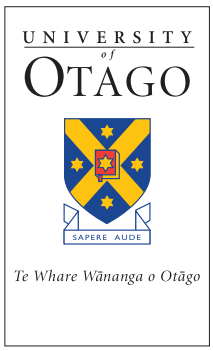


Combining Multiple Precision-Boosted Classifiers for Indoor-Outdoor Scene Classification

Da Deng
Jianhua Zhang

The Information Science Discussion Paper Series

Number 2006/09
May 2006
ISSN 1177-455X

University of Otago

Department of Information Science

The Department of Information Science is one of seven departments that make up the School of Business at the University of Otago. The department offers courses of study leading to a major in Information Science within the BCom, BA and BSc degrees. In addition to undergraduate teaching, the department is also strongly involved in postgraduate research programmes leading to MCom, MA, MSc and PhD degrees. Research projects in spatial information processing, connectionist-based information systems, software engineering and software development, information engineering and database, software metrics, distributed information systems, multimedia information systems and information systems security are particularly well supported.

The views expressed in this paper are not necessarily those of the department as a whole. The accuracy of the information presented in this paper is the sole responsibility of the authors.

Correspondence

This paper represents work to date and may not necessarily form the basis for the authors' final conclusions relating to this topic. It is likely, however, that the paper will appear in some form in a journal or in conference proceedings in the near future. The authors would be pleased to receive correspondence in connection with any of the issues raised in this paper, or for subsequent publication details. Please write directly to the authors at the address provided below. (Details of final journal/conference publication venues for these papers are also provided on the Department's publications web pages: http://www.otago.ac.nz/informationscience/pubs/). Any other correspondence concerning the Series should be sent to the DPS Coordinator.

Department of Information Science
University of Otago
P O Box 56
Dunedin
NEW ZEALAND

Fax: +64 3 479 8311
email: dps@infoscience.otago.ac.nz
www: http://www.otago.ac.nz/informationscience/

Combining Multiple Precision-Boosted Classifiers for Indoor-Outdoor Scene Classification

Da Deng and Jianhua Zhang

Department of Information Science, University of Otago,
P.O. Box 56, Dunedin, New Zealand

Abstract

Along with the progress of the content-based image retrieval research and the development of the MPEG-7 XM feature descriptors, there has been an increasing research interest on object recognition and semantics extraction from images and videos. In this paper, we revisit an old problem of indoor versus outdoor scene classification. By introducing a precision-boosted combination scheme of multiple classifiers trained on several global and regional feature descriptors, our experiment has led to better results compared with conventional approaches.

Key words: scene classification, classifier combination

1 Introduction

With the rapid development of information and communication technologies, the ever-growing use of multimedia data has brought many technical challenges, especially for data compression and information retrieval to be done effectively, efficiently and flexibly. In recent years, an intensive research effort has been focused on content-based image retrieval (CBIR) [1]. CBIR was proposed to overcome the shortcomings of traditional annotation-based retrieval system for images and videos. It aims at effective multimedia asset management and efficient information retrieval by automatically indexing image or video storage based on their low-level visual features such as colour, texture, shape and regions. Being under rigorous research worldwide for more than a decade, CBIR has made significant contribution to the research and development of multimedia systems.

[1] The author thanks the support of Grant UOOX0208 from the FRST, New Zealand.

However, CBIR's reliance on low-level features alone can also result in poor retrieval quality and lack of semantic representation of image indexes. Although techniques such as joint feature histograms and relevance feedback etc. have been investigated to more or less improve the retrieval quality, it has been realised that the bottleneck remains at the semantic gap [1] between the simplicity of visual features available and the richness of hidden semantics. An intelligent multimedia asset management system should require the capability of automatic interpretation of visual features so that semantic objects or concepts can be extracted and used for indexing and retrieval purposes. There are many research works that try to explore the relationship between low-level features or their combinations and the corresponding semantic concepts. For instance, in [2], colour semantics of art works are used for image retrieval based on perceptual concepts on colour quantities and sensation, such as warmth, harmony and anguish.

