Content Based Image Retrieval using Semantic Visual Categories

C.P. Town D. Sinclair

AT&T Laboratories Cambridge
24a Trumpington Street
Cambridge, England

Abstract

This paper demonstrates an approach to content based image retrieval founded on the semantically meaningful labelling of images by high level visual categories. The image labelling is achieved by means of a set of trained neural network classifiers which map segmented image region descriptors onto semantic class membership terms.

It is argued that the semantic terms give a good estimate of the salient features which are important for discrimination in image retrieval. Furthermore, it is shown that the choice of visual categories such as grass or sky which mirror high level human perception allows the implementation of intuitive and versatile query composition interfaces and a variety of image similarity metrics for content based retrieval.

1 Introduction

Advances in technology such as digital cameras, scanners, storage media, and large online picture archives have led to a proliferation of both professional and personal collections of digital images. However, many of these lack effective, consistent and scalable indexing schemes which are intuitive to the user and do not rely on extensive manual annotations. It is consequently very difficult to organise, browse, or retrieve images based on their visual content.

In order to derive image content descriptors which more closely resemble people's own perceptions, vision researchers are increasingly realising the importance of *learning* [4]. There are also a number of schemes which rely on classifying segmented image regions, for example [2] and [11]. However, in most cases the classification is based primarily on low-level region descriptors rather than semantically more relevant visual categories (IBM's QBIC [7] system is an example of this).

This paper demonstrates that it is possible to reliably label segmented regions of general images with a set of high-level visual categories of which humans already have an innate understanding. The approach taken here hinges on an iterative mapping from a vast problem domain with a large number of non-descriptive and highly localised features (the original pixel images) to an increasingly smaller feature space with associated semantics (a covering set of classified regions). The main stages are:

Image segmentation: digital images are segmented into non-overlapping regions and sets of properties are computed for each region.

Classification: the standardised and normalised region descriptors are fed into artificial neural network classifiers, one for each visual category.

Query composition and retrieval: the semantic labelling is used to define similarity metrics for content based image retrieval.

Although this method relies on a pre-defined set of semantic categories and a sufficiently large body of corresponding manually classified data for supervised neural network training, these potential disadvantages are more than compensated for by the expressive power and flexibility of user-transparent semantic indexing terms. This is an important difference between the methods put forward in this paper and those classification and visual keyword schemes which largely rely on the automatic generation of more abstract (and hence less intuitive) categories for content description ([6], [5]).

2 Image segmentation

The segmentation scheme chosen to supply region information to the trainable classifiers in this paper is given in [12] and [13]. Initially full three colour edge detection is performed, seed points for region growing are generated from the peaks in the distance transform of the edge image, and regions are then grown agglomeratively from seed points with gates on colour difference with respect to the boundary colour and mean colour across the region. Edges act as hard boundaries during the region growing process. A texture model based on discrete ridge features is also used. Features are clustered and the resulting clusters used to unify regions which share significant portions of the same feature cluster.

The segmentation scheme returns a region map together with internal region properties. Region properties include the following (see [12]:

- Region label, area, and boundary length
- Colour centre (R,G,B) and colour covariance matrix (3x3)
- Nonant membership histogram (how the pixels in a region are distributed across the image when viewed as split into 3x3 non-overlapping sub-rectangles)
- Texture feature orientation and density descriptors
- Gross region shape descriptors based upon area second moments

Figure 1 shows a sample image from the Corel picture library and its segmentation.

The number of segmented regions depends on image size and visual content, but has the desirable property that most of the image area is commonly contained within a few dozen regions which closely correspond to the salient features of the picture at the conceptual granularity of the semantic categories used here.

