Digital Documents Classification for Optimized Processing and Rendering

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

Abstract

The problem addressed in this paper is the high-level problem of distinguishing among photographs, graphics, texts and compound documents. To cope with the great variety of compound documents we have designed a hierarchical classification strategy which first classifies images as compound or not-compound by verifying the homogeneity of the sub-images in terms of low-level features. Not-compound images are then classified as photographs, graphics or texts. Results of our experiments on a database of over 35000 images collected from various sources will be reported and discussed in the final paper.

Introduction

Digital imaging workflows have become increasingly complicated in the last few years. Many factors have driven the increased complexity of this arena: many different kinds of imaging devices are now available (Inkjet and Laser Printers, Scanners, Digital Copiers, Digital Still Cameras, Internet Faxes, Monitors, and Multifunctional products), and for each type of device there are many different subcategories (taking printers, for example, we have High and Low-end, Networked and standalone, PC-centric and Peer-to-peer products, …). Different driver-peripherals couples may also partition features differently, and functionalities and complex design vectors, such as speed, resolution, or the user-interface, must also be taken into account. Consequently next generation designs in this field must address several issues, such as versatility (devices must have more and more features, and be easier to use), data size (increased resolution means more data to manage, calling for better compression and data representation schemes), quality, processing speed and ease of insertion of devices in complex home and office networks (interoperability, plug-and-play, cross-device optimization).

We believe that content-based image classification will play an important role here: being able to properly classify text, graphics, photo and compound images will allow the unsupervised optimization of image data size and rendering intent using specific processing strategies. In this paper we address the problem of distinguishing among photographs, graphics, texts and compound documents using low-level features, such as color, edge distribution, and image composition. These low-level features were derived from a general purpose image indexing library, and, in designing such a library we have considered perceptual similarity (the feature distance between two images are large only if the images are not "similar"), efficiency (the features can be rapidly computed) and economy (their dimensions must be small in order not to affect classification efficiency).

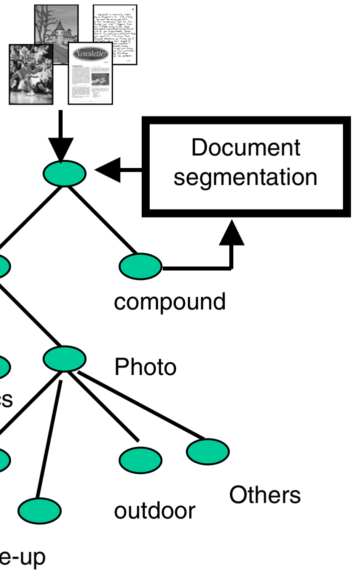

Figure 1. Hierarchical classification strategy.

Related Works

There have been few efforts to automate the classification of digital color documents to date. Athitsos and Swain,[2] and Gever et al.,[10] have proposed automated systems for distinguishing photographs and graphics on the Word Wide Web. Schettini et. al.[12,13] have designed a method for distinguishing photographs from graphics and texts purely on the basis of low-level feature analysis. Szummer and Picard[16] have constructed algorithms for indoor/outdoor image classification. Vailaya et al.[18] 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.

Schettini et. al.[14] related low-level visual features to semantic photo categories, such as indoor, outdoor and close-up, using CART classifiers. Specifically, they have designed and experimentally compared several classification strategies, producing a classifier that can provide a reasonably good performance and robustness on a database of over 7400 photos.

Classification Strategy

The hierarchical strategy we propose has been designed to address the high-level problem of distinguishing among photographs, graphics, texts, and compound documents and it is based on the use of tree classifiers built with the CART methodology.[3]

