Document zone classification using machine learning

(EXTENDED ABSTRACT)

Stuart Inglis Ian H. Witten

Department of Computer Science, University of Waikato, Hamilton, New Zealand
email: {singlis,ihw}@cs.waikato.ac.nz

When processing document images, an important step is classifying the zones they contain into meaningful categories such as text, halftone pictures, line drawings, and mathematical formulae. A character recognition system, for example, may confine its attention to zones that are classified as text, while in an image compressor may employ specialized techniques and models for zones such as halftone pictures. The penalty for incorrect classification may range from incorrect interpretation to reduced efficiency. In any case, the higher the classification accuracy, the better the results.

This classification problem is not new. For example, Wang and Srihari [1] described a method for zone classification that gave 100% accuracy on the images they used for testing. But these images contained just 41 zones—and one of the categories occurred only once. Previous approaches to the problem generally describe a new set of features and assign classes using some linear weighted formula or nonlinear heuristic. In contrast, our work uses pre-defined features and investigates the application of standard machine learning methods, using a large publicly-available document database as a source of training and test data.

Database The University of Washington database contains 1001 document images, scanned at 300 dpi, that have been segmented into 13 831 different zones [2]. The zones have been classified manually into the twelve categories shown in Table 1 (which also gives the number of zones for each). Our work aims to discriminate the text, halftone and drawing zone types. Text zones, which are by far the most frequent, represent paragraphs, bibliography sections, page numbers and document titles. Many images have small non-zero skew angles, and in some cases text is oriented vertically instead of horizontally.

Features For each zone in the database we calculate seven features. All but the last pair of features relate to the zone's *components*, which are connected groups of black pixels. The features are:

- zone density—number of components per unit area;
- component density—mean density of black pixels in the components' bounding boxes;
- aspect ratio—mean ratio of height to width of the components' bounding boxes;
- circularity—the square of the perimeter divided by the area, averaged over all components;

- separation—mean distance between a component and its nearest neighbor;
- F_{short} and Fl_{ong} measures.

F_{short} and Fl_{ong} are textural measures based on the number of black and white horizontal runs of pixels; they emphasize short and long runs respectively [1].

Zone type	Frequency
text	12216
drawing	461
halftone	154
math	511
ruling	314
table	134
logo	15
map	14
advertisement	8
not-clear	2
announcement	1
seal	1
Total	13831

Table 1. Zones in the standard database

Machine learning methods learn how to classify instances automatically into known categories. The process involves a training stage followed by a testing stage. Two broad classes of machine learning methods are *rule-based* and *instance-based* methods. The former derive a rule set that is evaluated to determine the class of an unknown instance, while the latter retain a list of examples of each class and match unknown instances by seeking the closest known example. Examples of rule-based methods are C4.5 [3], Induct [4] and FOIL [5]; examples of instance-based methods are IB1 [6] and K* [7]. Our work is performed using the WEKA machine learning workbench, which incorporates these five methods. As well as providing facilities for interactive exploration, it allows experiments to be run in batch mode and assists with report generation [8].

Results A preliminary experiment on discrimination between the text, halftone and drawing zone types used a subset of 1500 text zones along with the full complement of drawing and halftone zones, giving a total of 2100 examples. Test and training sets were obtained by dividing this set in half, and the results were averaged over 25 runs. Table 2 shows the results. The five ML methods give similar results of around 90% correct, with C4.5 achieving 93%. The standard deviation of the classification percentage is given in brackets.

Table 3 gives a breakdown of the misclassifications made by C4.5. For example, of the 740 text zones in the test set, 717 (96.9%) were correctly classified as text, 20 (2.7%) were erroneously classified as drawings, and 3 (0.4%) were erroneously classified as halftones. The largest percentage errors are for halftones misclassified as drawings (11%) and drawings misclassified as text (9%).