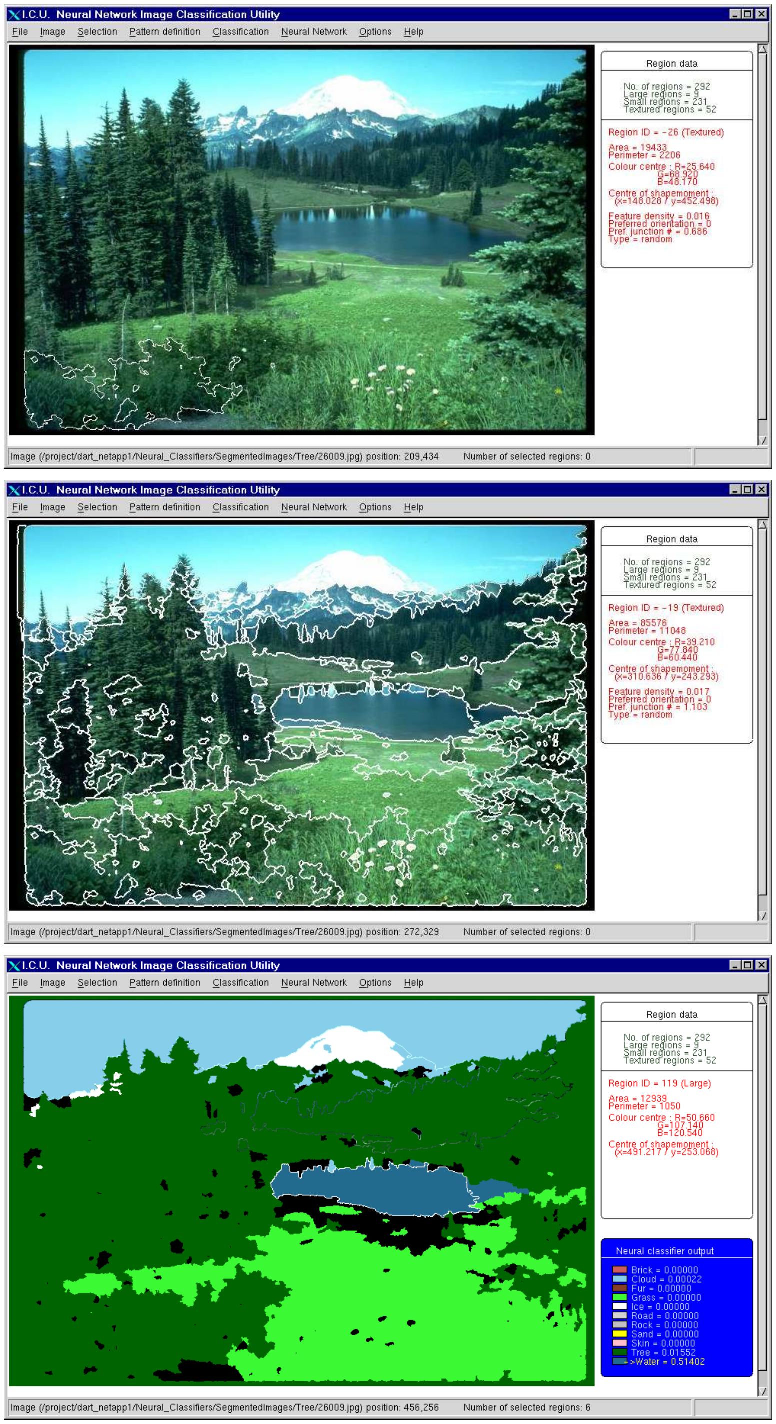

Figure 1: *This example shows the use of the tool "ICU" developed for this project. An image is loaded, the segmentation is performed and region boundaries are displayed. Then a set of neural network classifiers is applied to the image and the results are shown graphically by colouring regions according to their semantic class membership.*

3 Trainable classifiers

3.1 Neural net architectures and algorithms

Artificial neural network methods offer a variety of desirable properties, such as tolerance to noisy or incomplete input, generalisation from training data, and the ability to model almost any finite-dimensional vector function on a compact set, given a sufficient number of adaptable parameters (connection weights). However, they also suffer from a lack of conceptual transparency which makes it difficult to analyse or verify their performance.

For purposes such as classification and pattern processing, two families of networks are particularly prevalent: the general Multi-Layer Perceptron (MLP) and Radial Basis Functions (RBF), both of which were used extensively in this project. The main software package that was used to create and train classifiers is the freely available "Stuttgart Neural Network Simulator" (SNNS [14]).

This section gives a very basic overview of the primary neural network architectures and learning algorithms used in training the classifiers for visual categories. The interested reader is referred to standard literature in the field such as [1] and [10].

3.1.1 'Multi-layer Perceptrons

MLPs are feedforward neural nets with at least one hidden layer that are trained using some form of "supervised learning", most commonly a version of backpropagation. Hidden and output layer neurons operate by computing the weighted sum over all inputs x they receive multiplied by corresponding connection values w. A constant bias α is then added to this expression and the output of each neuron is set equal to the result of an activation function Φ. For i inputs, j hidden neurons, and k outputs this can therefore be written as (c.f. [9]) :

$$y_k = \Phi_o\{\alpha_k + \sum_j w_{jk}\Phi_h(\alpha_j + \sum_i w_{ij}x_i)\} \tag{1}$$

An MLP is trained by adapting the parameters (weights) according to some error measure defined by means of a *training set* of target input-output tuples. In the context of this paper, best results were obtained by using the RPROP ("Resilient backpropagation") and SCG ("Scaled Conjugate Gradients") methods.

3.1.2 Radial Basis Function networks

Radial Basis Function (RBF) networks are based on the concept of approximating a (piece-wise continuous) function by means of a finite number of basis functions. They usually have only one hidden layer and instead of inner products the Euclidean distance between the input vector and the weight vector is used in computing activation.

Training is usually a two stage process: first the input-to-hidden layer weights, which determine the centres of the radial basis functions, are set using an "unsupervised" method such as clustering. In the second stage the weights from hidden layer to outputs are adapted using a supervised training technique.

3.2 Choice of semantic categories

The project explores the paradigm of building classifiers to label image regions according to visual categories of which humans already have an innate understanding. This gives an excellent model from which to build an intuitive query composition and content based image retrieval system. The primary criterion in choosing a set of categories was to ensure that they are sufficiently well-defined in terms of the region descriptors and yet general enough to give meaningful semantic associations.

