Semantic Based Image Retrieval: A Probabilistic Approach

Ben Bradshaw
Microsoft Research,
St. George House,
1 Guildhall St.,
Cambridge, England.
+44 1223 744808
bbradsha@microsoft.com

ABSTRACT

This paper describes an approach to image retrieval based on the underlying semantics of images. To extract these semantics a hierarchical, probabilistic approach is proposed. The labels that are extracted in this case are man-made, natural, inside and outside. The hierarchical framework combines class likelihood probability estimates across a number of levels to form a posterior estimate of the probability of class membership. Unlike previous work in this field, the proposed algorithm can determine probabilities at any point in the scene and only a small number of images are required to train the system. To illustrate the potential of such an approach a prototype image retrieval system has been developed, initial results from this system are given in this paper.

Keywords

Image retrieval, semantic image analysis, image statistics.

1. INTRODUCTION

Content-based Image Retrieval (CBIR) has become an active area of research in the last 5-10 years because of the increasingly large volumes of electronically stored information and the corresponding requirement for high performance systems to access and manipulate this information.

This paper describes an architecture that accurately generates localised semantic labels (in a probabilistic setting). Specifically, the problem of probabilistically labelling images, or parts of images, as being made up of man-made (e.g. buildings, roads, cars) or natural (e.g. clouds, forests, streams) objects is addressed, as is the problem of determining whether an image was taken inside or outside. The advantages of the presented approach are that it generates localised probabilities, only requires a few 100 images to train and is computationally efficient.

The first section of this paper reviews the current research into CBIR systems and the extraction of semantic content from images. Following this review, Sections 2-4 describe the proposed method for extracting semantic information. Section 5 quantitatively examines results from the method and Section 6 describes and evaluates a prototype image retrieval system based on the research in this paper.

1.1 Content-based Image Retrieval

The first CBIR systems (classed by the author as 'first generation'), indexed images based on low-level features such as colour and texture. Examples of such systems are the IBM Query By Image Content (QBIC) system (see Flickner *et al.*, [2]) the Virage system (see Gupta *et al.*, [4]) and the VisualSEEk system (see Smith *et al.*, [14]). Each of these systems allows the user to specify a query in a number of ways based on the low-level features extracted by the system. The user is also often allowed to specify how much weight to attach to each of these features. It is now recognised that this explicit knowledge of the *low-level* feature space does not help the user formulate a query. The user (be they naive *or* experienced) finds it hard to determine which of the low-level features are appropriate for a given query.

In recent years there has been a proliferation in CBIR systems (which the author classifies as 'second generation') which deliberately hide the low level features from the user. Instead of specifying texture and colour combinations the user supplies an example image and asks for similar ones (termed Query By Example, QBE). Although this alleviates the problem of knowing which low level features are important for a given query it immediately introduces another one. Namely, the user must already have a good example of what they want prior to initiating the query. Examples of such systems are the 'Texture of textures' system devised by Debonet and Viola, [1] and the MARS system developed at the University of Illinois (see Rui *et al.* [13]). Note that the idea of introducing relevance feedback mechanisms into image retrieval systems, was first demonstrated in the latter system.

It is the author's belief that the next (third) generation of image retrieval systems will address the limitations of the second generation systems by replacing the low-level image feature space with a higher-level semantic space. Query formulation can then be performed using these higher level se-

mantics, these being much more understandable to the user than the low level image features used previously.

Given the large amounts of research into image retrieval it is surprising to find out how little research has been undertaken into understanding how users organise photos or image collections. To the author's knowledge the only published works addressing this are Rodden [11] who concentrated on home users (and hence is most relevant to our research) and Jose *et al.* [5] who concentrated on design professionals' needs in a work environment. The former outlines the results of interviews with a number of people who have large collections of home photographs and their experiences in organising them, whilst the latter undertakes a comparative evaluation of two variants of an image retrieval system in a work environment. It is interesting to note that the user's interviewed by Rodden stated that one of the *least* useful searching mechanisms for photos would be by colour or texture.

1.2 Extracting semantic content

There are a number of papers that address the issue of extracting semantic content from images. One of the first was Gorkani and Picard, [3] who attempt to discriminate 'city' from 'landscape' scenes using a texture orientation approach that is based on a multi-scale steerable pyramid operating on 128×128 pixel blocks across an image. Yiu [20] uses an identical texture extraction approach but introduces colour information to try and classify indoor and outdoor scenes. The procedure used is based on a nearest neighbour approach combined with a support vector machine classifier.

Szummer and Picard [15] address the same problem but combine a number of feature extraction techniques the best of which is a combination of Ohta colour space histograms (see Ohta, [10]) and textural features based on a multi-resolution simultaneous autoregressive model (see Mao and Jain, [8]).

Two other papers are closer to the work presented here. The first by Torralba and Oliva [16] describes an algorithm that attempts to determine a set of real-valued 'semantic axes' in their chosen feature space. They recognise the importance of being able to assign real-values to each image in relation to each semantic label, rather than the more common binary classification approach, but do not extend these real-values to a probabilistic representation. The second paper, by Vailaya *et al.* [17] describes a system that performs a hierarchical categorisation of images using a Bayesian framework resulting in probabilistic labels for the images.

Note that all of the systems described above only output one binary or real value per image. The primary contributions of this paper are the localisation of the semantic labelling to a *small area* (if required) in an image (rather than the entire image), the low number of images required to train the system and the low computational complexity involved in performing classification. All these features are important when considering the problem of retrieval from a large set of images. The localisation ability of the labelling is illustrated in Figure 1 where the labels have been used to segment the image. Another important contribution is the representation of the localised labels by probabilities (a localised equivalent of the global probabilistic values obtained by the system described by Vailaya *et al.*). This gives a principled approach to combining results across a number of categories.

