Logo Detection in Document Images

Steve Seiden* Michael Dillencourt* Sandy Irani*
Roland Borrey† Timothy Murphy†

February 12, 1997

Abstract

This paper focuses on the problem of classifying segments of bitonal document images according to whether or not they are likely to contain a logo. A simple set of features is extracted from each segment, based primarily on statistics about the connected components of black pixels within the segment. The features are then used to construct a set of rules to determine the classification.

1 Introduction

This paper describes one part of a larger project whose goal is to analyze electronic documents and to determine their source based on an identifying logo. We divide this problem into three basic steps:

Segmentation The document is partitioned into smaller images, which contain logically related portions of the document (e.g., small text, picture, large text, logo, etc.)

Logo Detection Features are extracted from the segments which are then used to classify the segments based on whether or not they are likely to contain a logo.

Logo Matching Segments which have been identified as likely to contain logos are matched against a large database of logos to determine the exact identity of the logo.

This paper focuses on the middle step, logo detection. Since logo matching is likely to be a computationally intensive procedure, it is important to limit the number of segments which must be matched so that a document can be processed in a reasonable amount of time. Thus, the goal of the logo detection phase is to prune the set of segments which must be further analyzed by the logo matching phase. False negatives are more of a problem than false positives, since a false negative means that a logo is missed while a false positive means only

*Supported in part by Kofax Image Products Corporation and Microelectronics Innovation and Computer Research Opportunities. Address: Dept. of Info. and Comp. Science, University of California, Irvine, CA 92697-3425. Email:{sseiden,irani,dillenco}@ics.uci.edu

†Kofax Image Products, 3 Jenner Street, Irvine, CA 92718-3807. Email: {rborrey,tmurphy}@kofax.com

Number of Components	Aspect Ratio of Segment Bounding Box
Average Component Area	Segment Height
Average Component Height	Segment Width
Standard Deviation of Component Height	Maximum Height Relative to Segment Height
Average Component Width	Maximum Width Relative to Segment Width
Maximum Component Area	Segment Density
Maximum Component Height	Standard Deviation of Horizontal Density
Maximum Component Width	Standard Deviation of Vertical Density

Figure 1: *List of features used.*

that some amount of time and computer resources will be wasted by running that portion of the image through a logo matching procedure. Our goal is to determine which features of a segment are most useful in distinguishing logo from non-logo segments and the effectiveness of relatively simple features. It is desirable to keep the logo detection phase as independent as possible of any specific library of logos. In addition, in order to keep our method independent of assumptions about a document's layout, we do not use any information about the location of a segment on the original page.

Logo recognition has also been addressed by Doerman, Rivlin, and Weiss in [2]. They concentrate on the matching phase, using a much more complex set of geometric invariants.

2 Feature Extraction

Our system scans documents and produces a binary, or bitonal, image for each document. Each document is then run through a segmentation procedure, resulting in a set of image segments. The segmentation algorithm which we use is a top-down, hierarchical, X-Y tree segmentation scheme [3]. Once the segments are obtained, they are stored in a large set of segments and are no longer associated with the originating document.

From each segment, we produce a set of features, mostly derived from statistics about the connected components of the black pixels. The most computationally intensive portion of this process is calculating the connected components. Once these are obtained, determining the features is relatively straightforward. Although these features are quite simple, they are remarkably effective in distinguishing between logos and non-logos. Figure 1 lists the set of features used. All measurements of size are expressed in units of inches or square inches in order to keep the features independent of the resolution of the image.

3 Experiment

To test our approach, we collected a set of 130 images, mostly business documents such as letters and billing statements. We then segmented each image to produce a set of segments. Each segment in the set was hand classified according to whether it was a logo or a non-logo.

We augmented this collection of segments with the logos from the University of Maryland Logo Database [1]. There are 255 logo segments and 3907 non-logo segments. The reason for the numeric imbalance is that most business documents have many segments but very few logos (usually only one).