In total, 11 such categories were chosen initially:
Brick – man-made stonework such as brick walls, tiles, cobbles.
Cloud – regions of sky that are not simply of uniform colour, e.g. partially overcast sky.
Fur – animal fur and feathers.
Grass – grass-like vegetation of all sorts.
Ice – patches of ice and snow.
Road - smooth road surfaces, pavements, runways.
Rock – naturally occurring rocks, stones and pebbles.
Sand – gravel, sand, sandstone.
Skin – human and animal skin of all shades.
Tree – foliage (summer and autumn), tree bark.
Water – bodies of water that have surface structure (i.e. do not simply represent reflective surfaces).

These should in no way be taken as an exhaustive list of all possible visual categories. In particular, the current set does not explicitly represent many artificial objects (although the set has since been extended successfully by categories such as "internal walls" and "cloth"), since these have an especially high degree of variability with respect to region properties such as colour variation. The category set is therefore primarily suited to outdoor scenes.

However, the most important properties of categories and classifiers used for purposes such as image retrieval are *class separation* (the reduction of intra-class vs inter-class variability) and *consistency* rather than a perfect and complete categorisation of all the types of visual content one might encounter. The latter is inherently infeasible, as in isolated regions, tree and grass or ice and cloudy sky may be indistinguishable. Only context within an image permits disambiguation, and this requires a separate and quite distinct type of domain specific knowledge.

Intra-class variability in visual categories such as the ones defined above can still be quite high and one has to contend with a number of difficulties arising from variations in light conditions, reflections, focus, resolution, scale, occlusions, and so forth.

The choice of the set of visual categories for which to train classifiers is task dependent. Sufficiently low level categories might be thought of as undergoing a transition from noun to adjective (e.g. instead of "wall" one might assign labels such as "grey, smooth, flat, artificial"), giving the prospect of a very flexible basis language for composite object recognition. The choice presented here reflects the emphasis on region rather than object classification.

3.3 Training process

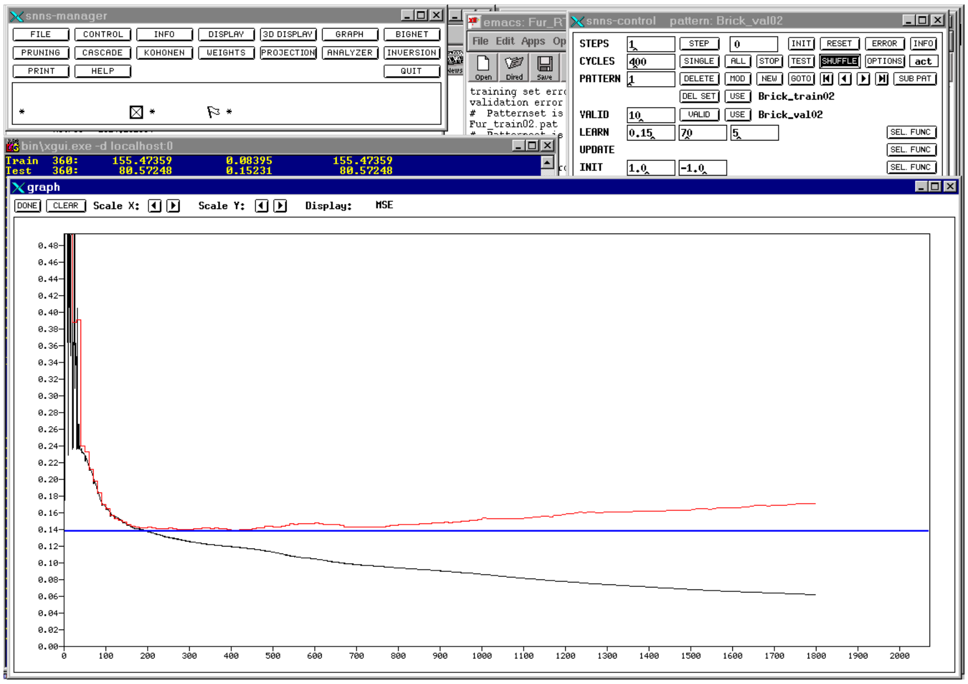

Figure 2: *SNNS training of a simple MLP neural network using the RPROP learning algorithm.* **Black** *: mean-squared error (MSE) wrt training set,* **Red** *: periodic calculation of the validation set error. Training should stop a few cycles after the validation error has become minimal and starts to rise again.*

3.3.1 Overview and training samples

The intrinsic variability of visual categories means that in order to create classifiers which are applicable in the general case, one must in effect approximate a very complex non-linear mapping from image descriptors to semantic class membership terms. Manually constructing such classifiers is a daunting task, and most attempts at building classifiers by hand are likely to involve highly subjective heuristics that impose strong constraints on the range of images to which they can usefully be applied.

In order to create statistically sound neural network classifiers, the three most important factors which have to be assessed are:

- *Model complexity*: the type and structure of the network in terms of the number of layers, units, and most importantly, the number of adaptable parameters (connection weights).

- *Training data*: the number of input–target output pairs used for classifier training.

- *Training cycles*: the number of iterations (and other learning algorithm parameters) needed to ensure that the model converges to the trend underlying the data.