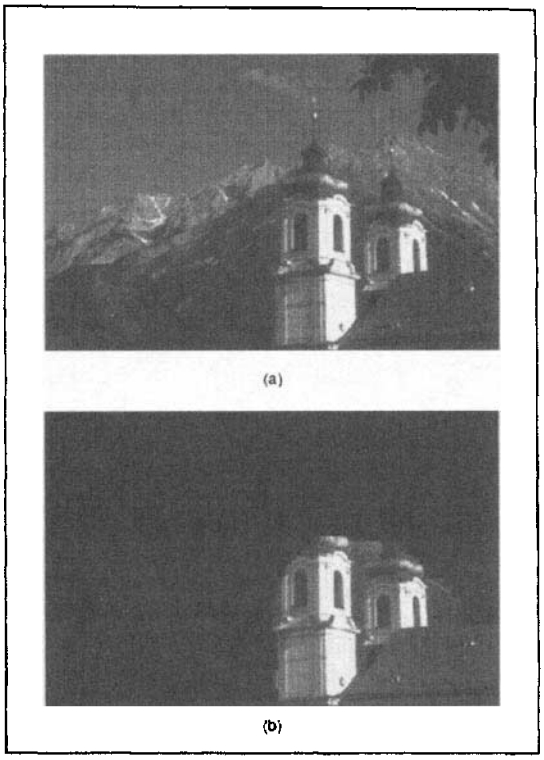

Figure 1: The proposed scheme can segment images based on the probabilistic labels. (a): Original image extracted from the Corel Gallery 1,000,000 collection. (b): Segmentation of the image highlighting those areas that are believed to be man-made. The threshold was set to only highlight those areas having probabilities greater than 0.5.

2. OVERVIEW

The technique outlined in this paper allows probabilistic, localised labelling of images. To achieve this the system aggregates data over a number of different block sizes and then combines the knowledge obtained from each of these levels. This gives rise to a significantly more robust algorithm than a single level approach. In this paper two sets of semantic labels are considered. The first set attempts to label small areas of the image as either natural or man-made (subsequently termed a 'local sampling procedure') whereas the second attempts to label the entire image as either having been taken from inside a building or outside (subsequently termed a 'global sampling procedure').

The local sampling procedure extracts data from different sized blocks from the image, each of which is centred at the current sampling point. Figure 2 illustrates this procedure. In using this approach, samples can be extracted at any position across the image, in the natural/man-made labelling the sample 'grid' was set to have a 16×16 pixel spac-

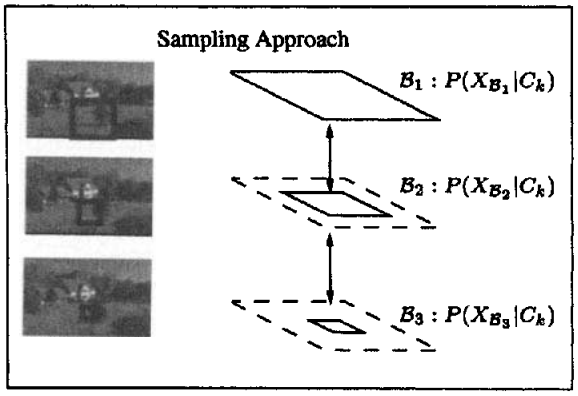

Figure 2: When classifying locally, blocks contributing to the same sample point are all centred at the same pixel position. At each level a feature vector from the corresponding block is extracted. After dimensionality reduction the likelihood of class membership can be estimated.

ing. Note that in so doing blocks from adjacent samples will overlap (thus applying an implicit smoothness constraint on the results of the labelling). Table 1 defines the levels used in both the local and global sampling procedures and their corresponding sizes. Note that all images used are either 256 × 384 pixels or 384 × 256 pixels in size, and have been extracted from the Corel Gallery 1,000,000 collection.

Table 1: Definition of levels and block sizes.

Level	Block Size
0	Whole image[1]
1	128 × 128
2	64 × 64
3	32 × 32
4	16 × 16

In order to obtain probabilistic labels probability densities representing the conditional class likelihoods must be estimated at each level. To achieve this a set of feature vectors is obtained for each class (in this case the natural/man-made categorisation is considered) at each level. This is done by selecting a set of natural and a set of man-made homogeneous images from which feature vectors can be extracted[2]. The feature vectors have 26 dimensions (these are described below); to make the task of estimating probability densities easier, dimensionality reduction is performed and then class likelihood densities estimated.

Once class conditional probability distributions have been determined labelling of an image can be performed. Feature vectors are extracted from all blocks at all levels. Class likelihoods are then estimated which, by the application of Bayes rule, allows posterior probabilities to be determined. In so doing, it is assumed that the class likelihoods at each level are statistically independent of each other. Note that in the two binary classification problems considered in this paper (natural/man-made and inside/outside) this results in probabilities (denoted as P(·)) such that:
P(Class 1) = 1 - P(Class 2).

One point that should also be mentioned is that the implicit smoothness constraint mentioned above implies that the proposed approach is unlikely to correctly label particularly small objects of either class (i.e. objects less than 16 × 16 pixels in size).

The following sections describe in detail how the feature vectors are extracted and the class likelihoods estimated.

3. EXTRACTING FEATURE VECTORS

3.1 Colour extraction

There are many different models to choose from when considering colour extraction. The primary aim when extracting colour is to obtain a set of values which are as decorrelated from each other as possible. Two common models used to achieve this are the Ohta colour model (see Ohta, [10]) and the HSV colour model (Hue, Saturation and Value). On comparing these methods, it was found that both the Ohta and HSV components had very similar correlation properties with both sets of components being significantly more decorrelated than those of the RGB model. Although they both give very similar results the Ohta transformation is more easily computed than the HSV transformation and so the former was used throughout the research described in this paper.

Subsequently, the three components of the Ohta colour transformation corresponding to image I are denoted as I_{o1}, I_{o2} and I_{o3}. The first of these corresponds to the luminance information, the latter two correspond to the chrominance information.

The chrominance information corresponding to block $\mathcal{B}_l$ (see Figure 2) is determined as follows:

$$C_1(\mathcal{B}_l) = \int_{\mathcal{B}_l} I_{o2}(\underline{r})\, d\underline{r} \quad (1)$$

$$C_2(\mathcal{B}_l) = \int_{\mathcal{B}_l} I_{o3}(\underline{r})\, d\underline{r} \quad (2)$$

where $\underline{r}$ denotes a particular pixel position in the image.