Although being still far away to offer any mature solutions in bridging the semantic gap, CBIR has activated the research on image analysis and laid down a sound basis of low level visual feature extraction schemes that are representative and powerful for image indexing and retrieval. For instance, the MPEG-7 core experiments [3][4] has proposed a rich set of low level features whose robustness has been tested with a huge amount of image data. It is therefore desirable to use machine learning and pattern recognition techniques to model these low-level features and semantic concepts. Because of the difficulty in achieving image understanding in general, this is usually limited within specific domains. In [5], an approach was proposed to learn abstract concepts (indoor versus outdoor) from low-level features based on classification combination. Each training image was manually assigned a semantic label, and then divided into fixed-size sub-blocks. Several kinds of visual features on colour, texture and frequency were extracted from the sub-blocks each to train a classifier separately, and then the classified results of all classifiers were combined based on a majority voting rule to determine the high-level semantic properties of a testing image. It was demonstrated that combining multiple weak features with a *k*-nearest neighbour (*k*-NN) classifier can produce better results than using a single 'good' feature. However, due to the rigid partition of the image into fixed-size blocks, there is no control for a block to possibly correspond to any meaningful object. Rather a block may contain several objects or parts of different objects, causing the feature set extracted from blocks unable to retain image semantics reliably, and hence the accuracy of scene classification is affected. In [6] some low level global features from the MPEG-7 core experiments suggested in [3] were extracted to train different classifiers such as *k*-NN and support vector machines for different semantic categories, and then results of the individual classifiers were combined into a final classification based on several strategies. In [7], an image is also divided into rigidly split blocks (4×4), and low-level features in each block are used to train a 2-D multiple-resolution HMM models that produce a semantic concept dictionary. Because of the rigid block partition, these however may inherit the weakness as in citesz98.

In this paper, we proposed a new approach to tackle the indoor-outdoor scene classification problem given in [5]. Both global and local visual feature were extracted from the scene image. For local features we use segmented regions of homogeneity instead of fixed-size blocks generated by brute-force partitioning. Each type feature set trains an individual classifier. We then calculate the precision of the classifier corresponding to each semantic category by cross validation, and use it to tune the Bayesian posterior probability and assign the image to the class with the maximum probability. Our experiment demonstrates that this approach successfully improves the classification rate of indoor-outdoor scenes.

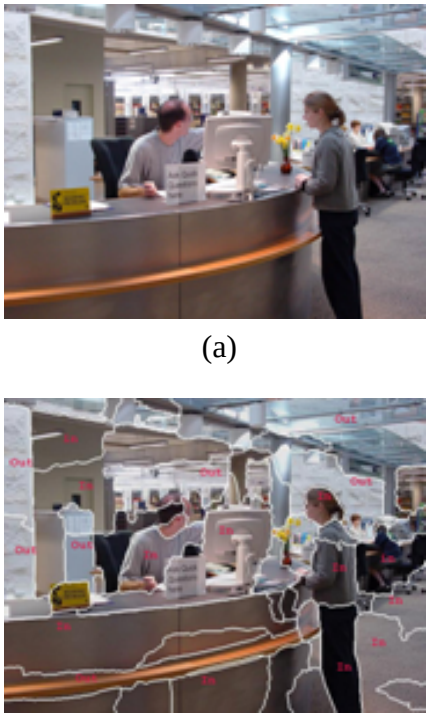
The rest of the paper is thus organised. In Section 2 we go through the feature extraction process and briefly introduce some feature schemes adopted for the scene classification problem. In Section 3 we present a new method to combine multiple classifiers using precision boosting. Empirical results are presented and discussed in Section 4. Finally we conclude with some discussion on the future work.

2 Feature Extraction

2.1 Global features

Human visual perceptions are very sensitive to colours. Colour histograms have been the most versatile features used in CBIR since they are representative and robust to resolution variation, translation and rotation. We adopt the LUV colour space in computing the colour histograms as it models well the human perception on colour similarity. For the global colour histogram, we quantize each channel of LUV into 5 bins and then compute the colour histogram as the global colour feature. The selection of the LUV space is also based on our finding that classification validation based on LUV is consistently better than that on the commonly used RGB space.