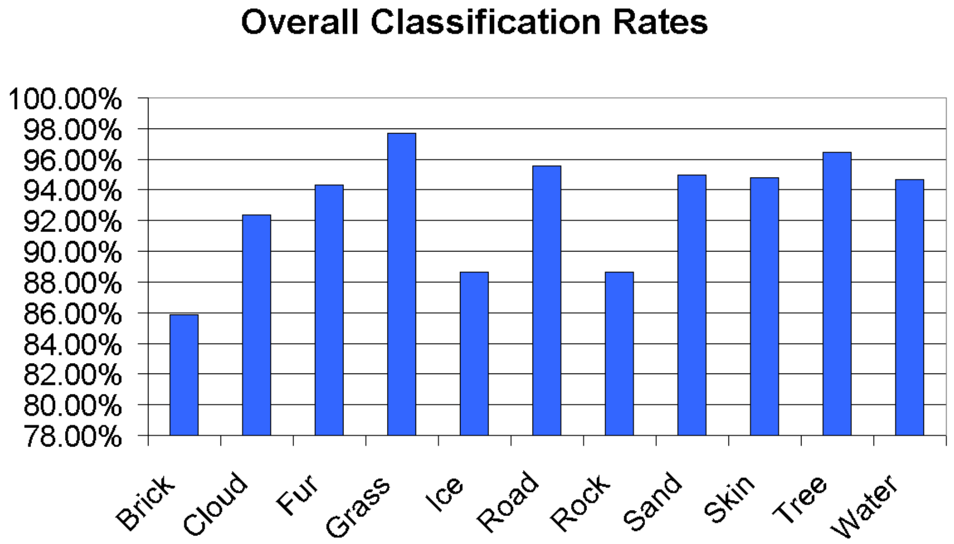

Figure 3: *Overall percentages of correct classification achieved for the visual categories on the test set.*

Too few training samples or parameters will leave the problem underdetermined and produce classifiers incapable of generalising from the data, and insufficient training does not allow convergence to a good minimum of the error measure.

In order to overcome these pitfalls, a very substantial amount of manually classified data was created, corresponding to over 4×10^4 segmented image regions chosen from over 2000 Corel picture library images. This data set was then divided into *training*, *validation* and *test* sets. A simple preprocessing stage was applied to the region descriptions output by the segmentation scheme which reduces the dimensionality of shape and texture descriptors and ensures that variation within different parameters (*e.g.* colour and texture) is of approximately the same magnitude.

3.3.2 Neural nets for classification

A separate classifier was created for each of the 11 categories as opposed to a single one with 11 outputs, primarily for reasons of better inter-class separation and because it allows one to add categories to the existing set later on or only consider a subset of the categories for a given task.

The general aim has been to create "conservative" classifiers, i.e. such that the error (false positive)

$$P(Classified \mid notMemberOfClass)$$

is smaller than the error (false negative)

$$P(notClassified \mid MemberOfClass)$$

and to achieve good class separation, i.e. regions will generally be labelled as belonging to only one (or none) of the visual categories with a high degree of certainty to yield an unambiguous classification.

Early on it became apparent that a larger number of adaptable weights (within an order of magnitude of the number of training samples) than normally recommended for statistical methods and artificial neural network training would yield better results. This is plausible, since correlations between input variables can greatly reduce the effective degrees of freedom of the input space.

Feature extraction in neural networks can generally be improved by applying prior knowledge of the input space. This was achieved by incorporating the notion of *receptive fields* (subsets of a given hidden layer are connected only to conceptually related inputs) and connecting these to a hierarchy of hidden layers of progressively smaller size to break up the mapping from input space to output space into several steps of decreasing dimensionality and increasing generality. For example, in most of the MLP classifiers inputs derived from colour histograms are initially fed to a particular group of first level hidden layer neurons which are in turn connected to the same subset of the second level layer as first level hidden neurons connected to inputs representing e.g. a 3x3 colour covariance matrix. Only at the penultimate stage are hidden neurons which encode distinct region features (colour, texture and shape) combined and then mapped to a single output. This approach is also biologically more plausible, since it introduces a neighbourhood relation between neurons of the same layer, especially if the receptive field areas are made to overlap according to the proximity of corresponding neurons.

3.3.3 Learning and stopping method

Several learning algorithms were then used to train classifiers. In order to ensure good results while maintaining the ability to generalise, a variation on the method of *early stopping* was used. The basic principle as shown in figure 2 is to use a dedicated validation set and terminate training when the network error computed for this set is about to increase. As has been pointed out by e.g. Ripley [9], this approach suffers from a number of drawbacks, since the validation error may have several poor minima and early stopping can therefore result in under-training. The approach adopted for this paper was to estimate the gradient of the validation error. If the validation error minimum was found to be sufficiently shallow, then training continued for a number of cycles determined by a validation error threshold.

3.3.4 Classification results

Figure 3 shows the classification performance for each of the original set of visual categories as measured by means of a test set. For 9 out of the 11 visual categories, best results where achieved by MLP neural networks with three hidden layers of up to 2000 neurons constructed according to the idea of "receptive fields" as described in the previous section and trained using the RPROP algorithm. In the other two cases, namely "cloud" and "skin", fully connected MLPs with two hidden layers trained with the SCG method achieved the best test set results.

