A Indoor/Outdoor/Close-up Photo Classifier

R. Schettini°, C. Brambilla*, A. Valsasna°, M.De Ponti§
°ITIM, *IAMI, Consiglio Nazionale delle Ricerche,
Milano, Italy
§STMicroelectronics TPA Group, Printer Division,
Agrate Brianza, Italy

Abstract

Annotating images with a description of the content can be useful in processing images, by taking into account the scene depicted. We show here that it is possible to relate low-level visual features to semantic photo categories, such as indoor, outdoor and close-up, using CART classifiers. We have designed and experimentally compared several classification strategies, producing a classifier that can provide a reasonably good performance and on generic photographs no matter how acquired.

Introduction

The classification of photographs in semantic categories is an unresolved challenge in multimedia and imaging communities. Annotating images with a description of the content can facilitate the organization, storage and retrieval of image databases. It can also be useful in processing images, by taking into account the scene depicted, in intelligent scanners, digital cameras, photocopiers, and printers. The most appropriate strategies of image enhancement, color processing, compression, and rendering algorithms could be automatically adopted by the system (in a completely unsupervised manner) if the image content were automatically and reliably inferred by analyzing its low-level features, that is, features that can be computed without any a-priori knowledge of the subject depicted.

But there have been few efforts to automate the classification of digital color documents to date. Athitsos and Swain,[1] and Gever et al.,[2] have proposed automated systems for distinguishing photographs and graphics on the Word Wide Web. Schettini et. al.[3,4] have designed a method for distinguishing photographs from graphics and texts purely on the basis of low-level feature analysis. Szummer and Picard[5] have constructed algorithms for indoor/outdoor image classification. Vailaya et al.[6] have considered the hierarchical classification of vacation images: at the highest level the images are sorted into indoor/outdoor classes, outdoor images are then assigned to city/landscape classes, and finally landscape images are classified in sunset, forest, and mountain categories.

We present here our experimentation on indoor/ outdoor/close-up image classification. More specifically, we report the performance of different classification strategies based on the use of tree classifiers and exploiting low-level image features, such as color and texture distributions, to describe the image content.

Image Classification

To perform the classification we used tree classifiers constructed according to the CART methodology.[3,4,7] Briefly, these are classifiers produced by recursively partitioning the predictor space, each split being formed by conditions regarding to the predictor values. In tree terminology subsets are called nodes: the predictor space is the root node, terminal subsets are terminal nodes, and so on. Once a tree has been constructed, a class is assigned to each of the terminal nodes, and when a new case is processed by the tree, its predicted class is the class associated with the terminal node into which the case finally moves on the basis of its predictor values. The construction process is based on training sets of cases of known class. In the two experiments described here the predictors are the features indexing the whole image, and those indexing its subblocks, and the training sets are composed of images whose semantic class is known.

Tree classifiers compare well with other consolidated classifiers. Many simulation studies have shown their accuracy to be very good, often close to the achievable optimum. Moreover, they provide a clear understanding of the conditions that drive the classification process. Finally, they imply no distributional assumptions for the predictors, and can handle both quantitative and qualitative predictors in a very natural way.

Since in high dimensional and very complex problems, as is the case here, it is practically impossible, no matter how powerful the chosen class of classifiers, to obtain in one step good results in terms of accuracy we decided to perform the classification by also using what is called a 'perturbing and combining' method.[8,9] Methods of this kind, which generate in various ways multiple versions of a base classifier and use these to derive an aggregate classifier, have proved very successful in improving accuracy. We used bagging (bootstrap aggregating), since it is particularly effective when the classifiers are unstable, as trees are, that is, when small perturbations in the training sets, or in the construction process of the classifiers, may result in significant changes in the resulting prediction. With bagging the multiple versions of the base classifier are formed by making bootstrap replicates of the training set and using them as new training sets. The aggregation is made by majority vote. In any particular bootstrap replicate each element of the training set may appear repeated times, or not at all, since the replicates are obtained by resampling with replacement. Figure 1 shows how the resulting classifier, called the bagged classifier, is obtained.