3.2 Texture extraction

The texture extraction approach is based on the complex wavelet transform (CWT). The CWT, developed by Kingsbury [6], is an efficient way of implementing a set of critically sampled Gabor-like wavelets. Gabor wavelets/filters have been used by a number of authors investigating both semantic content and classification problems, for example see Wood [19], Torralba [16], Rubner *et al.* [12] or Wiskott *et al.* [18]. They have also been shown to be similar in nature to the function of simple cells in the primary visual cortex of

[1]Note that level 0 is only used when classifying on a global scale such as in the Inside/Outside case and is not used when classifying locally as in the Natural/Manmade case.

[2]In this paper, the term 'homogeneous' refers to scenes that only contain one class of image data (i.e. in the natural/man-made case the image consists of completely natural or completely man-made objects) whereas 'inhomogeneous' refers to scenes containing both classes.

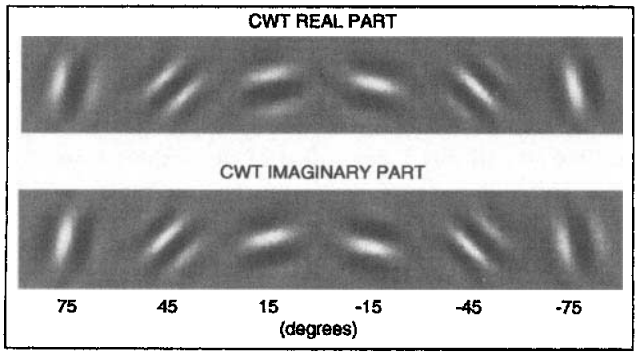

Figure 3: Illustration of the Gabor-like nature of the CWT which gives rise to directionally selective sets of filters.

primates (see Kruizinga [7] and Marčelja [9]). In this paper the CWT was used rather than Gabor wavelets because of the significantly reduced computational load[3].

Figure 3 illustrates the impulse responses obtained from the CWT at scale 4 in the decomposition. As described above these are Gabor-like in nature providing directional selectivity with 6 orientations at each scale. To succinctly represent the proposed procedure the following notation is introduced. The wavelet function at scale s, and orientation θ is denoted as ϕ_s^θ. The orientation can take one of six values $\Theta = \{15°, 45°, 75°, -75°, -45°, -15°\}$. The θ in the following text refers to an index into this vector i.e. $\theta \in \mathcal{I} : \{1 \ldots 6\}$. The response across the luminance image I_{o1}, extracted using the Ohta transformation, to each of the wavelet functions is determined as follows[4]:

$$I_s^\theta = I_{o1} * \phi_s^\theta \tag{3}$$

The $o1$ has been dropped to aid clarity. The energy response to the wavelet function at scale s, and orientation θ is defined for block $\mathcal{B}_l$ as:

$$\mathcal{T}_s^\theta(\mathcal{B}_l) = \int_{\mathcal{B}_l} (I_s^\theta(\underline{r}))^2 \, \mathrm{d}\underline{r} \tag{4}$$

In this paper, wavelet functions corresponding to scales 1-4 are used with 6 orientations at each scale giving rise to 24 texture based features per feature vector.

The feature vectors have 26 dimensions, 24 texture features and 2 colour features. Using the terms defined above, the feature vector at a particular block $\mathcal{B}_l$ is found by concatenating the texture based features with the colour based features as follows:

$$\mathbf{X}_{\mathcal{B}_l} = [\mathcal{T}_1^1(\mathcal{B}_l), \mathcal{T}_1^2(\mathcal{B}_l), \ldots, \mathcal{T}_4^6(\mathcal{B}_l), \mathcal{C}_1(\mathcal{B}_l), \mathcal{C}_2(\mathcal{B}_l)] \tag{5}$$

[3]Instead of requiring 2 dimensional convolutional operations the CWT combines results obtained from computationally efficient 1 dimensional convolutional operations.
[4]The $*$ symbol denotes the convolution operator

4. PROBABILISTIC LABELLING PROCEDURE

Prior to describing how the class likelihoods are estimated the following nomenclature is defined. The two possible classes are denoted as C_k where $k \in 1, 2$ (in the natural/man-made case C_1 corresponds to natural, C_2 corresponds to man-made). Probabilities are denoted as $P(\cdot)$ and probability densities as $p(\cdot)$. Fisher's discriminant approach is used to reduce the dimensionality of the feature space. This technique gives a vector at each level l, denoted subsequently as $\mathbf{p}_l$, on to which feature vectors from that level are projected. This discriminative approach gives the 'best' vector onto which to project feature vectors in the sense that the projection maximises inter-class separation whilst minimising intra-class distance. This results in a 1 dimensional space, i.e. the projections onto the vector gives scalar values. The value obtained from projecting the feature vector extracted from block $\mathcal{B}_l$ onto $\mathbf{p}_l$ is denoted as $X_{\mathcal{B}_l} = \mathbf{p}_l^{\mathrm{T}} \mathbf{X}_{\mathcal{B}_l}$.

Prior probabilities of class membership $P(C_k)$, likelihoods of class membership at a given block $P(X_{\mathcal{B}_l}|C_k)$ and a way of combining these likelihoods and priors is now required. Estimating the priors, $P(C_k)$, presents no problem; assuming that there is no knowledge about the images to be analysed they can be set at 0.5. However, to estimate the likelihoods at each position and level across the image, probability density estimation must be performed, this is discussed next.

4.1 Estimating the class likelihoods

The projection onto $\mathbf{p}_l$ at each level results in scalar values which makes probability density estimation very easy but throws away extra information that could possibly be used to discriminate between the two classes. This point is returned to in the summary.

Having determined the vector $\mathbf{p}_l$ for each level, probability densities for the class likelihoods can be estimated. The likelihoods are modelled using normal distributions this being based on the fact that when the dimensionality reduction step is undertaken it approximates to summing a set of independent, random variables and thus the central limit theorem can be invoked. The feature vectors for both classes are projected onto $\mathbf{p}_l$ and the mean and variance of the 1D normal distributions then found using the maximum likelihood approach. Figure 4 illustrates this class likelihood estimation process for the first three levels of the model in the natural/man-made case.