Luminance Edge histogram descriptor (EHD) [3], another feature descriptor widely used in CBIR research, captures the spatial distribution of five directional edges. It is found to be quite effective in representing natural image, therefore is adopted here. To obtain the EHD, edge filters are first applied to detect the edges of the 2×2 pixel blocks, such as vertical, horizontal, 45 diagonal, 135 diagonal edges, non-directed edges and no edge. Grouping the edge information of all blocks generates an edge histogram with 6 bins. An image is partitioned into 4×4 sub-images, each generating a sub-image edge histogram.Concatenating the edge histogram vectors results in a global edge histogram of 96 dimensions. We used the MPEG-7 core experiment code [8] for this purpose.



(a)

(b)

Fig. 1. An example scene image: (a) original, and (b) segmented.

2.2 Image segmentation

Apart from the global visual features, regional visual characteristics are also considered here. Because image segmentation can extract homogenous regions of better semantic integrity, one hope it will produce improved modelling capability on the image semantics compared with the approaches using fixed-size blocks or global histogram features only. The image is segmented into colour-texture homogeneous regions using the JSEG algorithm [9]. JSEG quantises the image into colour-maps, and then spatial segmentation and region-growing methods are used to merge similar regions to generate the final segmentation. One can tune the colour quantisation parameters and the merging threshold to control the segmentation outcome. We used a moderate merging threshold of 0.3 in our study. Figure 1 shows the segmentation result of an example 'indoor' scene with this setting.

Once the scene image is segmented, local colour features and texture features can be extracted from the segments and later used as training samples for different classifiers.

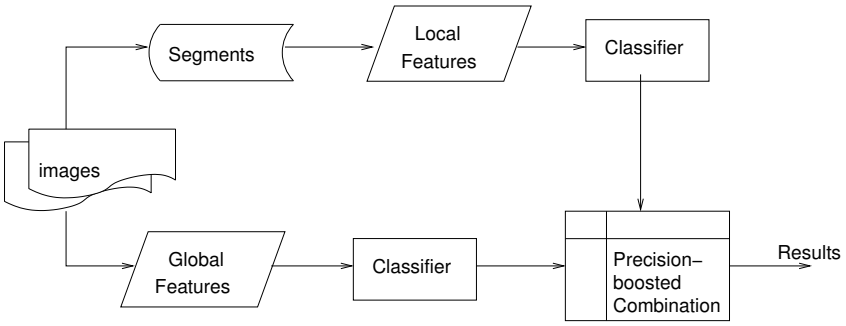


Fig. 2. The overall system diagram

2.3 Local features from image segments

For the colour histograms of segmented regions, as there exist fewer colours in each region, we can quantize colours in finer granularities. Unlike the uniform quantisation used in the global colour histogram, we quantise each channel with the same interval. Thus the L channel has 20 bins, U has 70 bins and V has 42 bins, all corresponding to the different ranges in each channel: L ($0 \sim 100$), U ($-134 \sim 220$) and V ($-140 \sim 122$). These are then concatenated into a total of 132 bins.

The computation of the local edge histogram over a segmented region can be tricky due to its arbitrary shape as shown in Figure 1. We compute the local edge histogram by grouping the edge histogram of the image blocks fallen into the segmented region and normalize it. The local edge histogram vector with 6 bins reflects the edge information since segmented regions have colour-texture homogeneity. Those regions that are too small to contain any image blocks are ignored as we assume that these small regions give no significant semantic indication.

3 Classification models and their combination

At this stage we have extracted four feature sets: two global and two regional. Each of these feature set will train an individual classifier. As revealed by many studies, in real world pattern classification problems it is usually hard to find and rely on a best classifier based on some good feature; a more reliable approach is to derive a consensus decision by combining the classification outcome of each classifier [10]. A Bayesian classification combination scheme is adopted, from which a precision boosting process is introduced for classification combination. The overall computational framework is shown in Figure 2.