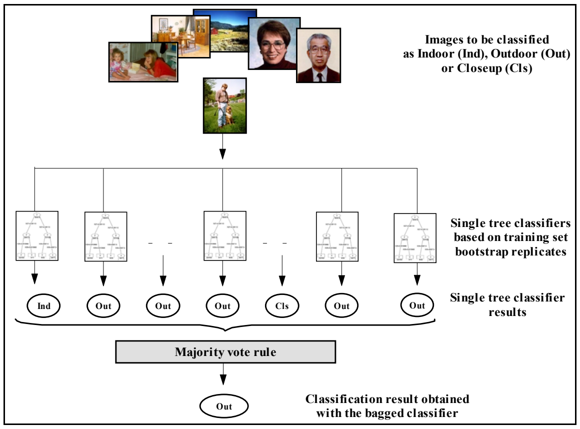

Figure 1. *The scheme of the bagged classifier.*

To provide a measure of confidence in the classification results and still greater accuracy, we applied an ambiguity rejection rule[10] to the bagged classifier: the classification obtained by means of the majority vote is rejected if the percentage of trees that contribute to it is lower than a given threshold. In this way only those results to which the classifier assigns a given confidence, as set by the threshold, are accepted. The rule is 'global' in the sense that it is constant over the feature space.

Image Description Using Low-level Features

We have used the following features to index the whole images and the subblocks identified by a 4 × 4 equally spaced grid:

- **Color Distribution,** described in terms of the moments of inertia (i.e. the mean, variance, skewness and kurtosis) of the distribution of hue, saturation and value.[11] The features of mean (E_i), variance (σ_i) and skewness (S_i) and Kurtosis (k_i)can be computed for the i-th color channel of the image I as follows:

$$E_i(I) = \frac{1}{XY} \sum_{x,y} I_i(x,y) \tag{1}$$

$$\sigma_i(I) = \sqrt{\frac{1}{XY} \sum_{x,y} (I_i(x,y) - E_i(I))^2} \tag{2}$$

$$s_i(I) = \sqrt[3]{\frac{1}{XY} \sum_{x,y} (I_i(x,y) - E_i(I))^3} \tag{3}$$

$$k_i(I) = \sqrt[2]{\frac{1}{XY} \sum_{x,y} \frac{(I_i(x,y) - E_i(I))^4}{(I_i(x,y) - E_i(I))^2}} \tag{4}$$

- **Edge Distribution,** the statistical information on image edges extracted by Canny's algorithm[12]:
  ♦ i) the percentages of low, medium, and high contrast edge pixels in the image;
  ♦ ii) the parametric thresholds on the gradient strength corresponding to medium and high contrast edges;
  ♦ iii) the number of connected regions identified by closed high contrast contours;
  ♦ iv) the percentage of medium contrast edge pixels connected to high contrast edges;
  ♦ v) the histogram of edge direction quantized in 18 bins;
- **Wavelets**. Multiresolution wavelet analysis provides representations of the image data in which both spatial and frequency information are present. In multiresolution wavelet analysis we have four bands for each level of resolution: a low-pass filtered version of the processed image, and three bands of details. Each band corresponds to a coefficient matrix one forth the size of the processed image. In our procedure the features are extracted from the luminance image using a three-step Daubechies multiresolution wavelet expansion producing ten sub-bands.[13] Two energy features, the mean and variance, are then computed for each subband;
- **Texture**. The estimate of texture features was based on the Neighborhood GrayTone Difference Matrix, i.e. coarseness, contrast, busyness, complexity, and strength[14,15];
- **Image Composition**. The HSV color space was partitioned into eleven color zones corresponding to basic color names (red, orange, yellow, green, blue, purple, pink, brown, black, gray and white). This partition was defined and validated empirically by different groups of examiners.[20] The spatial composition of the color regions identified by the process of quantization was described in terms of:
  ♦ fragmentation (the number of color regions),
  ♦ distribution of the color regions with respect to the center of the image, and
  ♦ distribution of the color regions with respect to the *x* axis, and with respect to the *y* axis.
  ♦ Interested readers may find a description of these features in references 16 and 19.
- **Skin Pixels**, the percentage of skin pixels. We used a statistical skin color detector based on the r, g chromaticities of the pixel; a training set of 30,000 color skin data was used to model the probability distribution of skin color.[17]
- **Spatial Chromatic Histogram** (SCH)**,**[18] extended histograms that preserve not only information about the