4.2 Combining class likelihoods

To estimate the probability of class membership, given data at a number of levels, it is assumed that the likelihoods at each level are statistically independent of each other; this assumption is addressed later in this section. In the following discussion the block index $\mathcal{B}$ is dropped to aid clarity. Given that data has been extracted from a number of block sizes, 1 to L, at a given sampling point the posterior conditioned on this data is:

$$P(C_k|X_1, \ldots, X_L) = \frac{P(X_1, \ldots, X_L|C_k)}{P(X_1, \ldots, X_L)} P(C_k) \tag{6}$$

$$\approx \frac{\prod_{l=1}^{L} P(X_l|C_k)}{P(X_1, \ldots, X_L)} P(C_k) \tag{7}$$

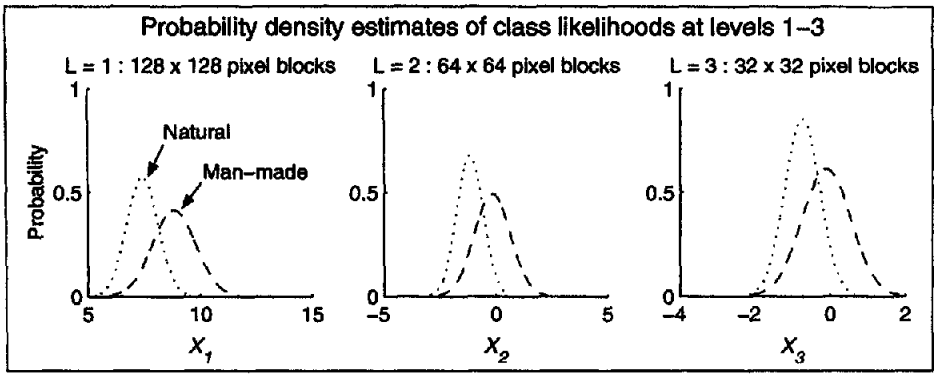

Figure 4: Estimates of the class likelihood probability densities. These correspond to the first three levels of the architecture obtained from the projected data X_1, X_2, X_3. Dotted: natural class likelihood $p(X_1|C_1)$. Dashed : Man-made class likelihood $p(X_1|C_2)$.

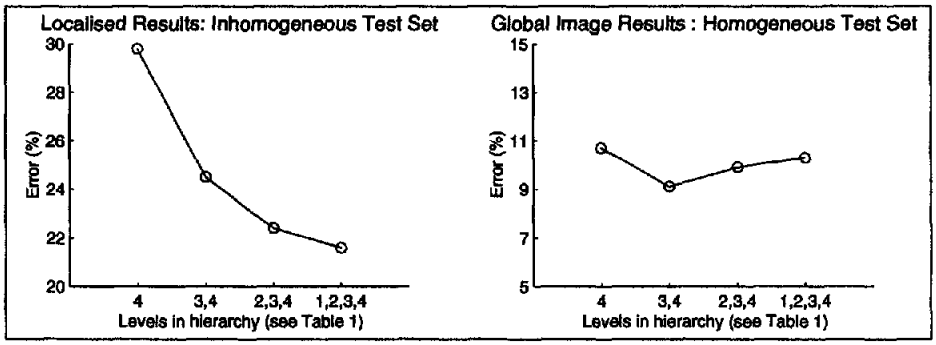

Figure 5: Natural/Man-made Results: Advantage gained from multi-level architectures. The left graph illustrates the classification error for all the samples (i.e. localised within each image). The right graph illustrates the error when the localised results for the homogeneous test set are averaged across each image to obtain a single classification result per image.

The denominator can be evaluated using the chain rule as follows:

$$P(X_1, \ldots, X_L) = P(X_L|X_{L-1}, \ldots, X_1) \ldots P(X_2|X_1)P(X_1) \quad (8)$$

Each factor in this expansion of the denominator can be evaluated by marginalising over the class variable and this then allows Equation 7 to be evaluated in a top down, recursive, manner.

As described above the likelihoods at each level are assumed to be statistically independent of each other. The reason that this assumption is made is to reduce the amount of data required to train the system. Thus, the system described in this paper can be trained using 100's of images rather than 1000's (as is required in other techniques).

5. RESULTS

5.1 Natural/Man-made

As described above, to estimate the class likelihoods a set of labelled data is required and so 120 'natural' and 120 'man-made' scenes were taken from the Corel Gallery 1000000 collection. Feature vectors for each of these sets of images were extracted for levels 1 to 4 (128 × 128 to 16 × 16 pixel block size)[5]. Following this, the optimum vector to project onto at each of the levels was determined using Fisher's linear discriminant and estimation of the probability densities then performed. The resulting densities for the first three levels are shown in Figure 4.

Two test sets of images were used to assess the system. The first set consists of 240 homogeneous images. In this set there were 120 images containing only natural objects and 120 images containing only man-made objects. The second inhomogeneous test set consisted of 125 images containing both natural and man-made objects. These were hand-segmented to allow analysis of how well the system performed at a local scale. The images for both test sets were taken from the Corel Gallery 1000000 collection. Note that none of the images in either of the test sets were in the labelled set of images used to generate the class likelihoods.

Experiments were undertaken to evaluate the performance of a variety of architectures. Classification was based on thresholding the posterior probabilities such that samples

[5]To ensure that the system correctly identified sky as 'natural', those areas of the man-made scenes containing significant amounts of sky were ignored.