3.1 Global scene classification using k-NN

In this work we use *k*-NN classifiers as *k*-NN is a simple, non-parametric classification method but has been effective in solving many classification problems, especially in difficult situations where no sufficient data is available for parametric estimation. Given a training set, upon receiving a new instance, a *k*-NN classifier identifies k nearest neighbour samples in a codebook and typically assigns the class label with the largest number of neighbours to the new instance. An additional advantage of using k-NN classifiers is that one can assign straightforwardly a soft class membership using the following approximation of the posterior probability:

$$P(w_i|f) = k_i/k \tag{1}$$

where w_i denotes the i-th class, f the feature code, k_i is the number of occurrence of class i among the k nearest neighbours. The classification decision is thus made:

$$c = \arg\max_i\{P(w_i|f)|i = 1, 2, ..., m\} \tag{2}$$

3.2 Region-based scene classification

Image regions segmented are likely to relate directly to semantic objects. If these objects can labelled reliably then one can expect the whole scene can be more accurately classified. While the global feature corresponds to an image class directly, the local features only represent the local properties of image regions, therefore another step of decision making is required to combine region labels into a scene category. For the two-class scene classification problem, we adopt a simple approach and assume class labelling of each segment is inherited from its parent image during training. With this groundtruth dataset we then train k-NN classifiers on the features extracted from the segments so as to match regional features to 'indoor' or 'outdoor' labels. For the classification of the whole image, it is first segmented into regions, and then *k*-NN classification on each region is done using Eq.(2). The final classification for the whole image can be obtained via majority voting of the labelled regions. For instance, assume there are N_{in} regions labelled as 'indoor' and N_{out} as 'outdoor' in an image. The probability for the image to be labelled as 'indoor' is

$$P(indoor|f) = \frac{N_{in}}{N_{in} + N_{out}} \tag{3}$$

where f denotes a regional feature used to classify the segmented regions and then the whole image.

3.3 Classifier validation

The performance of a classifier can be evaluated with precision and recall values. In our experiments, we perform a leave-one-out cross validation over the data set and compute the precision on indoor and outdoor classes separately.

3.4 Precision-boosted Bayesian multiclassifier combination

There are now four classifiers, two based on global colour and edge features and the other two on regional colour and edge features. As revealed by our empirical results to be shown later, none of these features leads to a very strong classifier. It is therefore necessary to investigate the combination of these individual classification decisions that gives a consensus decision. Various combination rules, such as product rule, sum rule, min rule, max rule, median rule, and majority voting can be used for this purpose [10].

However, there is another observation that the performance of these classifiers differ. Some are stronger, but others weaker, as indicated by validation results. It is our concern to tune the classifier combination so that strong classifiers gain more weights in a Bayesian combination scheme.

Assume a sample image x is associated with a feature set $F = \{f_1, f_2, f_3, .., .f_n\}$. For the sake of simplicity we further assume features in F are independent from each other. The probability that a sample belongs to class w_c ($c \in C = \{c_1, c_2, ..., c_m\}$) is:

$$P(w_c|x) = P(w_c|F) \tag{4}$$

From the Bayes' rule, we have:

$$\begin{aligned} P(w_c|F) &= \frac{p(F|w_c)P(w_c)}{p(F)} \\ &= \frac{P(w_c)\prod_{j=1}^{n} p(f_j|w_c)}{\sum_{i\in C} P(w_i)\prod_{j=1}^{n} p(f_j|w_i)} \end{aligned} \tag{5}$$

As

$$p(f_j|w_i) = \frac{P(w_i|f_j)p(f_j)}{P(w_i)},$$

and assuming all priori probabilities $P(w_i)$ are equal between classes, Eq.(5) can

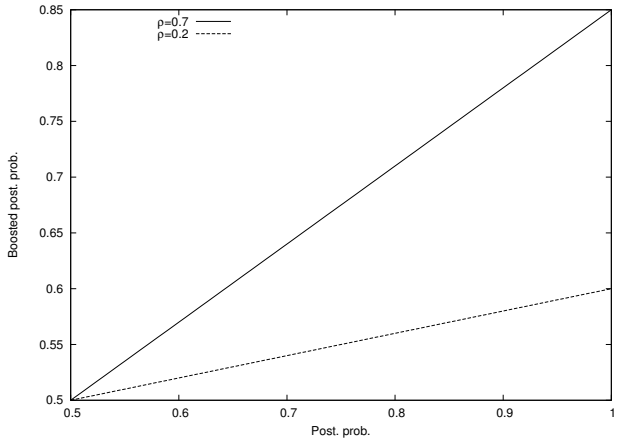


Fig. 3. Two examples of posterior probability boosting according to the precision of the classifier.