color content of the image, but also the spatial distribution of each color within the image. For each color, the entry in a SCH is composed of three values: the ratio of pixels in the image of the considered color(h), the baricenter (in relative coordinates) of the spatial distribution of the color (**b**), and the standard deviation of the distribution of color (σ). Combining histogram and spatial information requires a new distance function. Given two Spatial Chromatic Histograms H and H′ having c bins, the distance is computed as follows. Let

$$M_i = \min\left(\mathrm{h}_{\mathrm{H}}(\mathrm{i}) - \mathrm{h}_{\mathrm{H'}}(\mathrm{i})\right), \qquad (5)$$

where $h(i)$ is the ratio of pixels having color i, and

$$S_i = \frac{\sqrt{2} - \mathrm{d}\left(\mathrm{b}_{\mathrm{H}}(\mathrm{i}), \mathrm{b}_{\mathrm{H'}}(\mathrm{i})\right)}{\sqrt{2}} + \frac{\min\left(\sigma_{\mathrm{H}}(\mathrm{i}), \sigma_{\mathrm{H'}}(\mathrm{i})\right)}{\max\left(\sigma_{\mathrm{H}}(\mathrm{i}), \sigma_{\mathrm{H'}}(\mathrm{i})\right)} \qquad (6)$$

The distance is defined as

$$\mathrm{D}(\mathrm{H}, \mathrm{H'}) = \sum_{\mathrm{i}=1}^{\mathrm{c}} M_i S_i \qquad (7)$$

While all the features must be computed for the images in the training sets, only the features actually used by the classifier need to be computed for images in the test sets, and for new images processed by the classifiers. In general the features used in the classifiers we obtained are less then one third of the original ones.

Experimental Results

As said the problem was to classify a digital image as indoor, outdoor, and close-up. The indoor class included photographs of rooms, groups of persons, and details in which the context also indicated that they were taken inside. The outdoor class included natural landscapes, buildings, and city shots and details, in which the context concurred to indicate that the photographs were taken outside. The close-up class included portraits and photos of objects in which the context supplied no information of where the photo was taken.

The image database used in our experiments contained over 7400 images collected from various sources, such as images downloaded from the web, or acquired by scanner. It included some 1700 indoor images, 4200 outdoor images, and 1600 close-ups. All this material varied in size (ranging from 150 × 150 pixels to 900x900 pixels), resolution, and tonal depth.

In our first experiment we used as predictors the following features computed on the whole image: wavelet coefficients, statistical information and the direction histogram of image edges, texture estimators, the spatial blob composition, spatial-chromatic histogram, and skin detector.

Tables 1 and 2 show the classification accuracy achieved using a single tree classifier on the training and test sets respectively. The training set was equally distributed among the typologies present in the three classes, and contained about 4100 images (1100 indoor, 2100 outdoor and 900 close-up).

The test set contained some 3300 photos (600 indoor, 2000 outdoor and 700 close-up) which had not been utilized in the training set.

Table 1. Classification accuracy obtained on the training set using a single tree.

		Predicted class		
		Indoor	Outdoor	Closeup
True class	Indoor	0.96	0.02	0.02
	Outdoor	0.04	0.93	0.03
	Closeup	0.01	0.01	0.98

Table 2. Classification accuracy obtained on the test set using a single tree.