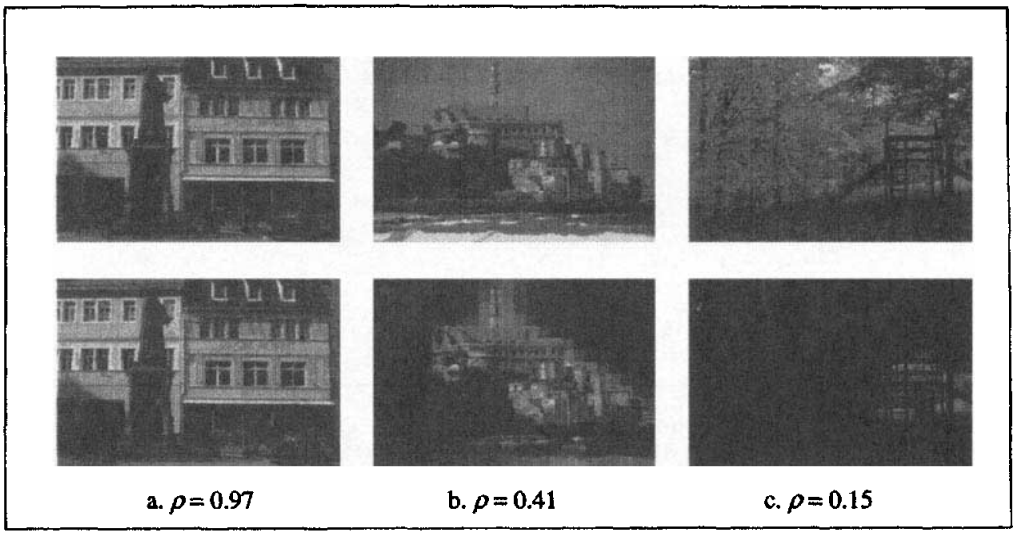

Figure 6: Examples of Natural/Man-made Results: Top row: Original images. Bottom row: The brightness of each image block has been weighted with the posterior probability of being man-made. ρ : proportion of samples classified as being man-made.

with posteriors > 0.5 were 'in-class' otherwise they were 'out-of-class'.

Figure 5 shows the classification error for the samples obtained from the inhomogeneous test set and also shows the error obtained for the homogeneous test set when the localised results were averaged across each image to give a single classification result per image. These graphs illustrate two points.

- For images containing both natural and man-made objects, adding extra levels significantly increases the performance of the algorithm (as illustrated by the left graph). Using only level 4, for example, results in a classification error for the inhomogeneous test set of 29.8% whereas combining levels 1,2,3 and 4 reduces this to 21.6%.
- By averaging the localised results (as illustrated in the right graph), the system can accurately classify homogeneous images. The error in this classification is approximately 10%.

Figure 6 shows the results from a number of images having used an architecture consisting of levels 1,2,3 and 4. These examples illustrate that the proportion of samples, ρ, corresponds well with the proportion of the scene containing man-made objects.

5.2 Inside/Outside

The problem of classifying images as having been taken inside or outside is different to that of the natural/man-made problem in that there is no requirement to obtain localised results (i.e. the classification is a global property of the image). To take account of this an extra level was added to the hierarchy containing information aggregated across the

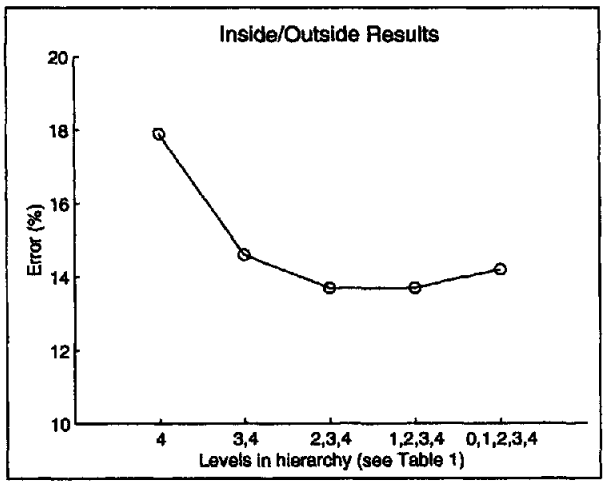

Figure 7: Inside/Outside Results: These results illustrate that the optimum performance is achieved when using at least 3 levels in the hierarchy.

entire image (see Table 1). 120 'inside' and 120 'outside' scenes were taken from the Corel Gallery 1000000 collection. Feature vectors for each of these sets of images were extracted for levels 0 to 4 (entire image through to 16 × 16 pixel block size). As with the natural/man-made case the optimum vector to project onto at each of the levels was determined using Fisher's linear discriminant and estimation of the probability densities then performed. Note that because inside/outside is a global property of the image samples were extracted using a 32 × 32 grid (rather than the 16 × 16 grid used in the natural/man-made case).

A test set of 240 images (none of which were in the training set) was used to determine the performance of the system.

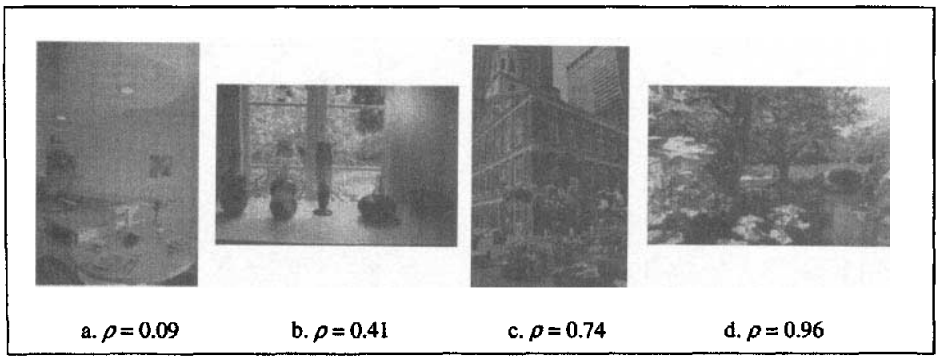

Figure 8: Examples of Inside/Outside Results. ρ : proportion of samples classified as being outside.