then be simplified as

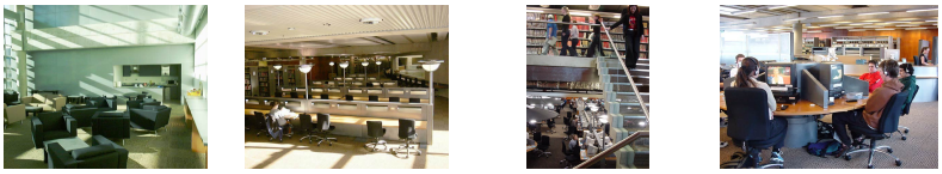
$$\begin{aligned} P(w_c|F) &= \frac{P(w_c)\prod_{j=1}^{n}\frac{P(w_c|f_j)p(f_j)}{P(w_c)}}{\sum_{i\in C}P(w_i)\prod_{j=1}^{n}\frac{P(w_i|f_j)p(f_j)}{P(w_i)}} \\ &= \frac{\prod_{j=1}^{n}P(w_c|f_j)}{\sum_{i\in C}\prod_{j=1}^{n}P(w_i|f_j)} \end{aligned} \tag{6}$$

This means that the combined classification result can be worked out from the posterior probabilities from individual classifiers using the above product rule.

As seen from Eq.(3), the posterior probabilities can be estimated from the k-NN results, but this approximation ability is in question when classifiers performance varies. Consequently we should give the classifier with high precision more weight when combining multiple classification decisions. Hence we boost the posterior probability of each classifier (if it is greater than 0.5) as:

$$P_b(w_c|f_j) = \frac{1}{2} + [P(w_c|f_j) - \frac{1}{2}]\rho(w_c, f_j), \tag{7}$$

where $\rho(w_c, f_j)$ is the precision of classifier j on class c obtained from cross validation. Figure 3 gives two examples of this precision-based boosting. The posterior probabilities in Eq.(6) can be better estimated with the precision-boosted value $P_b(w_c|f_j)$ and therefore is hopeful to give more accurate classification.



(a)

(b)

Fig. 4. Sample scene images: (a) indoor, and (b) outdoor.

A normalisation step is then carried out after the boosting:

$$P(w_c|f_j) = \frac{P_b(w_c|f_j)}{\sum_{l}^{n} P_b(w_c|f_l)}. \tag{8}$$

4 Empirical study

4.1 Groundtruth dataset

Due to the lack of benchmark data for the scene classification problem, we have to use our own image database as groundtruth. The image database consists of 153 photos that are taken during and after the construction of the Information Services Building (ISB) of University of Otago. A variety of pictures of constructional sites, completed building with outdoor background, indoor scenes of close-ups of library users or architectural structures etc. are included. Some of these indoor and outdoor images, as shown in Figure 4, have visually similar components especially of the architecture. It is challenging to classify these scenes with a high accuracy.

For training and testing purposes we hand-labelled each image as ‘indoor’ or ‘outdoor’, resulting in 60 (39%) images labelled as indoor scenes, and 93 (61%) images labelled as outdoor scenes. After segmentation, there are 1420 indoor segments and 1682 outdoor segments as training samples. The priori probabilities are close enough for their difference to be ignored.

4.2 Experiment results

In our experiment, we perform leave-one-out cross validation over the training samples and compute the precision and recall rates separately. For the global feature, we have two classifiers based on colour and edge features. Figure 5 shows the classification precision and recall on the *indoor* and *outdoor* classes. As we can see, the performances of the classifiers vary on indoor and outdoor scenes. For instance, the LUV classifier has relevant lower precision in indoor classification and higher precision in outdoor classification. Different classifiers also give different classification precision. One classifier may have a better precision but a worse recall. Yet to simplify the scenario we concentrate on the precision performance of these classifiers.

