnt. Learnability theory [62] suggests that this kind of assumption is not always valid. While some classes of concept can provably be learnt to a given degree of accuracy in a finite time by the repeated presentation of examples, the extent of the set of learnable classes is not at present known. In the longer term, however, one suspects that the problems of updating and extending complex model-based approaches such as Forsyth's to cover more than a toy subset of object classes will prove insuperable unless they too make use of some form of adaptive learning. It is notable that Brooks, the developer of the first model-based scene interpreter ACRONYM, now favours a very different architecture for intelligent systems, based on cooperation between large numbers of simple autonomous agents [27].

8. Future prospects for intelligent image retrieval

Current research into semantic-level image retrieval is clearly at a very early stage, and it will be a long time before any generally useful systems emerge. As indicated above, both currently-favoured approaches have their strengths and weaknesses. Model-based approaches to object classification can be powerful, but are often very limited in scope and lack the ability to learn from experience. They need to develop automatic mechanisms for adapting the structure and content of their models as they encounter new examples and counter-examples of the objects they have been designed to classify. And one suspects that they will need more powerful knowledge-structuring methods to cope with any realistic domain of object types. Statistical approaches have the enormous potential advantage of being able to learn from experience — provided the learning process does not cease at the design stage, but continues to incorporate new knowledge gained from user feedback. However, current techniques for incorporating such feedback make little or no attempt to structure their knowledge bases, which limits their ability to reason with higher-level knowledge. They are crucially dependent on the value of the input they receive from users. Since they lack any mechanism to control the quality of user input, they inevitably incorporate feedback from all users — even if idiosyncratic or malicious — in the same fashion. This suggests that the effectiveness of such systems could be markedly improved by better initial structuring of their knowledge bases, and the use of more appropriate learning paradigms, as outlined by Brooks et al. [27].

It is far from certain that any current approach will lead to effective semantic image retrieval. Santini and Jain, for example, have argued that true semantic retrieval is unachievable, as users' views on which images are relevant to a particular query are so variable that no generalization is possible [63]. They therefore propose an enhanced form of relevance feedback, in which users can visualise search results in several dimensions, exploring the relationships between underlying primitive features and their current information needs. From this perspective, it can actually be detrimental to equip a system with complex semantic models or the ability to learn permanent associations between semantic concepts and primitive features, as this simply interferes with users' freedom to manipulate search space. Only time will tell if this proves to be a realistic view.

In the opinion of this reviewer, such a view is unduly pessimistic. The advances made over the past three years provide grounds for at least cautious optimism about future prospects for semantic image retrieval. It has to be conceded, though, that the problems remain formidable, and the nature of the underlying mechanisms needed to support semantic image retrieval remain the subject of speculation. Analysis of the work reviewed above suggests the following conjectures about these mechanisms. Although none is readily quantifiable, or susceptible to mathematical proof, they should all be empirically testable in the same sense as Newell and Simon's physical symbol system hypothesis [26] — they can be confirmed or refuted by the development of future systems.

- Image retrieval at the semantic level can be achieved only by reference to some knowledge base of prior experience. Evidence for this hypothesis comes from consideration of the systems reviewed above. In every case, it is possible to identify either a repository of past knowledge about the domain in which they operate, or a mechanism for building up such a store. Such a feature is seldom, if ever, seen in primitive-level systems.
- The higher the level of interpretation required to answer a semantic query, the larger and more complex the knowledge base and reasoning mechanisms needed to perform this interpretation. Consideration of the examples discussed in Section 2 above suggests that there are clear differences between the extent of reasoning needed to identify an image of (say) a chair, and a child's birthday party. One can therefore postulate that the reasoning capabilities required for the former task would be a subset of those needed to accomplish the latter.
- Successful semantic retrieval involving images of complex objects or scenes requires an adaptive system capable of learning from experience. Models of almost any type of object need to be built up and refined over a period of time in the light of experience. To expect system designers to get every detail of their models right first time is unreasonable. Even if they were to do so, another problem would remain. The visual properties of almost any human artefact are likely to change — perhaps quite frequently — over time. It is simply not feasible for system designers to modify an object model every time a visually different example of that object is encountered. The system itself has to be able to adapt.

If the conjectures outlined above do turn out to reflect the processes underlying semantic image retrieval, one can expect a significant proportion of further advances in this field come from the domain of AI. As indicated above, a number of specific AI techniques have already been applied to image retrieval at the semantic level, including rule-based reasoning, neural networks, and genetic algorithms. However, these have often been applied to peripheral aspects of the problem in hand, and there appears to be considerable scope for the more systematic application of AI techniques and concepts.