Method	Classification %	
C4.5	93.3	(± 0.9)
IB1	90.8	(± 0.6)
FOIL	89.2	(± 1.1)
K*	87.6	(± 1.0)
Induct	84.3	(± 5.5)

Table 2. Differentiating text, halftone and drawings

Zone class	Text	Drawing	Halftone
Text	717	20	3
Drawing	20	192	18
Halftone	0	9	69

Table 3. Breakdown of classifications for text, drawing and halftone

This pattern of misclassifications shows that very few errors are made between text and halftone, and indeed the same techniques can be used to distinguish between text and halftone to an accuracy of over 99%. Distinguishing text from non-text is somewhat less successful, and can be performed at around the 96–97% level of accuracy. A much larger error rate, of 12% or more, is encountered when distinguishing halftone from drawing.

Discussion Some of the errors can be attributed to peculiarities in the classifications in the University of Washington database. For example, a large proportion of the images classified as halftone in fact have no halftone patterns visible—due perhaps to repeated photocopying with a resolution less than the halftone grid. Many contain very large expanses of black or white, again possibly caused by photocopying. An extreme example is a photograph of a face which comprises just five large components, but is nevertheless classified as a halftone picture. Other images contain textual labels and descriptions such as figure captions and numbers.

Almost all of the “drawing” zones are flowcharts, diagrams, and graphs that contain some text. In some cases the zone is virtually entirely text, and a few lines or arrows that border the text presumably account for its classification as a drawing. Some of the other “drawing” examples were marred by a large number of small specks, which is normally one of the attributes of halftones.

Conclusion Our methodology for document zone classification is to take a large standard database of images and derive feature values from each zone. Instead of devising features and adjusting parameters to yield good results on particular test files, a comprehensive set of features is passed to a machine learning scheme, which picks the most important of them automatically.

A 93% success rate is achieved with a three-way document zone classification. This improves to 97% when just two classes—text and non-text—are used. Text can be distinguished from halftones with 99% accuracy. Distinguishing halftones from drawings can be done with only about 88% accuracy. Future work includes the use of more descriptive features to help separate these two classes.

With an pre-classified database and a workbench such as WEKA that allows easy experimentation using different machine learning schemes, accurate rules can quickly be derived for document zone classification. There does not seem to be a great deal of difference in performance between the various machine learning schemes, and the widely-available C4.5 program does best of all. This methodology keeps attention firmly focused on creating and analyzing good features, rather than on *ad hoc* heuristics for combining them into an overall judgment.

Bibliography

[1] D. Wang, and S.N. Srihari, “Classification of Newspaper Image Blocks Using Texture Analysis,” *Computer Vision, Graphics and Image Processing 47,* pp. 327–352, 1989.

[2] University of Washington English Document Image Database I, Seattle, WA, USA, 1993.

[3] J.R. Quinlan, C4.5: *Programs for Machine Learning,* Morgan Kaufmann, 1992.

[4] B.R. Gaines, “The tradeoff between knowledge and data in knowledge acquisition in knowledge discovery in databases,” *AAAI Press*, pp. 491–505, 1991.

[5] J.R. Quinlan and R.M. Cameron-Jones, “FOIL: a midterm report,” *Proc European Conference on Machine Learning,* pp. 3–20. Springer Verlag, 1993.

[6] D.W. Aha, D. Kibler, and M.K. Albert, “Instance-based learning algorithms.” *Machine Learning 6*, pp. 37–66, 1991.

[7] J.G. Cleary and L. Trigg, “K*: an instance-based learner using an entropic distance measure.” *Proc Int Conference on Machine Learning,* Morgan Kaufmann, 1995.

[8] G. Holmes, A. Donkin, and I.H. Witten, “WEKA: a machine learning workbench,” *Proc Australian and New Zealand Conference on Intelligent Information Systems*, Brisbane, Australia, pp. 357–361, 1994, URL: http://www.cs.waikato.ac.nz/~ml.