Our experiments then consisted of choosing a random subset of the segments (the *training set*), deriving a set of rules based on the training set, and testing the rules on the remaining segments (the *test set*). The rules are derived using Quinlan's C4.5 rule-based classification system [4].

The fact that there are many more non-logo segments than logo segments poses a potential problem: if the training set contains more non-logos than logos, the learning component might improve its performance by biasing the rules to increase the number of segments it classifies as non-logos. As indicated in the introduction, this would be undesirable, because false negatives are, from a practical viewpoint, worse than false negatives. We avoid this problem by choosing training sets that have the same number of logos as non-logos. That is, we sample by first deciding on a fraction x, then picking a random set of logos of size $x \cdot N$ where N is the total number of logo segments, and then picking a random set of non-logo segments of size $x \cdot N$.

For each given value of x, the experiment is run 100 times. The values measured are the average percentage of logos misclassified and the average percentage of non-logos misclassified. As x increases, the size of the training set increases, and we would expect the quality of the set of rules to improve and the number of misclassifications to drop.

4 Results and Interpretation

Figure 2 shows the observed relationship between the number of misclassified logos and the size of the training set. Even for a fairly small training set (30 logos out of 255, or about 12%), the number of misclassified logos is less than 10%. As the size of the training set increases, the number of misclassified logos shows a downward trend, but it remains above 6%.

The histogram in Figure 3 gives more information about the distribution of misclassified segments. This figure was generated by running 1000 trials with a training set of 150 logo segments. For each segment, we kept track of the number of times that it was in the test set, and then calculated the fraction of these times that it was misclassified. It can be seen from Figure 3 that a few logo segments are frequently misclassified, while most logo segments are correctly classified most of the time. For example, over half the logos are always classified correctly, 85% are misclassified less than 10% of the time, and 95% are misclassified less than 25% of the time.

Figure 4 shows several examples of logo segments that are frequently misclassified. The logos in Figures 4(b), (d), and (e) are difficult to recognize because they are purely textual logos. The reason why Figures 4(a), (c), and (f) are difficult appears to be the fact that each of these images contains a number fairly small components, and hence exhibits some of the properties of small text as measured by our feature vector.

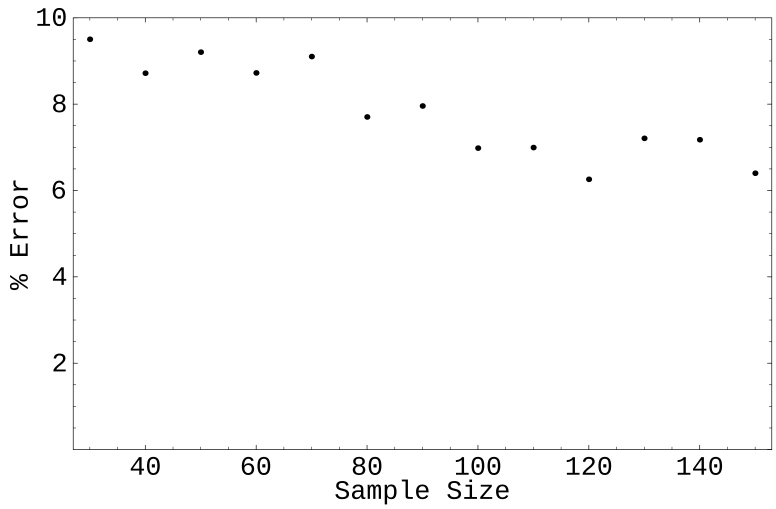

Figure 2: Percentage of logos misclassified as sample size varies.

5 Conclusion and Future Work

The preliminary results described here are encouraging. Our experiments indicate that logo detection, using only the simple set of features listed in Figure 1, can be done with a false negative rate well under 10%. We believe that these results can be improved considerably, and we suggest some approaches here.