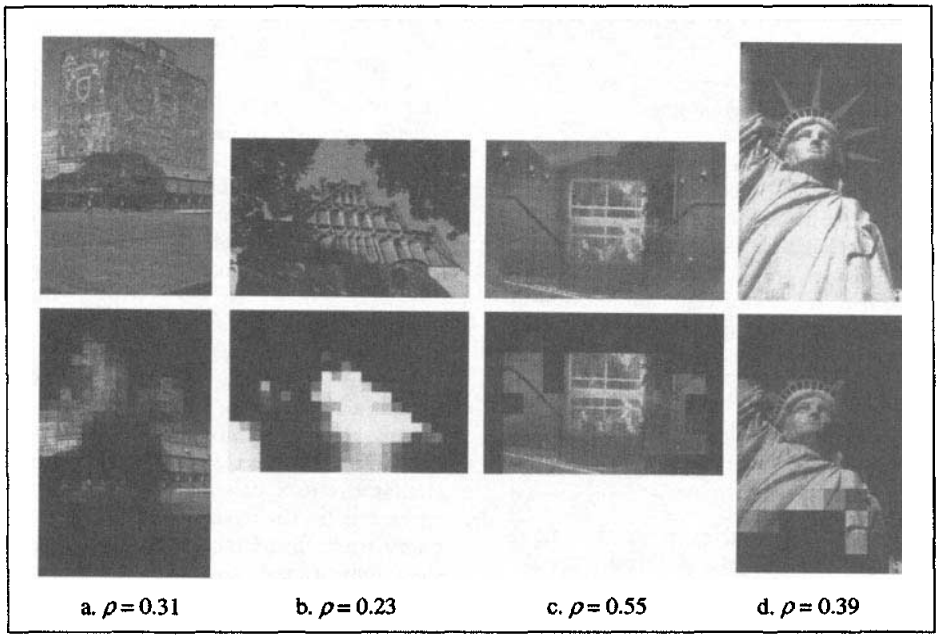

Figure 9: Four examples of errors in classification. Top row: Original images. Bottom row: Original images weighted by the relevant probabilities. (a),(b): Man-made classification; (c),(d): Outside classification. Note that in (b) the original image has been omitted to aid clarity.

Figure 7 shows the results illustrating that the optimum architecture consists of at least 3 levels giving rise to a classification accuracy of 86.3%. As with the natural/man-made case, using multiple levels gives better results although interestingly using the whole image (i.e. Level 0) degrades the performance slightly. This indicates that using features aggregated across the entire image can reduce the accuracy of an algorithm even in the cases where *global* attributes are trying to be extracted. Figure 8 shows a number of examples (having used an architecture consisting of levels 2,3, and 4).

5.3 Classification Errors

Figure 9 shows examples of where the classification process has gone wrong. In 9(a) the colours painted onto the building lead to about half of the building being incorrectly labelled as natural, thus biasing the overall result. In 9(b), the shadows cast onto the building from the surrounding trees are also incorrectly classified as natural (in this case the image has been omitted for clarity). In 9(c), the system incorrectly labels the swimming pool and the window as corresponding to objects that occur in 'outside' scenes and thus the overall classification is biased towards an 'outside' result. 9(d) illustrates that the system labels dark, low texture areas as corresponding to images that have been taken inside which in this case results in an incorrect classification.

6. IMAGE RETRIEVAL APPLICATION

To assess whether a system based on the algorithms outlined above can actually work in practice, a prototype image retrieval application was built using the author's digital photo archive as a database. A useful feature associated with digital cameras is that it is often possible to extract the date that each photo was taken, this extra information was used in the retrieval procedure as explained below. The archive consists of 800 images taken over the past 12 months. A daytime/night-time binary classification was added using a very similar procedure to that of the inside/outside to give extra information useful for retrieval. The resulting indexing

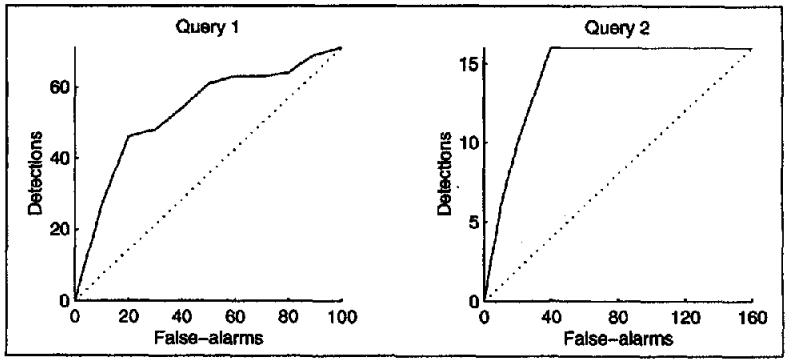

Figure 10: Receiver-Operator Characteristics: Query 1 : 'Find pictures of my holiday in Wales'. Images ordered in terms of P(Outside)×P(Natural). Query 2 : 'Find pictures of the evening spent with friends in a bar in Boston'. Images ordered in terms of P(Inside). The dotted lines indicate the results that would be obtained by random selection.

procedure for each image consists of two steps.

1. Rescaling the image. This is done firstly to reduce the computational complexity and secondly to ensure that the image data extracted from the photos corresponds to approximately the same area as that obtained from the images from the Corel collection. Thus, the system rescales each image to ensure that the maximum dimension is no greater than 400 pixels.

2. Processing the image and subsequently storing the extracted probabilities and also the date that the photo was taken.

Unlike the majority of image retrieval systems that have to store the actual feature vectors in the database, the resulting 'signature' per image in this case is extremely small consisting of a date and three probability values.

During retrieval the user is asked to specify the dates between which the image was likely to have been taken. This significantly reduces the number of photos that need to be searched. The user can then indicate whether the photo was taken inside or outside. If it was taken outside, the user can also indicate whether the image consists of mainly natural or man-made objects and/or whether the image was taken during the daytime or at night.

To give an indication of how well the system performs, the results of two queries presented to this prototype system are shown in Figures 11 and 12. In both cases the same sets of dates were chosen during which the author went on holiday to both Boston and Wales; this reduces the search from the entire database (800 images) to 171 images. The first query is based on trying to find pictures from the holiday in Wales and, knowing that the images contain a lot of scenery, the system is asked to return images in decreasing order of P(Outside'ness and 'Natural'ness. Out of the 171 images, 71 match this query (i.e. were taken whilst in Wales). The Receiver Operator Characteristic (a plot of the false-alarms vs detections) is shown in Figure 10 and the top 24 images, as returned by the system, are shown in Figure 11. Of the first 24 images, 17 are relevant to the query.

The second query is based on extracting images taken on a specific occasion, namely an evening spent with friends in a bar in Boston. In this case the images are ordered in terms P(Inside). Only 16 of the 171 images match this query, again the Receiver Operator Characteristic is shown in Figure 10 and the top 24 images are shown in Figure 12. Of the first 24 images, 9 are relevant to this query.

The prototype application presented above illustrates the potential of a semantic-based image retrieval application. The results show that combining date metadata with even this small number of categories gives a system that allows the user to find different types of images quickly and easily. The assumption on which this semantic retrieval is based is that the user can map from the query to the semantics extracted by the system (in query 1, illustrated above, the query was: 'Find pictures from the holiday in Wales' and the semantics used to retrieve images were 'outside-ness' and 'natural-ness'). Note that this mapping from query to semantic features is much more intuitive than the mapping from query to low-level image features.