		Predicted class		
		Indoor	Outdoor	Closeup
True class	Indoor	0.72	0.14	0.15
	Outdoor	0.11	0.81	0.07
	Closeup	0.05	0.10	0.86

Tables 3 and 4 show, instead, the classification accuracy achieved on the training and test sets respectively, using a bagged classifier obtained by aggregating the trees based on 25 bootstrap replicates of the training set. As expected, the use of the bagged classifier produced a marked improvement in classification accuracy: by 7%, 9% and 4% in the test set, for the indoor, outdoor and close-up classes respectively. The aggregation of a larger number of trees brought no significant improvement.

Table 3 Classification accuracy obtained on the training set using the bagged classifier.

		Predicted class		
		Indoor	Outdoor	Closeup
True class	Indoor	0.99	0.01	0.00
	Outdoor	0.01	0.98	0.01
	Closeup	0.00	0.00	1.00

Table 4. Classification accuracy obtained on the test set using the bagged classifier.

		Predicted class		
		Indoor	Outdoor	Closeup
True class	Indoor	0.79	0.12	0.09
	Outdoor	0.06	0.90	0.04
	Closeup	0.02	0.08	0.90

The misclassified images are generally photographs that are either overexposed or underexposed, or with a background that provide little information about the class to which the images belong. Indoor images misclassified as outdoor often show a window, while outdoor images misclassified as indoor are images of building details, with little outdoor background. The misclassification of close-up images as indoor or outdoor and viceversa we consider acceptable: it simply reveals the overlapping between the close-up and the other categories.

We then performed a second experiment to extract the local inter-class differences, using as predictors the following features computed on 4 × 4 subblocks of the image: wavelet coefficients, statistical information and the direction histogram of image edges, texture estimators, the moment of inertia of the HSV distribution, and the skin detector. The images of the training and test sets are those used in the previous experiment.

Tables 5 and 6 show the classification accuracy achieved using a single classifier.

Table 5. Classification accuracy obtained on the training set using a single tree, trained with subblock indexing.

		Predicted class		
		Indoor	Outdoor	Closeup
True class	Indoor	0,98	0,00	0,02
	Outdoor	0,03	0,95	0,02
	Closeup	0,01	0,01	0,98

Table 6. Classification accuracy obtained on the test set using a single tree, trained with subblock indexing.

		Predicted class		
		Indoor	Outdoor	Closeup
True class	Indoor	0.66	0.20	0.14
	Outdoor	0.13	0.80	0.07
	Closeup	0.06	0.07	0.87

Tables 7 and 8, instead, register the accuracy reached on the training and test sets respectively using the bagged classifier.

Table 7. Classification accuracy obtained on the training set using the bagged classifier, in the case of subblock indexing.

		Predicted class		
		Indoor	Outdoor	Closeup
True class	Indoor	1.00	0.00	0.00
	Outdoor	0.00	0.99	0.01
	Closeup	0.00	0.00	1.00

Table 8. Classification accuracy obtained on the test set using the bagged classifier, in the case of subblock indexing.

		Predicted class		
		Indoor	Outdoor	Closeup
True class	Indoor	0.81	0.13	0.06
	Outdoor	0.06	0.90	0.04
	Closeup	0.03	0.05	0.92

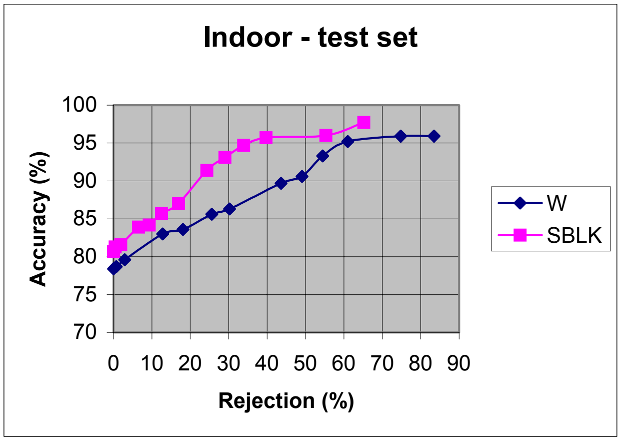

***Figure 2**. Accuracy-rejection trade-offs for the indoor class test set for both the whole (W) and the subblock (SBLK) indexing.*

In this experiment as well, the application of the bagged classifier improved classification accuracy: by 15%, 10% and 5% for the indoor, outdoor and close-up classes of the test set respectively.