One major problem illustrated by Figure 4 is the problem of detecting purely textual logos. One way to address this problem would be to throw the problem back onto the machine-learning component by performing a three-way classification into non-logos, graphic logos, and textual logos. Another possible solution would be to preprocess text segments using OCR techniques, in effect merging the detection and matching phases for textual logos.

A second problem is distinguishing logos that consist of several small connected components, such as Figures 4(a), (c), and (f). One possible approach to this problem is to augment the feature vector to include more aggregate data about the components (e.g., standard deviation of aspect ratio.) Topological information about the components (e.g., average number of holes) or periodicity information (e.g., Fourier coefficients for the vertical and horizontal density histograms) may also be useful in distinguishing graphical logos from pure text. Another useful feature might be the area of the bounding box of the largest component as a fraction of the entire segment area; this feature might help us distinguish Figure 4(c) from pure text.

There are various ways to take advantage of positional information that we have yet to exploit. For example, the position of a segment on a page may be a significant clue as to whether it is a logo. Also, merging adjacent segments in the XY-tree may lead to improved results, since some of the misclassification occurred on logos which where improperly

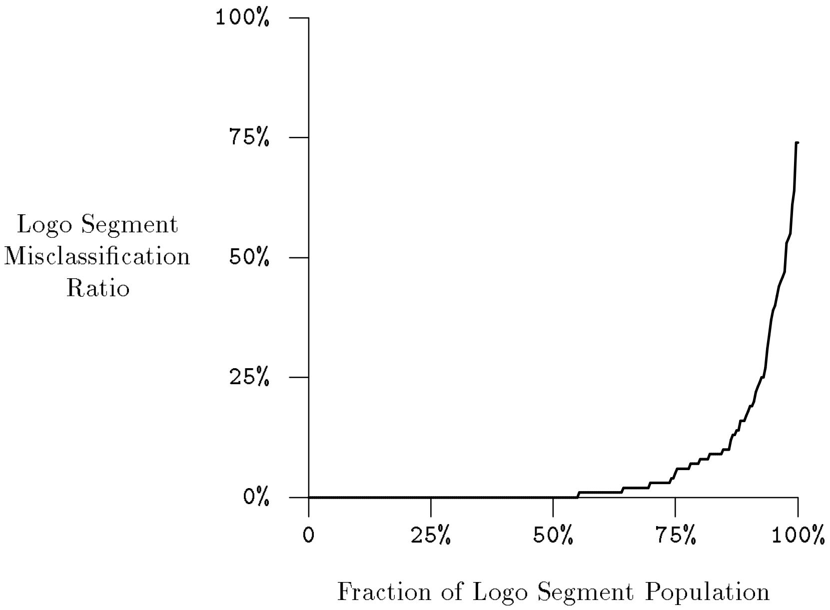

Figure 3: Histograms of misclassified logos

segmented. For example, Figure 4(a) is actually not an entire logo; rather, it is the left portion of the logo shown in Figure 5, which our system did a slightly better job of recognizing.

References

[1] D.S. Doermann, University of Maryland, Document Processing Group Logo Database, URL `http://documents.cfar.umd.edu/resources/database/UMDlogo.html`

[2] D.S. Doermann, E. Rivlin, I. Weiss. Logo recognition using geometric invariants. *Proceedings of the Second International Conference on Document Analysis and Recognition*, 1993. p. 894-7.

[3] G. Nagy and S. Seth. Hierarchical representation of optically scanned documents. *Seventh International Conference on Pattern Recognition*, vol. 1, 1984, pp. 347-9.

[4] J. R. Quinlan. *C4.5: Programs For Machine Learning.* Morgan Kaufmann, San Francisco, CA, 1993.

a)

b) SOUTHERN CALIFORNIA EDISON

c) IATA

d) AIR LINK (U.S.A.) INC.

e) MOTOROLA

f)

Figure 4: Examples of difficult logos.

KOFAX

Figure 5: Complete logo of Figure 4(a)