7. SUMMARY

In this paper a probabilistic, multiple level approach to the semantic labelling of images has been proposed. The architecture is based on modelling class likelihoods at each of the levels separately and combining these to form an overall estimate of the posterior, conditioned on the data. The results that have been presented illustrate that using multiple levels significantly increases the accuracy of the posterior probabilities.

The binary semantic categories that have been investigated are natural/man-made and inside/outside. The results illustrate that the proposed technique can classify images with an accuracy of between 86-91% given a training set of only a couple of hundred images. The method outlined in this paper compares well with other, previously published, techniques such as that proposed by Vailaya *et al.*. Note that it is difficult to make quantitative comparisons between different retrieval systems because of the lack of consistent test sets of images. It is also acknowledged that the numbers of images used in the analyses presented in this paper are small. To this end, a much larger test set of images is currently being

gathered and hand-labelled to allow more exhaustive testing of the algorithms described in this paper.

A prototype photo archive image retrieval system based on the probabilistic labelling procedure has been tested using the author's digital image archive. Using the date information combined with the semantic labels gave surprisingly good results and indicate that this is a very useful approach to image retrieval. It is clear that more needs to be learnt about how users would like to interact with such systems.

The primary limitations of the proposed system are a. it can only discriminate between classes that are linearly separable and b. the assumption of statistical independence between levels in the classification algorithm. Noting these limitations, there are a number of interesting areas of research to persue following from the work presented in this paper:

- Using more sophisticated techniques for the probability density estimation task such as kernel methods or Bayesian Belief Networks etc. These techniques are likely to improve upon the results presented here (and address the limitations of the current algorithm as described above).
- Finding other categories that are of use in this domain. An obvious category that is required in the photo archive domain is 'People'/'No people'. This leads onto more specific categories such as trying to locate particular types of object.
- Incorporating relevance feedback and Query By Example techniques into the system.

8. ACKNOWLEDGEMENTS

All the images used in this paper were extracted from the Corel Gallery 1,000,000 collection. Special thanks to Andrew Blake for helpful comments regarding this paper and to Nick Kingsbury for comments regarding the description of the CWT and for the 'donation' of Figure 3.

9. REFERENCES

[1] J. S. DeBonet. Novel statistical multiresolution techniques for image synthesis, discrimnation and recognition. Master's thesis, M.I.T. Learning and Vision Group, AI Lab,, 1997.

[2] M. Flickner, H. Sawnhey, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker. *Intelligent Multimedia Information Retrieval*, chapter Query by image and video content: The QBIC system, pages 8–22. AAAI Press, 1997.

[3] M. M. Gorkani and R. W. Picard. Texture orientation for sorting photos 'at a glance'. In *Proc. of the IEEE Int. Conf. on Pattern Recognition*, October 1994.

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

[5] J. M. Jose and J. Furner. Spatial querying for image retrieval: a user oriented evaluation. In *21st SIGIR Int. Conf. on Research and Development in Information Retrieval*. ACM, August 1998.

[6] N. G. Kingsbury. The dual-tree complex wavelet transform: A new efficient tool for image restoration and enhancement. In *EUSIPCO'98*, volume 1, pages 319–322. EURASIP, 1998.

[7] P. Kruizinga and N. Petkov. Nonlinear operator for oriented texture. *IEEE Trans. on Image Processing*, 8(10):1395–1407, October 1999.

[8] J. Mao and A. Jain. Texture classification and segmentation using multiresolution simultaneous autoregressive models. *Pat. Rec.*, 25(2):173–188, 1992.

[9] S. Marčelja. Mathematical description of the response of simple cortical cells. *Journal of the Optical Society of America*, 70:1297–1300, 1980.

[10] Y. Ohta, T. Kanade, and T. Sakai. Colour information for region segmentation. *Computer Graphics and Image Processing*, 13:222–241, 1980.

[11] K. Rodden. How do people organise their photographs? In *IRSG 99, 21st Colloquim on Information Retrieval*, pages 142–152. British Computer Society, April 1999.

[12] Y. Rubner and C. Tomasi. Texture-based image retrieval without segmentation. In *ICCV'99*, Corfu, Greece, September 1999.

[13] Y. Rui, T. Huang, S. Mehrotra, and M. Ortega. A relevance feedback architecture in content-based multimedia information retrieval systems. In *Proc of IEEE Workshop on Content-based Access of Image and Video Libraries*, 1997. in conjunction with CVPR'97.

[14] J. Smith and S. Chang. *Intelligent Multimedia Information Retrieval*, chapter Querying by colour regions using the VisualSEEk content-based visual query system, pages 23–41. AAAI Press, 1997.

[15] M. Szummer and R. W. Picard. Indoor-outdoor image classification. In *IEEE Int. Work. on Content-based Access of Image and Vid. Databases*, January 1998.

[16] A. B. Torralba and A. Oliva. Semantic organisation of scenes using discriminant structural templates. In *ICCV'99*, Corfu, Greece, September 1999.

[17] A. Vailaya, M. Figueiredo, A. K. Jain, and H.J. Zhang. Content-based hierarchical classification of vacation images. In *IEEE Conf. on Multimedia Computing and Systems*, volume 1, pages 518–523, 1999.

[18] L. Wiskott, J. Fellous, N. Krüger, and C. von der Malsburg. Face recognition by elastic bunch graph matching. In *IEEE Int. Conf. on Image Processing*, volume 1, 1997.

[19] M. E. J. Wood, N. W. Campbell, and B. T. Thomas. Iterative refinement by relevance feedback in content-based digital image retrieval. In *6th ACM Int. Multimedia Conference*, Bristol, September 1998. http://www.acm.org/sigmm/MM98/.

[20] E. C. Yiu. Image classification using colour cues and texture orientation. Master's thesis, Dept EECS, MIT, 1996.