Adaptive learning is perhaps the prime example here. The potential of techniques such as case-based reasoning [64], explanation-based learning [65], reasoning by analogy [66] and conceptual clustering [67] to provide systematic learning capabilities for image retrieval systems remains largely untapped. The opportunities for developing truly intelligent image retrieval systems by combining techniques from the fields of image processing and artificial intelligence are considerable.

References

[1] T. Kato, Database architecture for content-based image retrieval, in: A.A. Jambardino, W.R. Niblack (Eds.), Image Storage and Retrieval Systems, Proc SPIE 1662, 1992, 112–123.

[2] W.R. Niblack et al., The QBIC project: querying images by color, texture and shape, IBM Research Report RJ-9203, 1993.

[3] A. Pentland et al., Photobook—tools for content-based manipulation of image databases, Storage and Retrieval for Image and Video Databases II, Proc SPIE 2185, 1994, pp. 34–47.

[4] F. Idris, S. Panchanathan, Review of image and video indexing techniques, J. Visual Commun. Image Representation 8 (2) (1997) 146–166.

[5] Y. Rui et al., Image retrieval current techniques, promising directions, and open issues, J. Visual Commun. Image Representat. 10 (1) (1999) 39–62.

[6] M. de Marsicoi et al., Indexing pictorial documents by their content: a survey of current techniques, Image and Vision Comput. 15 (1997) 119–141.

[7] M.J. Swain, D.H. Ballard, Color indexing, Int. J. Comput. Vision 7 (1) (1991) 11–32.

[8] J.R. Smith, S.F. Chang, Querying by color regions using the VisualSEEk content-based visual query system, in: M.T. Maybury (Ed.), Intelligent Multimedia Information Retrieval, AAAI Press, Menlo Park, CA, 1997, pp. 23–41.

[9] F. Liu, R.W. Picard, Periodicity directionality and randomness: Wold features for image modelling and retrieval, IEEE Trans. Pattern Anal. Mach. Intell. 18 (7) (1996) 722–733.

[10] B.S. Manjunath, W.Y. Ma, Texture features for browsing and retrieval of large image data, IEEE Trans. Pattern Anal. Mach. Intell. 18 (1996) 837–842.

[11] R. Mehrotra, J.E. Gary, Similar shape retrieval in shape data management, IEEE Comput. 28 (9) 57–62, Sep 1995.

[12] A.K. Jain, A. Vailaya, Image retrieval using color and shape, Pattern Recognition 29 (8) (1996) 1233–1244.

[13] V.N. Gudivada, V.V. Raghavan, Design and evaluation of algorithms for image retrieval by spatial similarity, ACM Trans. Inform. Systems 13 (2) (1995) 115–144.

[14] K.C. Liang, C.C.J. Kuo, Implementation and performance evaluation of a progressive image retrieval system, in: I.K. Sethi, R.C. Jain (Eds.), Storage and Retrieval for Image and Video Databases VI, Proc SPIE 3312, 1998, 37–48.

[15] S. Ravela, R. Manmatha, Retrieving images by appearance, Proceedings of IEEE International Conference on Computer Vision (ICCV98), Bombay, India, 1998, pp. 608–613.

[16] V.N. Gudivada, V.V. Raghavan, Content-based image retrieval systems, IEEE Comput. 28 (9) (1995) 18–22.

[17] J.P. Eakins, M.E Graham, Content-based image retrieval, JISC Technology Applications Programme Report 39, October 1999. Available online at http://www.unn.ac.uk/iidr/CBIR/report.html.

[18] L. Armitage, P.G.B. Enser, Analysis of user need in image archives, J. Inform. Sci. 23 (4) (1997) 287–299.

[19] J.P. Eakins, Techniques for image retrieval, Library and Information Briefings 85, British Library and South Bank University, London, 1998.

[20] G. Salton, The SMART retrieval system — experiments in automatic document processing, Prentice-Hall, New Jersey, 1971.

[21] Y. Rui et al., "Relevance feedback techniques in interactive content-based image retrieval" in: I.K. Sethi, R.C. Jain (Eds.), Storage and Retrieval for Image and Video Databases VI, Proc SPIE 3312, 1997, pp. 25–36.

[22] C. Meilhac et al., "Relevance feedback in Surfimage" Proceedings of Fourth IEEE Workshop on Applications of Computer Vision (WACV'98), 1998, pp. 266–267.