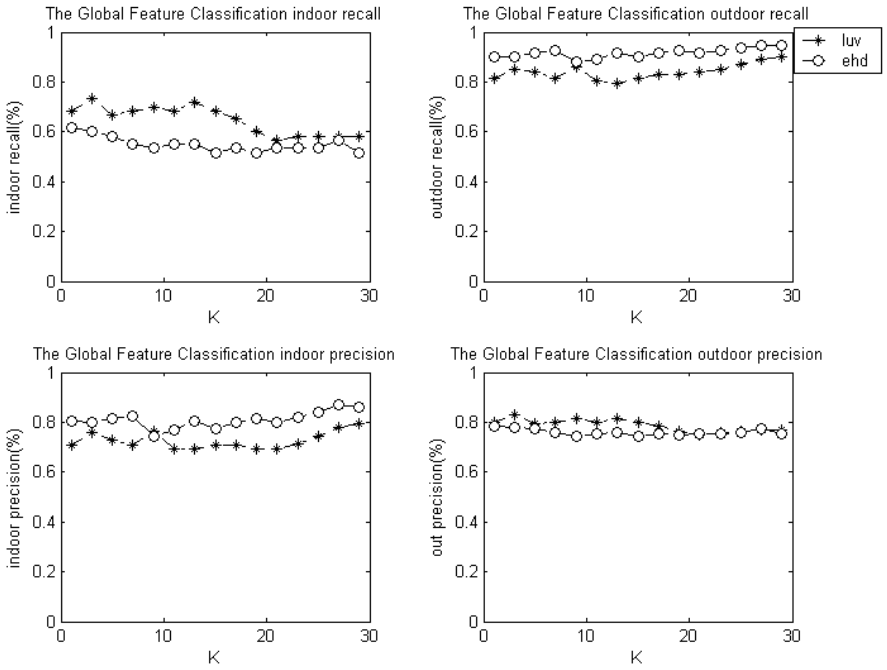


Fig. 5. Performance of the global classifiers.

EHD-based classifiers are also evaluated using a leave-one-out cross validation process. The precision of regional colour and regional EHD classifiers are shown in Figure 6.

A big difference of the performance of the classifiers between regional colour and edge features is observed. Regional colour classifiers work on outdoor images well but poorly on indoor images. Regional edge classifiers are just on the contrary. The reason may be that most segments within the outdoor image contain bright colours, such as sky, grass and buildings, while the colours of indoor image segments change inconsistently. On the other hand, most segments in indoor scenes have no edge, but the outdoor segments have edges in different directions. As indicated by the

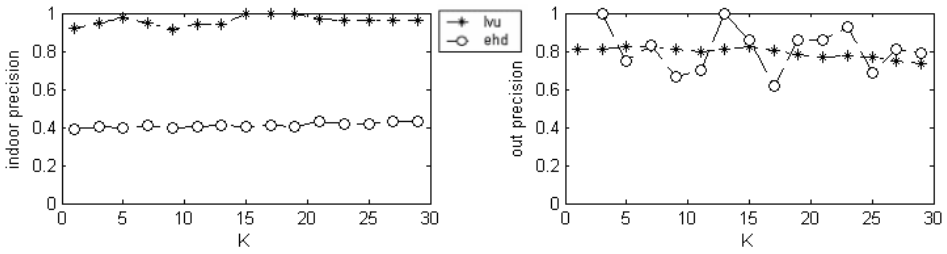


Fig. 6. Precisions of the regional classifiers.

performance metrics, no single classifier using either global or local features works well over all classes, which confirms the need for combining classification results from classifiers trained on both global and regional features. The cross validation results of combined classifiers are compared with those of individual classifiers, as shown in Figure 7.

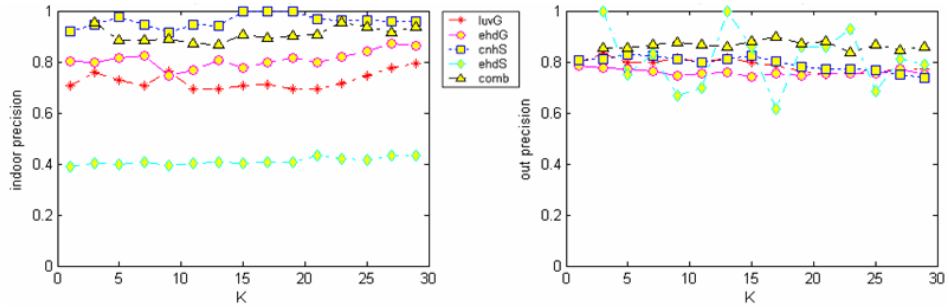


Fig. 7. Comparison of precision-boosted combination classifier with the individual classifier.