Cart classifiers are trees constructed by recursively partitioning the predictor space, each split being formed by conditions related to the predictor values. The process is binary: the predictor space and each subset of it are split exactly in two. In tree terminology the subsets are called nodes: the predictor space is the root node, terminal subsets are terminal nodes, and so on. Figure 2 shows an example of tree classifier. 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 our problem the predictors are the features indexing the images (the features used are listed in the following section), and the training sets are composed of images whose semantic class is known. The critical problems of the splitting process are essentially two: how to identify candidate splits, and how to define the goodness of the splits. Candidate splits are generated by a set of admissible questions regarding the values of the predictors, and differ according to the nature of the predictors themselves. The goodness of the splits depends basically on selecting the splits so that the data in the descendant nodes are purer than the data in the original ones. To do so, a function of impurity of the nodes is introduced, and the decrease in value of the chosen function produced by a split is taken as a measure of the goodness of the split itself. Since trees can be very large and overfit the data, the Cart methodology contemplates a pruning process based on the idea of finding a trade-off between the complexity and the accuracy of the trees. The overall performance of the trees is evaluated in terms of misclassification probability, or misclassification cost.

Recent work[4] has shown that the accuracy of Cart classifiers can be improved by perturbing and combining methods. This means generating multiple versions of a classifier by perturbing the training set, or the construction method, and then combining these multiple versions to produce a single classifier. The most natural way to combine different classifiers is voting. Hereafter a classifier obtained by perturbing and combining is called "classifier engine".

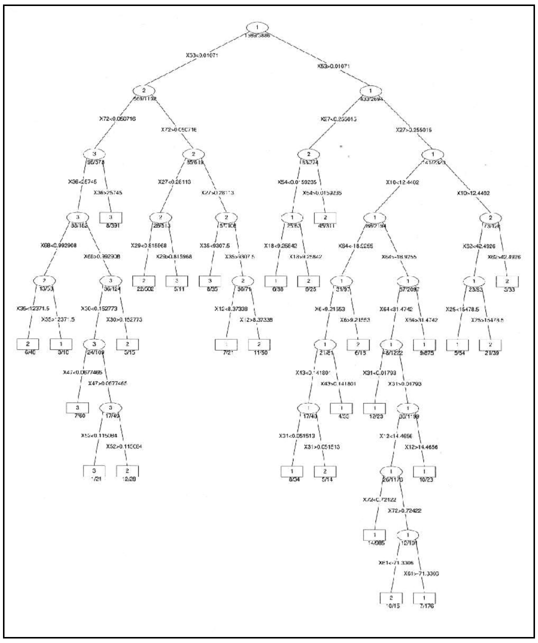

Figure 2. An example of tree classifier.

Tree classifiers built with the CART methodology present several advantages:

(i) they can handle the co-existence of different relationships between the features in different regions of the feature space in a very natural way;
(ii) they give a clear characterization of the conditions that determine when an image belongs to one class rather than to another, thereby detecting the most discriminant features for the problem addressed and unmasking redundancy;
(iii) they do not require assumptions about the probability distribution of the features;
(iv) they not only provide a classification rule, but also allow the assignment of a degree of confidence in the classification; and
(v) they may be very easily combined to derive an even more accurate classifier, as we have done.

The more straightforward way to address a classification problem with 4 classes would have been to use a 4-class classifier. However, the great variety and complexity of compound images would have required the definition of a huge training set without guaranteeing its completeness. Consequently we discarded this approach and defined the classification strategy described below and shown in Figure 3.

We first built and validated a "classifier engine" for the classification of photographs, graphics, and texts (the "classifier engine" was obtained by generating multiple tree classifiers and by combining these through a majority vote). We then used it to derive a compound vs. not-compound classifier: the images were subdivided into a given number of disjoint sub-images, and these were classified as photo, graphics, or text by the "classifier engine". A measure of confidence for the classification of each sub-images was provided by the percentage of trees, combined in the "classifier engine", that contributed to the result. The whole image was classified as compound if, with a "good" level of confidence, its sub-images were classified in at least two of the three different classes. Not-compound images were then globally classified as photograph, graphic or text. Images of dimensions (in pixels) smaller then a minimum threshold were excluded from the compound vs. not-compound classification, and classified directly globally as photograph, graphic, or text. This constraint was set because no strategy for compound document processing or analysis, such as region segmentation, or zone classification, could be useful or feasible in the case of images smaller then the chosen threshold.