RBF networks generally took less time to train but resulted in higher rates of misclassification. A possible reason for this is that using simple Euclidean vector distance between basis functions and input patterns inhibits discrimination in terms of particular subsets of region descriptors. The way RBF nets are implemented in SNNS also seems

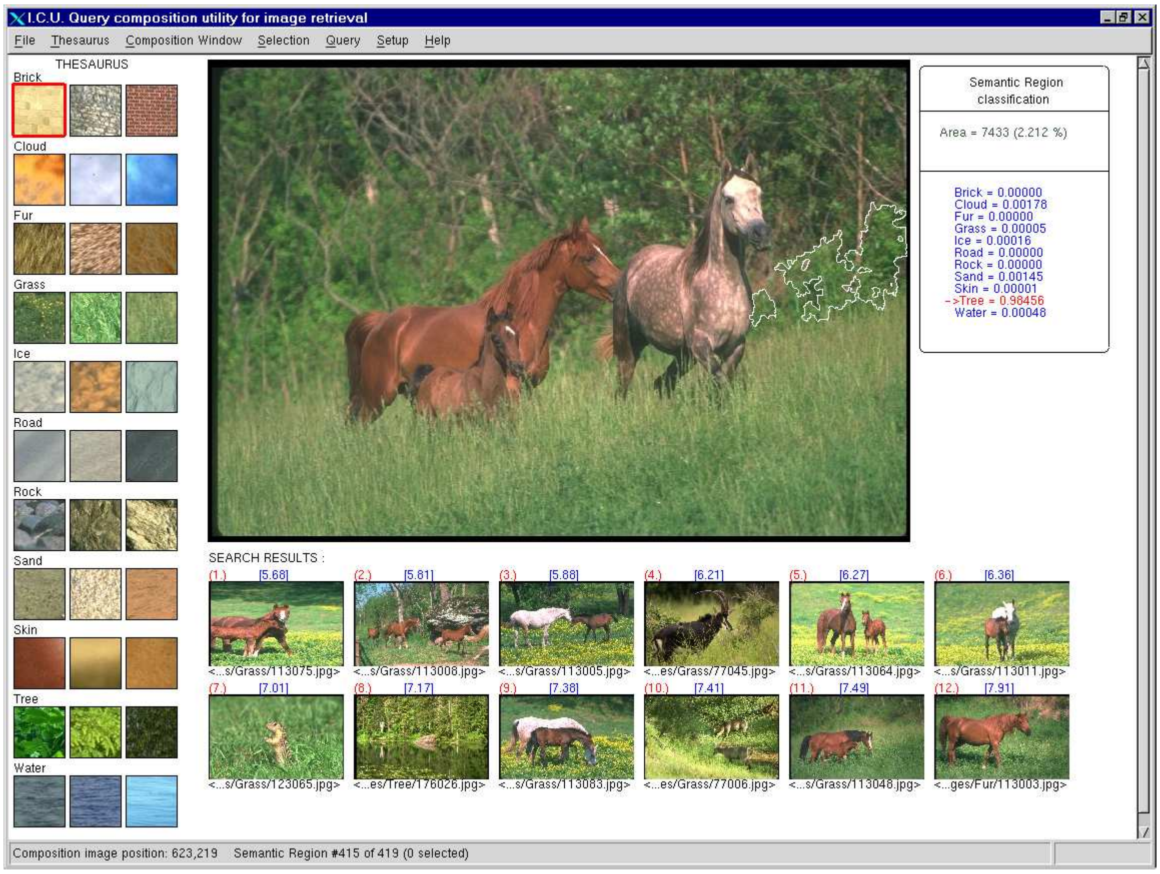

Figure 4: *User interface for the ICU_Query environment together with top search returns for a sample image query using a grid-based metric on a search corpus of approx. 20000 Corel images.*

somewhat cumbersome, and one has far less scope for architectural or algorithmic improvements than in the MLP case. Performance differences can be accounted for by the fact that some categories are inherently more variable than others, e.g. "water" exhibits widely varying visual properties according to lighting conditions, wind, nearby objects etc.. Such differences must also be reflected in providing a larger body of training data and greater model complexity.

4 Semantically labelled region based image retrieval

4.1 Query formulation and refinement

The choice of visual categories that mirror human perception allows highly intuitive query interfaces.