To achieve an optimal classification rate, four individual classifiers with high precision values are empirically chosen. They are:

- Global colour histogram (G-CH) classifier at $k = 9$,
- Global EHD (G-EHD) classifier at $k = 7$,
- Regional colour histogram (R-CH) classifier at $k = 13$ and
- Regional EHD (R-EHD) classifier at $k = 21$.

After implementing different combination rules for classification, the final classification accuracy values are listed in Table 1. It is clear that by using classification combination the scene classification accuracy is consistently improved, while the precision-based combination scheme gives the best result.

Some of the misclassified images are shown in Figure 8. These scenes are difficult to classify simply using colour and texture information, without the help of knowledge on the spatial layout or the structure of the buildings. Nevertheless, an overall accuracy about 90% scenes can be correctly classified, showing the effectiveness of our approach using content-based visual features and classifier combination.

We cannot compare these results directly with those reported in [5] as the dataset

Table 1
Classification accuracy using different classification schemes and different combination rules.

Classifier	Accuracy (%)
G-CH	79.7
G-EHD	77.8
R-CH	70.7
R-EHD classifier	44.4
Majority voting combination	80.4
Product combination	85.6
Precision-based combination	89.5

Fig. 8. Some misclassified images.

used there is a different one and is not available as a benchmark. However, our improvement of using precision-boosted classification combination over the conventional majority voting rule and product rule is obvious.

5 Conclusion

In this paper, we introduced a scene classification method using global and region features. A new classification combination rule using multiple precision-boosted classifiers is proposed and applied to the indoor-outdoor scene classification problem, where regional features of segmented image regions rather than fixed-size blocks are extracted together with global visual features. Empirical results have shown that the new classification combination scheme performs better than majority voting and the product rule.

While we have been focusing on a two-class problem, it is clear that this precision-boosted classification combination approach can be extended to multi-class problems. On the other hand, we hope to improve this approach by adopting more types of classifiers, and further test it on some benchmark image datasets. To do this we look forward to including more MPEG-7 XM features to enhance object classification capability. Further research will be on constructing some regional semantic models and then combining them to form a probabilistic semantic space of visual objects and visual concepts, so that more complicated scene classification and scene understanding can be achieved.

References

[1] A.W.M. Smeulders, M. Worring, and S. Santini and A. Gupta, "Content-based Image Retrieval of the end of the early years". *IEEE Trans on PAMI*, Vol. 22, No. 12, 2000, pp. 1349-1380.

[2] J.M. Corridoni, A.D. Bimbo, and P. Pala, "Image Retrieval by Colour Semantics", *Multimedia System*, Vol. 7, No. 3, 1999, pp.175-183.

[3] B.S. Manjunath, J. Ohm and V. Vinod, "Colour and Texture Descriptors", *IEEE Trans on Circuits and Systems for Video Technology*, Vol. 11, No. 6, 2001, pp.703-715.

[4] M. Bober, "MPEG-7 Visual Shape Descriptors", *IEEE Trans. on Circuits and Systems for Video Technology*, Vol. 11, 2001, pp.716-719.

[5] M. Szummer and R.W. Picard,"Indoor-Outdoor Image Classification," in *Proc. IEEE International Workshop on Content-based Access of Image and Video Databases*, 1998, pp.42-51.

[6] M. Soysal and A.A. Alatan, "Combining MPEG-7 Based Visual Experts For Reaching Semantics", in *Proc. of VLBV03*, Madrid, 2003.

[7] J. Li and J.Z. Wang, "Automatic Linguistic Indexing of Pictures by A Statistical Modelling Approach", *IEEE Trans. on PAMI*, vol. 25, No. 9, 2003, pp.1075-1088.

[8] MPEG-7 eXperimentation Model (XM), Institute for Integrated Systems, Munich University of Technology, Germany. URL http://www.lis.ei.tum.de/research/bv/topics/mmdb/e_mpeg7.html.

[9] Y. Deng and B.S. Manjunath, "Unsupervised segmentation of colour-texture regions in images and video", *IEEE Trans. on PAMI*, Vol. 23, 2001, pp.800-810.

[10] J. Kittler, Mohamad Hatef et al., "On Combination Classifiers", *IEEE Trans on PAMI*, Vol. 20, No. 3, 1998, pp.226-238.