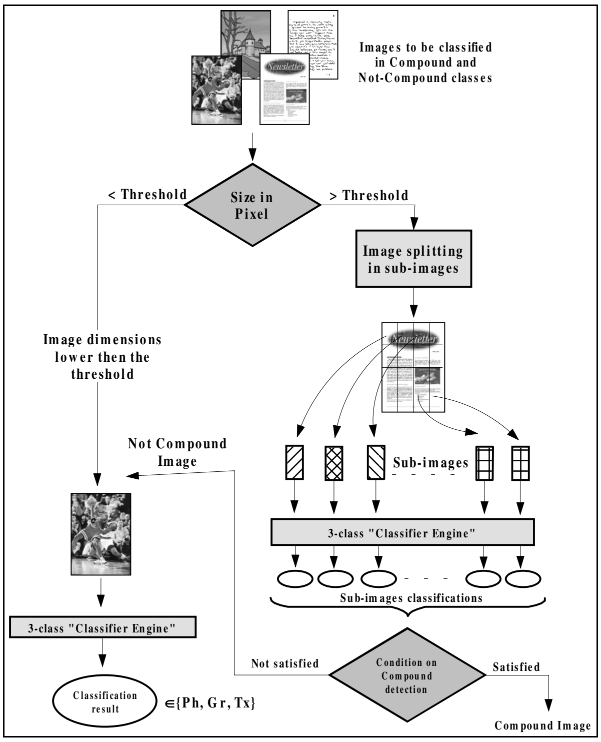
Images to be classified in Compound and Not-Compound classes
Size in Pixel
< Threshold
> Threshold
Image splitting in sub-images
Image dimensions lower then the threshold
Not Compound Image
Sub-images
3-class "Classifier Engine"
Sub-images classifications
3-class "Classifier Engine"
Not satisfied
Condition on Compound detection
Satisfied
Classification result
∈{Ph, Gr, Tx}
Compound Image

Figure 3. Compound document classification strategy.

The two biggest problems in the document subdivision performed for compound document detection are:

- a sub-image of a compound may also be a compound image itself, and then its classification in photo, graphics and text classes is a badly posed problem;
- a not-compound sub-image may be misclassified, while the whole image is not. Both these problems can be handled by prior analysis of the document to roughly detect the position of the homogeneous regions which may constitute it, and to utilize these as sub-images. Different subdivisions into sub-images could also be used, and the results of the corresponding classifications compared.

Image Indexing

We have used the following features to index the images:

- the image dimension;
- the moments of inertia, i.e. the mean, variance, skewness and kurtosis, of the distribution of the colors in terms of hue, saturation and value;[15]
- the percentage of "colored" and of "not-colored" pixels of the image;
- the statistical information on image edges extracted by Canny's algorithm: 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;[5]
- the mean and variance of the absolute values of the coefficients of the sub-images of the first three levels of the multi-resolution Daubechies wavelet transform of the luminance image;[6]
- the estimate of texture features based on the Neighborhood GrayTone Difference Matrix (NGTDM), i.e. coarseness, contrast, busyness, complexity, and strength;[1,7]
- the spatial composition of the color regions identified by a process of quantization in 11 colors: i) fragmentation (the number of color regions); ii) distribution of the color regions with respect to the center of the image; iii) distribution of the color regions with respect to the x axis, and with respect to the y axis;[7,8]
- the percentage of skin pixels detected by a skin region detector trained on a large amount of labeled skin data, e.g.[11,13]

Experimental Results

The image database used in our experiments consists of over 35000 images collected from various sources.

Compound and not-compound documents were correctly classified with an accuracy of 90% and 83% respectively. Among the not-compound documents, 10% of the photographs were misclassified as compound, while the figure for graphics and text was 20% (Table I). We also have observed that about 40% of the graphic and text images misclassified as compound, were misclassified by the 3-class classification as well.

Table II and Table III show the average classification accuracy of the “classifier engines” obtained on the training and test sets respectively.

Table I. Average classification accuracy of compound vs. not compound classification, evaluated on image classes.

		Predicted class	
		not Compound	Compound
True class	Photo	0.9	0.1
	Graphic	0.8	0.2
	Text	0.8	0.2
	Compound	0.1	0.9

Table II Average classification accuracy obtained on the training sets by using the "classifier engines".