The retrieval system shown in figure 4 operates entirely at the level of semantically labelled image regions and abstracts away from lower-level region properties and neural classifiers. Visual queries consist of *semantic regions* positioned in a composition window (i.e. an image region with a certain classification). It is possible to define a set of

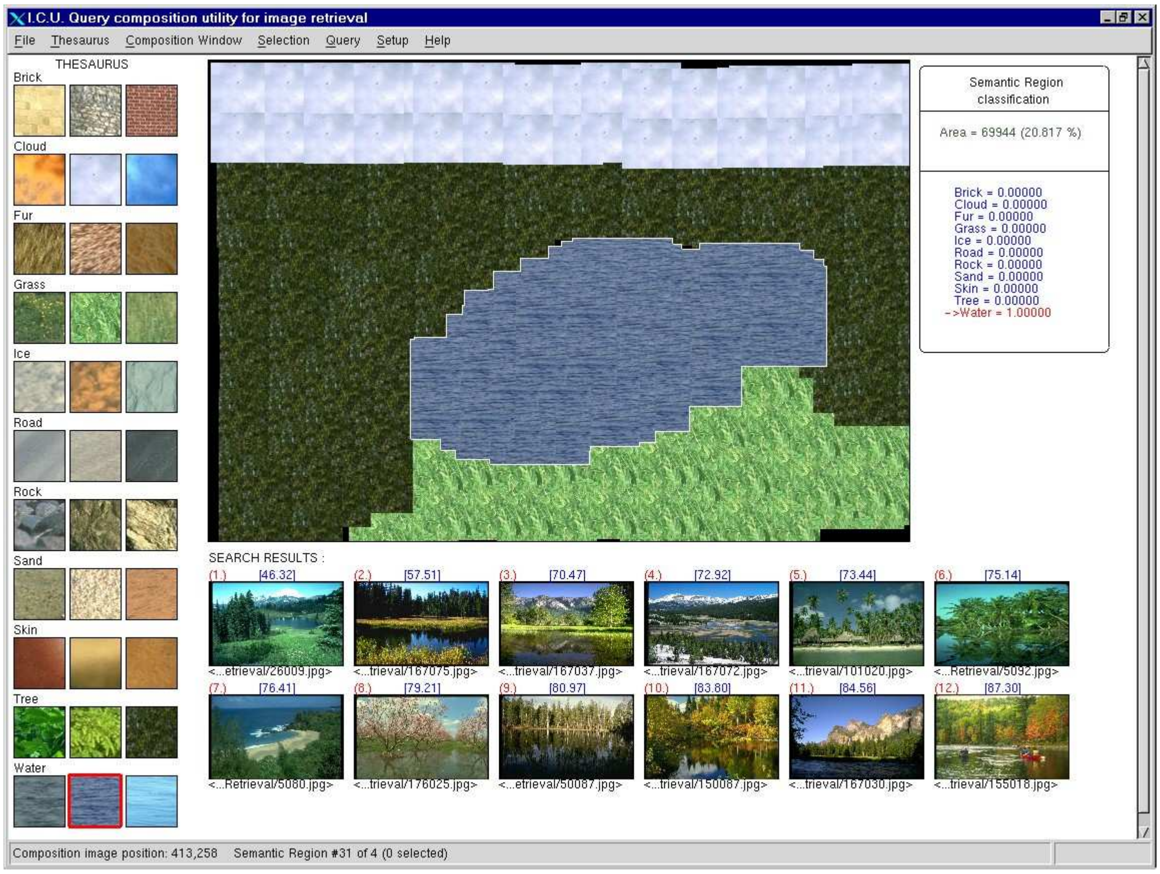

Figure 5: *Retrieval using a query composed by means of the thesaurus.*

both positive and negative retrieval targets by formulating queries using two principle operations:

Sample image: Any pre-classified image may be loaded as an initial search target (scaled according to the dimensions of the composition window).

Visual thesaurus: This consists of bitmap templates that have been individually assigned a certain classification, usually corresponding to membership of precisely one of the visual categories. A query can then be composed by drawing with a virtual paintbrush which copies the content of the selected thesaurus bitmap onto the composition window and simultaneously updates the underlying query matrix to create or modify semantic regions with the corresponding target classification. There is also a predefined default "don't care" category, which allows one to leave parts of the composition window unspecified.

The above methods can be arbitrarily combined, which is of particular use for query refinement. A user might have a vague recollection of an image he or she wishes to retrieve and can use the thesaurus to draw a very quick sketch. If the initial query is unsuccessful in retrieving the desired image, there is still a very high probability that the top retrieval results are somehow related to the desired outcome. The user could then use one of these to override parts of the original query, e.g. to provide target content at finer granularity or take account of the particular circumstance in which the image was taken, such as light conditions or elements of the surrounding scenery. It might also be useful to vary the similarity metric that is being used, for example by focussing on a particular part of the query and concentrating less on overall composition.

Sketch-based query composition tools have often been regarded as too cumbersome for real-world retrieval systems. However, the system developed here has shown that crude queries (figure 5) which can be constructed in a manner of seconds by ordinary users without requiring artistic skills or knowledge of the underlying system yield highly satisfactory results.

4.2 Similarity metrics

Relevance is assessed by means of a similarity metric, computed as the distance between a user specified selection of features in the query and corresponding features in each image of the search corpus. A small distance indicates a high degree of relevance. The most highly ranked images will be those which minimise the sum of distances for positive target queries and maximise the total distance for the negative queries. Targets can be weighted, the default being +1 for a positive query and -1 for a negative one.

Images are retrieved, ranked in order of decreasing relevance and the top returns are displayed in groups of 12. Relevance feedback consists of allowing the user to add, remove, or re-weight elements of the original query in light of the search result, e.g. by modifying target content based on desirable results or adding erroneous results as negative samples and restarting the retrieval process.