Moreover, the 4 × 4 subblock indexing improved classification accuracy by 2% for both the indoor and the close-up classes of the test set, with respect to the whole image indexing.

In general the subblock indexing experiment presented the same misclassification problems as the whole image indexing experiment.

Figures 2, 3 and 4 show the effects of the application of the rejection rule in the accuracy-rejected plane, with varying rejection thresholds, for the indoor, outdoor and close-up classes respectively. Each figure shows the experimental results, for both the whole (W) and the subblock (SBLK) indexing, for the test set.

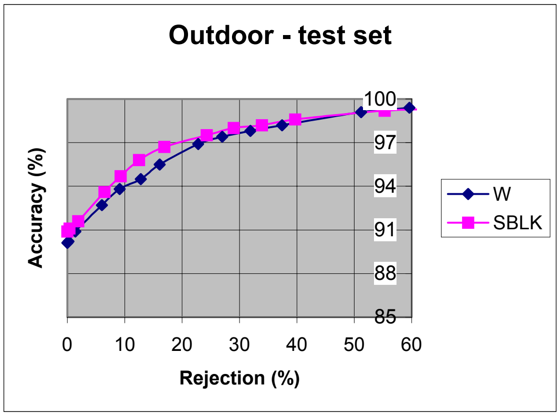

***Figure 3**. Accuracy-rejection trade-offs for the outdoor class test set for both the whole (W) and the subblock (SBLK) indexing.*

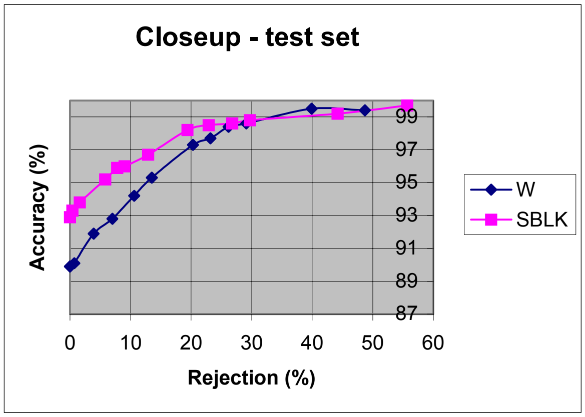

***Figure 4**. Accuracy-rejection trade-offs for the close-up class test set for both the whole (W) and the subblock (SBLK) indexing.*

The above figures reflect the better performance of the classifiers based on the 4 × 4 subblock image indexing.

Comparing figures 3 and 4 with figure 2, we note again the greater difficulty, already registered in Tables 4 and 8, in classifying indoor images. For both the outdoor and close-up classes, the 30% of rejected images corresponds to a near perfect classification accuracy.

Final Remarks

With the experiments described here, we have shown that it is possible to relate low-level visual features to semantic photo categories, such as indoor, outdoor and close-up, using CART classifiers. Specifically, we have designed and experimentally compared several classification strategies, producing a classifier that can provide a reasonably good performance and robustness not only on our database (over 7400 images collected from various sources) but also on generic photographs no matter how acquired. The results obtained have also allowed us to identify points that would benefit from further investigation:

- ***Image Indexing:*** We have seen that classification results were worst when significant parts of the images were occupied by skin regions, i.e., by people. As people may appear in any photograph of the classes considered, this information is not actually an discriminant in establishing the category to which a photo belongs. We are now refining the skin/people detector in order to make it much more stable (still assuming uncontrolled lighting conditions). We should then like to use significant skin regions to drive adaptive image partitions and then classify the parts of image that are not "occupied" by people. In fact, it is the context in which the subject is depicted that guides our interpretation of the scene.
- ***Classification Strategy***: The experiments performed have proved the feasibility of using the bagging method and the rejection rule to boost classification accuracy. These tools, however, could be further refined. In particular, we should like to create a more robust rejection rule by incorporating, together with the global measure proposed here, feature space-dependent information, such as the accuracy inside the terminal nodes.