[23] N. Beckmann, The R*-tree: an efficient and robust access method for points and rectangles, ACM SIGMOD Record 19 (2) (1990) 322–331.

[24] M. Wertheimer, "Untersuchungen zur Lehre von der Gestalt", Psycholog. Forschung 4 (1923) 301–350. (Translated as laws of organization in perceptual forms, in: W.D. Ellis (Ed.), A Sourcebook of Gestalt Psychology, Humanities Press, New York, 1950).

[25] G.F. Luger, W.A. Stubblefield, Artificial Intelligence, Addison-Wesley, Reading, MA, 1997.

[26] A. Newell, H.A. Simon, Computer science as empirical inquiry: symbols and search, Commun. ACM 19 (3) (1976) 113–126.

[27] R.A. Brooks et al., Alternative essences of intelligence, Proceedings of 15th National Conference on Artificial Intelligence (AAAI-98), 1998, pp. 961–968.

[28] E. Rosch et al., Basic objects in natural categories, Cognitive Psychol. 8 (1976) 382–439.

[29] D.A. Forsyth et al., "Finding pictures of objects in large collections of images", in: P.B. Heidorn, B. Sandore (Eds.), Digital Image Access and Retrieval: 1996 Clinic on Library Applications of Data Processing Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign, 1997, pp. 118–139.

[30] J.M. Buijs, M.S. Lew, Visual learning of simple semantics in ImageScape, VISUAL99: Third International Conference on Visual Information and Information Systems. Lecture Notes in Computer Science, Vol. 1614, Springer, Berlin, 1999, pp. 131–138.

[31] T. Hermes et al., Image retrieval for information systems, in: W.R. Niblack, R.C. Jain (Eds.), Storage and Retrieval for Image and Video Databases III, Proc SPIE 2420, 1995, pp. 394–405.

[32] A. Oliva et al., Global semantic classification of scenes using, power spectrum templates, CIR-99: The Challenge of Image Retrieval, Newcastle upon Tyne, UK, February 1999.

[33] M. Szummer, R. Picard, Indoor-outdoor image classification, IEEE International Workshop on Content-based Access of Image and Video Databases (CAIVD98), Bombay, India, 1998, pp. 42–51.

[34] P. Lipson et al., Configuration-based scene classification and image indexing, Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR-97), Puerto Rico, 1997, pp. 1007–1013.

[35] A. Vailaya et al., On image classification: city images vs landscapes, Pattern Recognition 31 (12) (1998) 1921–1936.

[36] S. Paek et al., Integration of visual and text-based approaches for the content labeling and classification of photographs, ACM SIGIR'99 Workshop on Multimedia Indexing and Retrieval, Berkeley, CA, August 1999.

[37] A. Vailaya, A.K. Jain, Incremental learning for Bayesian classification of images, IEEE International Conference on Image Processing(ICIP'99), Kobe, Japan, October 1999.

[38] A. Vailaya, A.K. Jain, Detecting sky and vegetation in outdoor images, in Storage and Retrieval for Media Databases 2000, Proc SPIE 3972, January 2000, pp. 411–420.

[39] R.A. Brooks, Model-based three-dimensional interpretations of two-dimensional images, IEEE Trans. Pattern Anal. Mach. Intell. 5 (2) (1983) 140–150.

[40] T. Matsuyama, V. Hwang, SIGMA: a knowledge-based aerial image understanding system, Plenum, New York, 1990.

[41] B.A. Draper et al., The SCHEMA system, Int. J. Comput Vision 2 (1989) 209–250.

[42] R.K. Srihari, Automatic indexing content-based retrieval of captioned images, IEEE Computer 28 (9) (1995) 49–56.

[43] G. Durand et al., Extraction of composite visual objects from audiovisual materials, in: S. Panchanathan et al. (Eds.), Multimedia Storage and Archiving Systems IV, Proc SPIE 3846, 1999, pp. 194–203.

[44] N.W. Campbell et al., Interpreting image databases by region classification, Pattern Recognition 30 (4) (1997) 555–567.

[45] A. Martinez, J.R. Serra, Semantic Access to a Database of Images: an approach to object-related image retrieval, Proceedings of IEEE Multimedia Systems (ICMCS), Florence, Italy, 1999, pp. 624–629.

[46] S. Belongie et al., Color and texture-based image segmentation using EM and its application to content-based image retrieval, Proceedings of IEEE International Conference on Computer Vision (ICCV-98), Bombay, India, 1998, pp. 675–682.