There are two principal ways of calculating similarity between an image and a query:

Segment classification search: the image and query are split into a grid of evenly spaced segments, and similarity is computed as either the sum over all grids of the Euclidean distance between *classification vectors*, or their cosine of correlation. By varying the number of grid rectangles (e.g. 1x1, 3x3, 5x5, and so on), one can take varying account of the spatial localisation of features.

Best region match: the user can select a set of semantic regions in the composition window and retrieve images based on the best match between the query regions and either a comparable subset or the entirety of regions in each corpus image. Regions may be selected manually by clicking on them or based on some criteria such as the largest regions covering a minimum or maximum percentage of image area, all regions of a certain size or classification, all regions in a certain part of the image, etc.. Similarity is once again based on comparing classification vectors, but other factors such as descriptors for the location, size, or shape of the regions in question can also be incorporated into the metric computation.

It is also possible to ignore some of the visual categories when calculating similarity or perform pre-processing on the classification vectors, namely by taking the L_∞ norm ("winner takes all") or standardising values to binary according to a classification threshold (0.6 is the default for positive classification).

Tests have shown that it is generally better to make use of the full classification of 11 categories. It has however proven effective to automatically assign confidence weights to region classification vectors. This uses prior knowledge of the corresponding visual categories (e.g. the conceptual similarity between "grass" and "tree") and simple numerical measures such as the number of peaks and troughs represented by the elements of the vector. For example, a vector which assigns high values only to a particular pair of related categories is likely to represent a greater certainty of classification than one which contains uniformly high or ambiguous probabilities for unrelated categories, and hence the former should be given a larger weight to reflect its greater usefulness for discrimination.

4.3 Work in progress

The semantic classification is currently being used to develop a text based query interface which maps search terms (e.g. "grass") onto visual categories and allows the corresponding target classification to be specified at greater precision by introducing modifiers for location (such as "upper left") or relative spatial proximity (e.g. "next to"). These terms can then be combined using boolean operators (and, or, xor, not).

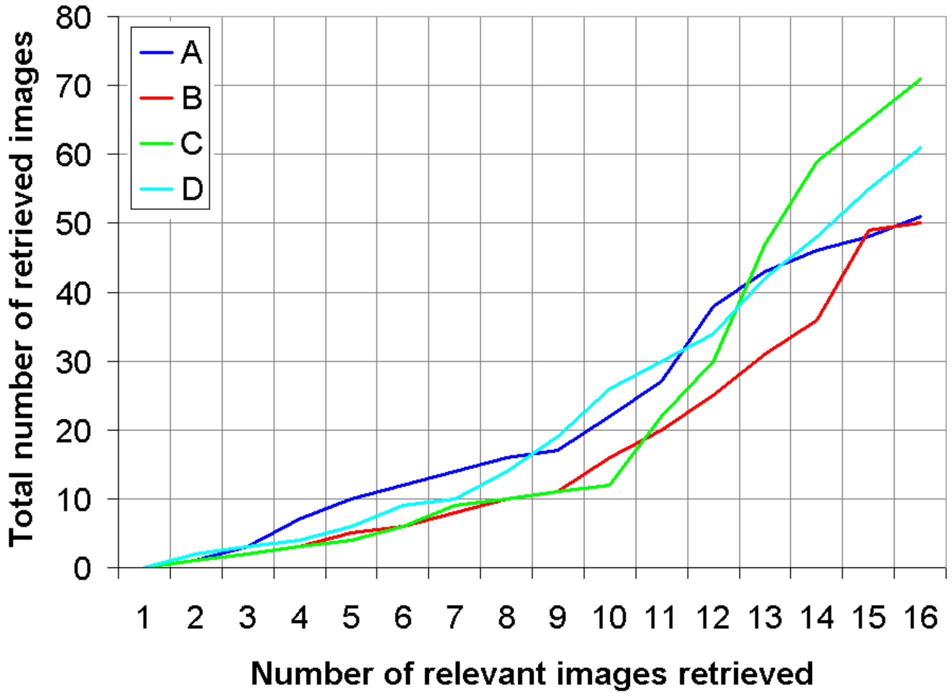

Figure 6: *Data from 4 retrieval experiments plotted as graphs showing the total number of images that were retrieved to find a given number of relevant images (the goal was to find the top 15 images deemed relevant for each query).*

5 Retrieval system evaluation

Evaluation was carried out using a highly heterogeneous collection consisting of over 1000 Corel Photo Library images and approximately 500 amateur home pictures.

Classification and retrieval results for the digital camera photographs were on the whole worse than those for the Corel images. This is to be expected, as the former often contain indoor scenes and materials outside the scope of the 11 visual categories, and may not be of the same professional quality as the latter. Moreover, the original training data was defined using Corel images.

However, an extended set of classifiers which includes the categories "cloth", "internal walls", "tarmac, and "wood" has yielded substantially improved results for general amateur imagery. The retrieval experiments discussed below were conducted using the original set of classifiers.

5.1 Comparison with related systems

Direct comparison is made difficult by the fact that many of these utilise very different segmentation and query schemes, as well as different (and sometimes proprietary) image databases.