		Predicted class		
		Photo	Graphic	Text
True class	Photo	0.99	0.1	0
	Graphic	0.03	0.94	0.03
	Text	0	0.1	0.99

Table III. Average classification accuracy obtained on the test sets by using the "classifier engines".

		Predicted class		
		Photo	Graphic	Text
True class	Photo	0.97	0.03	0
	Graphic	0.03	0.93	0.03
	Text	0	0.04	0.96

The photographs misclassified as graphics are mostly of small dimensions and low resolution, or object portraits with a uniform background; the graphics misclassified as photos are graphic illustrations with photo realistic intent, or smooth clip art; the graphics misclassified as texts are maps or tables with overlaid text, and the texts misclassified as graphics present a few colored words in large fonts or busy background.

Conclusions

We have presented a digital document classifier that can provide a reasonably good performance on a generic database of over 35000 images collected from various sources. We plan to refine the classification strategy in the near future and to integrate it in the system depicted in Figure 1.

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. M. Amadasun, R. King, Textural features corresponding to textural properties, IEEE Transaction on System, Man and Cybernetics, Vol. 19(5), pp. 1264-1274 (1989).
2. 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)
3. L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, Classification and Regression Trees, Wadsworth and Brooks/Cole, 1984.
4. L. Breiman, Bagging predictors, Machine learning, 26, 123-140 (1996).
5. J. Canny, A computational approach to edge detection, IEEE Trans. On Pattern Analysis and Machine Intelligence, IEEE-8, 679-698 (1986).
6. F. Idris, S. Panchanathan, Storage and retrieval of compressed images using wavelet vector quantization, Journal of Visual Languages and Computing, 8, 289-301 (1997).
7. G. Ciocca, I. Gagliardi, R. Schettini, Quicklook^2: An Integrated Multimedia system, Journal of Visual Languages and Computing, Vol 12(1), pp. 81-103, 2001 (SCI 5417).
8. L. Cinque, G. Ciocca, S. Levialdi, A. Pellicanò, R. Schettini, Color-based image retrieval using spatial-chromatic histograms, Image and Vision Computing, Vol. 19, pp. 879-986, 2001.
9. K.-C. Fan, C.-H. Liu, Y.-K. Wang, Segmentation and classification of mixed text/graphics/image documents, Pattern Recognition Letters, 15, 1201-1209, (1994).
10. T. Gevers, AWM Smeulders, PicTo Seek: combining color and shape invarinat features for image retrieval, IEEE Trans. On Image Processing, 19(1), 102-120 (2000).
11. Y. Miyake, H. Saitoh, H. Yaguchi, and N. TsukadaFacial Pattern detection and color correction from television picture for newspaper printing, Journal of Imaging Technology, 16, 165-169 (1990).
12. 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).
13. R. Schettini, G. Ciocca, A. Valsasna, C. Brambilla, M. De Ponti, A hierarchical classification strategy for digital documents, Pattern Recognition, 2002 (in print).
14. R. Schettini, A. Valsasna, C. Brambilla, M. De Ponti, A Indoor/Outdoor/Close-up Photo Classifier, Proc. IX Color Imaging Conference, Scottsdale (Arizona), pp. 35-40, 2001.
15. M. A. Stricker, M. Orengo, Similarity of Color Images. SPIE Storage and Retrieval for Image and Video Databases III Conference, (1995).
16. M. Szummer, R. Picard, Indoor-outdoor image classification, Proc. Int. Workshop on Content-Based Access of Image and Video databases, pp. 42-51 (1998).
17. H. Tamura, S. Mori, T. Yamawaki, Textural features corresponding to visual perception, IEEE Transaction on System, Man and Cybernetics, Vol. 8, pp. 460-473 (1978).
18. A. Vailaya, M. Figueiredo, A. K. Jain, and H.-J. Zhang, "Image Classification for Content-Based Indexing" , IEEE Transactions on Image Processing, Vol. 10(1), pp. 117-130 (2001).