Acknowledgement

This investigation was performed as a part of a ST Microelectronics' research contract (1999-2001). The authors thank ST Microelectronics for permission to present this paper.

References

1. V. Athitsos, M. Swain, Distinguishing photographs and graphics on the World Wide, Web. *Proc. Workshop in Content-based Access to Image and Video Libraries*, 10-17 (1997).
2. T. Gevers, AWM Smeulders, PicToSeek: combining color and shape invarinat features for image retrieval, *IEEE Trans. On Image Processing*, **19(1),** 102-120 (2000).
3. R. Schettini, C. Brambilla, G. Ciocca, M. De Ponti, Color image classification using tree classifiers, *Proc. VII Color Imaging Conference*: Scottsdale (Arizona), 269-272 (1999).
4. R. Schettini, C. Brambilla, A. Valsasna, M. De Ponti, Content-based image classification, Proc. Internet Imaging Conference, *Proceedings of SPIE* **3964** (G.B. Beretta, R. Schettini eds.), 28-33 (2000).
5. M Szummer, R. Picard, Indoor-outdoor image classification, *Proc. Int. Workshop on Content-Based Access of Image and Video databases*, 42-51 (1998).
6. A. Vailaya, M. Figueiredo, A. K. Jain, and H.-J. Zhang, Image classification for content-based indexing, *IEEE Transactions on Image Processing*, **10(1)**, 117-130 (2001).
7. L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, *Classification and Regression Trees*, Wadsworth and Brooks/Cole, 1984.
8. L. Breiman, Bagging predictors, *Machine learning*, **26**, 123-140 (1996).

9. L. Breiman, Arcing classifiers, *Annals of Statistics*, **26**, 801-849 (1998).
10. A. Vailaya and A. Jain, Reject option for VQ-based bayesian classification, *Proc. 15th International Conference on Pattern Recognition*, Barcelona (Spain), 2000.
11. M. A Stricker, M. Orengo, Similarity of color images, *SPIE Storage and Retrieval for Image and Video Databases III Conference*, (1995).
12. J. Canny, A computational approach to edge detection, *IEEE Trans. On Pattern Analysis and Machine Intelligence,* IEEE-**8,** 679-698 (1986).
13. P. Scheunders , S. Livens, G. Van de Wouwer, P. Vautrot, D. Van Dyck, Wavelet-based texture analysis, *International Journal Computer Science and Information Management.* wcc.ruca.ua.ac.be/~livens/WTA/, (1997).
14. M. Amadasun, R. King, Textural features corresponding to textural properties, *IEEE Transaction on System, Man and Cybernetics,* **19(5),** 1264-1274 (1989).
15. H. Tamura, S. Mori, T. Yamawaki, Textural features corresponding to visual perception, *IEEE Transaction on System, Man and Cybernetics,* **8,** 460-473 (1978).
16. P. Ciocca, R. Schettini, A relevance feedback mechanism for content-based retrieval, *Information Processing and Management,* **35,** 605-632 (1999).
17. Y. Miyake, H. Saitoh, H. Yaguchi, and N. Tsukada, Facial Pattern detection and color correction from television picture for newspaper printing, *Journal of Imaging Technology*, **16,** 165-169 (1990).
18. L. Cinque, G. Ciocca, S. Levialdi, A. Pellicanò, R. Schettini, Color-based image retrieval using spatial-chromatic histograms, *Image and Vision Computing*, 2001 (in print).
19. Y. H. Ang, S. H. Ong and Zhao Li, Retrieval of artifact images using multidimensional multiresolution features, *Computer & Graphics*, **20(1),** 51-59 (1996).
20. I. Gagliardi, R. Schettini, A method for the automatic indexing of color images for effective image retrieval, *The New Review of Hypermedia and Multimedia,* **3,** 201-224 (1997).