Blobworld [3] is a content-based image retrieval system developed at UC Berkeley. It uses image segmentation to identify relevant parts of the image. A probabilistic image model is used to perform retrieval based on the Expectation-Maximization algorithm.

Photobook [8] was developed at the MIT Media Lab. Retrieval can be based on various features, such as appearance, shape, and texture. The underlying principle is referred to as "semantics-preserving image compression" and makes use of eigen-decomposition together with a reduction in feature dimensionality to derive image descriptors.
Both systems rely on segmentation and region properties to derive feature vectors of reduced dimensionality and allow content based retrieval using a variety of mechanisms. Compared to the schemes developed in this project, the above systems are more generally applicable to a wider range of images. However, both a subjective comparison of retrieval performances and assessment of precision/recall statistics show that the system developed here does in fact yield similar results for professional quality images of outdoor scenes. Furthermore, the explicit use of visual categories with obvious meaning allows more intuitive interface and retrieval methods than abstract concepts such as colour and texture properties.

5.2 Query composition and retrieval experiments

Retrieval success can be explicitly quantified by means of *precision* (fraction of retrieved images which are relevant) and *recall* (fraction of relevant images that have been retrieved).

A number of retrieval experiments were conducted on the combined corpus of 1538 images. In each case a full manual relevance assessment of all the images in the collection was carried out and the top n retrieved images ($n = 0.5R$, where R is the total number of images that were deemed relevant to the query) were assessed to give relative values for precision versus recall, as plotted in figure 6.

The results achieved are quite good compared to similar content based retrieval systems. Figure 5 (graph D) also illustrates how the thesaurus may serve as a useful query composition tool.

6 Summary and conclusions

This paper demonstrates how artificial neural network classifiers can be trained to map segmented image region properties onto a semantic category membership vector. The construction of such classifiers relies on sound statistical principles to provide a sufficient body of sample data and build neural networks which are capable of performing feature extraction from the high-dimensional input space while maintaining their ability to generalise from the learning set.

The semantic labelling of image regions represents a very useful high-level descriptor which is sufficiently consistent that regions of similar content result in similar classification vectors even if the classification itself is imperfect.

The main benefit in choosing a pre-defined set of visual categories such as trees and water lies in the flexibility and intuitive nature of query composition interfaces and retrieval systems which can be developed based on it.

Acknowledgements

The authors would like to acknowledge directional guidance and encouragement from Prof Andy Hopper and Dr Ken Wood.

References

[1] C.M. Bishop. *Neural Networks for Pattern Recognition.* Oxford University Press, 1995.

[2] N. W. Campbell, W. P. J. Mackeown, B. T Thomas, and T. Troscianko. Interpreting image databases by region classification. *Pattern Recognition (Special Edition on Image Databases)*, 30(4):555–563, April 1997.

[3] C. Carson, M. Thomas, S. Belongie, J.M. Hellerstein, and J. Malik. Blobworld: A system for region-based image indexing and retrieval. In *Third Int. Conference on Visual Information Systems*, 1999.

[4] D. Forsyth, J. Malik, M. Fleck, and J. Ponce. Primitives, perceptual organisation and object recognition. Technical Report http://HTTP.CS.Berkeley.EDU/ daf/vr11.ps.Z, University of California at Berkeley, 1997.

[5] A. Jaimes and S. Chang. Model-based classification of visual informaton for content-based retrieval. In *SPIE Conference on Storage and Retrieval for Image and Video Databases*, 1999.

[6] Joo-Hwee Lim. Learnable visual keywords for image classification. In *Proceedings of the Fourth ACM International Conference on Digital Libraries*, 1999.

[7] W. Niblack. The qbic project: querying images by color, texture and shape. In *IBM Research Report RJ-9203*, 1993.

[8] A. Pentland, R. Picard, and S. Sclaroff. Photobook: Tools for content-based manipulation of image databases. In *SPIE Storage and Retrieval of Image and Video Databases II*, 1994.

[9] B. D. Ripley. Neural networks and related methods for classification. In *J. Roy. Statist. Soc*, pages 409–456, 1994.

[10] B.D. Ripley. *Pattern Recognition and Neural Networks.* Cambridge University Press, 1996.

[11] N. C. Rowe and B. Frew. *Automatic classification of objects in captioned descriptive photographs for retrieval*, chapter 4, pages 65–79. AAAI Press, 1997.

[12] D. Sinclair. Voronoi seeded colour image segmentation. Technical Report TR99-04, AT&T Laboratories Cambridge, 1999.

[13] D. Sinclair. Smooth region structure: folds, domes, bowls, ridges, valleys and slopes. In *Proc. Conf. Computer Vision and Pattern Recognition*, pages 389–394. IEEE Comput. Soc. Press, 2000.

[14] A. Zell, N. Mache, R. Hübner, M. Vogt, G. Mamier, M. Schmalzl, and T. Sommer. SNNS user manual, version 4.2. Technical report, Eberhard-Karls-Universität Tübingen, Wilhelm-Schickard-Institut für Informatik, Tübingen, Germany, 1998.