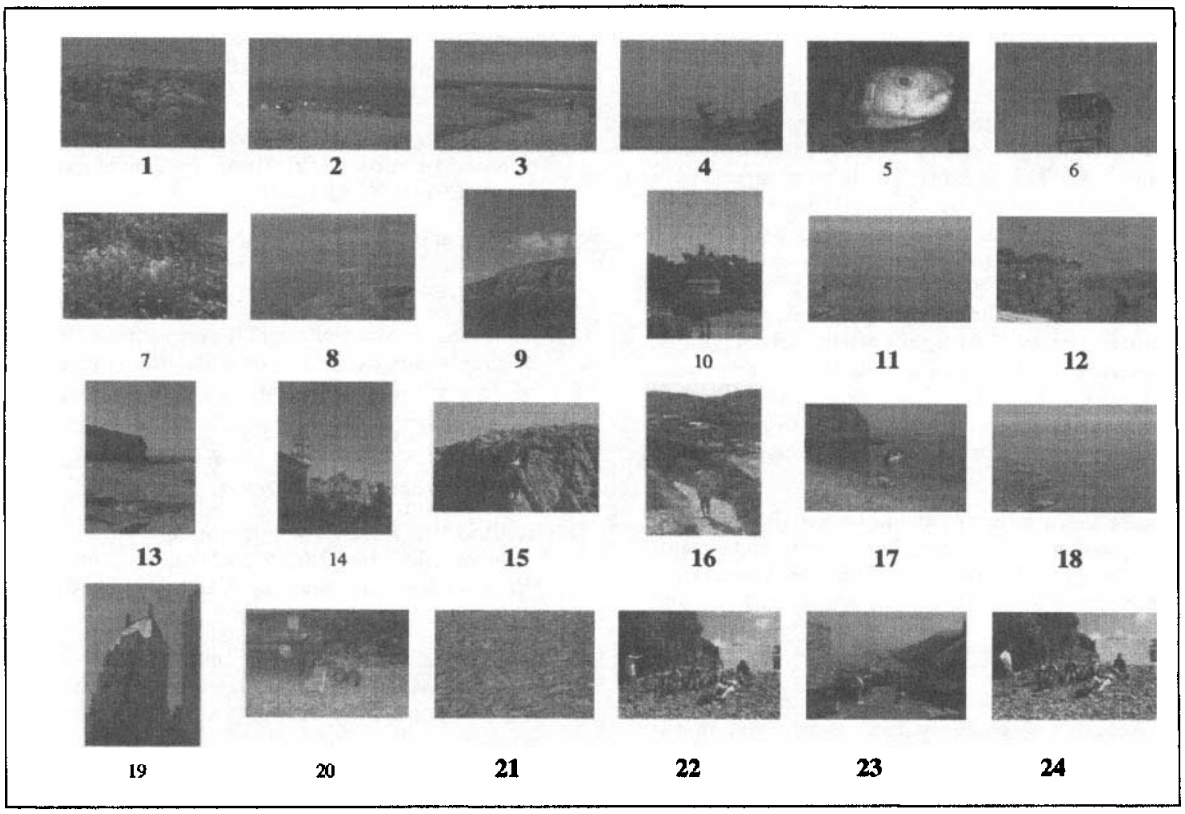

Figure 11: Image retrieval application: Query 1 : 'Find pictures of my holiday in Wales'. Selecting dates between May 1st and June 15th cuts down the search to 171 images which are then ordered in terms of P(Outside)×P(Natural). Out of the images shown here, 17 are relevant to the query. These are indicated by the larger, bold typeface numbers.

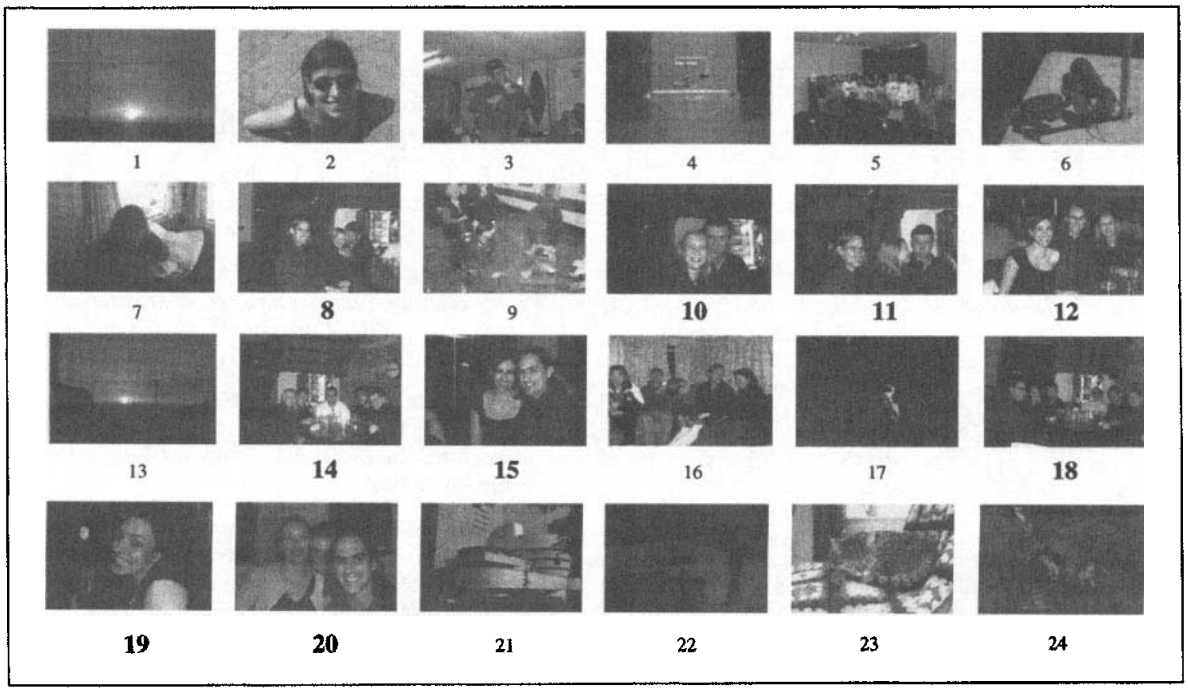

Figure 12: Image retrieval application: Query 2 : 'Find images of the evening spent with friends in a bar in Boston'. Selecting dates between May 1st and June 15th cuts down the search to 171 images which are then ordered in terms of P(Inside). 9 of the images shown here are relevant to the query. These are indicated by the larger, bold typeface numbers.