[47] T. Leung, J. Malik, Recognizing surfaces using three-dimensional textons, Seventh IEEE International Conference on Computer Vision (ICCV-99), Vol. 2, Corfu, Greece, 1999, pp. 1010–1017.

[48] C. Bregler, J. Malik, Learning appearance based models: mixtures of second moment experts, in: M.C. Mozer et al. (Eds.), Advances in Neural Information Processing Systems 9, MIT Press, 1997, pp. 845–851.

[49] H. Schneiderman, T. Kanade, Probabilistic modeling of local appearance and spatial relationships for object recognition, Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR 98), Santa Barbara, CA, 1998, pp. 45–51.

[50] H.A. Rowley et al., Neural network-based face detection, IEEE Trans. Pattern Anal. Mach. Intell. 20 (1) (1998) 23–38.

[51] B. Schiele, J.L. Crowley, The concept of visual classes for object classification, Proceedings of SCIA'97, Tenth Scandinavian Conference on Image Analysis, Lappeenranta, Finland, 1997, pp. 43–50.

[52] B. Schiele, J.L. Crowley, Object recognition using multidimensional receptive field histograms, Proceedings of Fourth European Conference on Computer Vision, Cambridge, UK, 1996, pp. 610–619.

[53] P.H. Lewis et al., Towards multimedia thesaurus support for media-based navigation, in: A.W.M. Smeulders, R.C. Jain (Eds.), Image Databases and Multimedia Search, World Scientific, Amsterdam, 1997, pp. 111–118.

[54] T. Minka, An image database browser that learns from user interaction, MIT Media Laboratory Technical Report No. 365, 1996.

[55] R.S. Michalski, A theory and methodology of inductive learning, Artificial Intelligence 20 (2) (1983) 111–161.

[56] G. Frederix, E.J. Pauwels, Automatic interpretation based on robust segmentation and shape extraction, VISUAL99: Third International Conference on Visual Information and Information Systems, Lecture Notes in Computer Science, Vol. 1614, Springer, Berlin, 1999, pp. 769–776.

[57] S.F. Chang et al., Semantic visual templates: linking visual features to semantics, in IEEE International Conference on Image Processing (ICIP'98), Chicago, Illinois, 1998, 531–535.

[58] M.E.J. Wood et al., "Iterative refinement by relevance feedback in content-based digital image retrieval", Proceedings of ACM Multimedia 98, Bristol, UK, 1998, 13–20.

[59] C.S. Lee et al., Information embedding based on users' relevance feedback for image retrieval, in: S. Panchanathan et al., (Eds.), Multimedia Storage and Archiving Systems IV, Proc SPIE 3846, 1999, 294–304.

[60] A. Jaimes, S.F. Chang, Model-based classification of visual information for content-based retrieval, Storage and Retrieval for Image and Video Databases, Proc SPIE 3656, 1999, pp. 402–414.

[61] D. McG Squire, T. Pun, A comparison of human and machine assessments of image similarity for the organization of image databases, Proceedings of SCIA'97, Tenth Scandinavian Conference on Image Analysis, Lappeenranta, Finland, 1997, pp. 51–58.

[62] L.G. Valiant, A theory of the learnable, Commun. ACM 27 (1984) 1134–1142.

[63] S. Santini, R.C. Jain, "Do images mean anything?" Proceedings of IEEE International Conference on Image Processing (ICIP-97), 1997, 564–567.

[64] J.L. Kolodner, Case-based Reasoning. Morgan Kaufmann, San Mateo, CA, 1993.

[65] T.M. Mitchell et al., Explanation-based generalization: a unifying view, Mach. Learning 1 (1) (1986) 47–80.

[66] R.P. Hall, Computational approaches to analogical reasoning: a comparative study, Artif. Intell. 39 (1) (1989) 39–120.

[67] D.H. Fisher, Knowledge acquisition via incremental conceptual clustering, Mach. Learning 2 (1987) 139–172.

About the Author—JOHN P. EAKINS is Principal Lecturer in Computing and Director of the Institute for Image Data Research at the University of Northumbria at Newcastle, United Kingdom. He graduated from the University of Cambridge in Natural Sciences, holds a Master's degree in Information Studies from Sheffield University, and a Doctorate in Computing Science from the University of Newcastle upon Tyne. His main research interests lie in the development and evaluation of CBIR systems. He has authored over 20 conference presentations and journal articles on this and related topics, and was co-chair of the 1998 and 1999 UK Challenge of Image Retrieval conferences.