DOCUMENT IMAGE ZONE CLASSIFICATION
A Simple High-Performance Approach

Daniel Keysers, Faisal Shafait
German Research Center for Artificial Intelligence (DFKI) GmbH, Kaiserslautern, Germany

Thomas M. Breuel
Technical University of Kaiserslautern, Germany

Keywords: Document Image Analysis, Zone Classification.

Abstract: We describe a simple, fast, and accurate system for document image zone classification — an important subproblem of document image analysis — that results from a detailed analysis of different features. Using a novel combination of known algorithms, we achieve a very competitive error rate of 1.46% ($n = 13811$) in comparison to (Wang et al., 2006) who report an error rate of 1.55% ($n = 24177$) using more complicated techniques. The experiments were performed on zones extracted from the widely used UW-III database, which is representative of images of scanned journal pages and contains ground-truthed real-world data.

1 INTRODUCTION

One important subtask of document image processing is the classification of blocks detected by the physical layout analysis system into one of a set of predefined classes. For example, we may want to distinguish between text blocks and drawings to pass the former to an OCR system and the latter to an image enhancer. For a detailed discussion of the task and its relevance please see e.g. (Wang et al., 2006).

During the design of our block classification system we noticed that among the approaches we found in the literature a detailed comparison of different features was usually not performed, and in particular we did not find a comparison that included features as they are typically used in other image classification or retrieval tasks. In this paper we address this shortcoming by comparing a large set of commonly used features for block classification and include in the comparison three features that are known to yield good performance in content-based image retrieval (CBIR) and are applicable to binary images (Deselaers et al., 2004). Interestingly, we found that the single feature with the best performance is the Tamura texture histogram, which belongs to this latter class. Another result we transfer from experience in the area of CBIR is that often a histogram is a more powerful feature than using statistics of a distribution like mean and variance only. We show that the use of histograms improves the performance for block classification significantly in our experiments. By combining a number of different features, we achieve a very competitive error rate of less than 1.5% on a data set of blocks extracted from the well-known University of Washington III (UW-III) database. In addition to the data used in prior work we include a class of ‘speckles’ blocks that often occur during photocopying and for which a correct classification can facilitate further processing of a document image. Figure 1 shows example block images for each of the eight types distinguished in our approach. We also present a very fast (but at 2.1% error slightly less accurate) classifier, using simple features and only a fraction of a second to classify one block on average on a standard PC.

2 RELATED WORK AND CONTRIBUTION

We briefly discuss some related work in this section, for a more detailed overview of related work in the field of document zone classification please refer to (Okun et al., 1999; Wang et al., 2006). Table 1 shows an overview of related results in zone classification.

Inglis and Witten (Inglis and Witten, 1995)

Keysers D., Shafait F. and M. Breuel T. (2007).
DOCUMENT IMAGE ZONE CLASSIFICATION - A Simple High-Performance Approach.
In *Proceedings of the Second International Conference on Computer Vision Theory and Applications - IU/MTSV*, pages 44-51

Table 1: Summary of UW zone classification error rates from the literature along with the number of pages, zones and block types used. Note that an exact comparison between all error rates is not possible.

reference	# pages	# zones	# types	error [%]
(Inglis and Witten, 1995)	1001	13831	3	6.7
(Liang et al., 1996)	979	13726	8	5.4
(Sivaramakrishnan et al., 1995)	979	13726	9	3.3
(Wang et al., 2000)	1600	24177	9	2.5
(Wang et al., 2006)	1600	24177	9	1.5
this work	713	13811	8	1.5

present a study of the zone classification problem as a machine learning problem. They use 13831 zones from the UW database and distinguish the three classes text, halftone, and drawing. Using seven features based on connected components and run lengths, the authors apply various machine learning techniques to the problem, of which the C4.5 decision tree performs best at 6.7% error rate.

The review paper by Okun et al. (Okun et al., 1999) succinctly summarizes the main approaches used for document zone classification in the 1990s. The predominant feature type is based on connected components (see also for example (Liang et al., 1996)) and run-length statistics. Other features used include the cross-correlation between scan-lines, vertical projection profiles, wavelet coefficients, learned masks, and the black pixel distribution. The most common classifier used is a neural network.

The widespread use of features based on connected components run-length statistics, combined with the simplicity of implementation of such features, led us to use these feature types in our experiments as well, comparing them to the use of features used in content-based image retrieval. Our CBIR features are based on the open source image retrieval system FIRE (Deselaers et al., 2004). We restrict our analysis for zone classification to those features that are promising for the analysis of binary images as described in the following section. (The overall most successful features in CBIR are usually based on color information.)

The most recent and detailed overview of the progress in document zone classification and a very accurate system is presented in (Wang et al., 2006). The authors use a decision tree classifier and model contextual dependencies for some zones. In our work we do not model zone context, although it is likely that a context model (which can be integrated in a similar way as presented by Wang et al.) would help the overall classification performance. Wang et al. use 24177 zones extracted from the UW-III database to evaluate their approach. In our experiments we use only 11804 labeled zones (plus 2007 additional zones of type 'speckles') extracted from the UW-III database because many zones occur in different versions in the database. In Section 5 we further illustrate this shortcoming and our approach to overcome it. As the authors use 9-fold cross-validation to obtain their results, it might be possible that the error rates they present (the best result is an overall error rate of 1.5%) may be influenced positively by this fact, because it is likely that instances of blocks of the same document occur in training and test set. In a similar direction, Wang et al. use one feature that "uses a statistical method to classify glyphs and was extensively trained on the UWCDROM-III document image database." It is not clear to us if this implies that the glyphs that occur in testing have also been used in the training of the glyph classifier.

We expand on the work presented in (Wang et al., 2006) in the following ways:

- We include a detailed feature comparison including a comparison with commonly used CBIR features. It turns out that the single best feature is the Tamura texture histogram which was not previously used for zone classification.
- We present results both for a simple nearest neighbor classifier and for a very fast linear classifier based on logistic regression and the maximum entropy criterion.
- We introduce a new class of blocks containing speckles that has not been labeled in the UW-III database. This typical class of noise is important to detect during the layout analysis especially for images of photocopied documents.
- We present results for the part of the UW-III database without using duplicates and achieve a similar error rate of 1.5%.
- We introduce the use of histograms for the measurements of connected components and run lengths and show that this leads to a performance increase.

3 FEATURE EXTRACTION

We extract the following features from each block, where features 1-3 are chosen based on their performance in CBIR (Deselaers et al., 2004) feature 4 was expected to help distinguish between the types 'drawing' and 'text' and features 5-9 were chosen based on their common use in block classification (Okun et al., 1999; Wang et al., 2006). Due to space limitations we refer the interested reader to the references for implementation details.

1. Tamura texture features histogram (TTFH)
2. Relational invariant feature histograms (RIFH)
3. Down-scaled images of size 32×32 (DSI)
4. The fill ratio, i.e. the ratio of the number of black pixels in a horizontally smeared (Wong et al., 1982) image to the area of the image (FR)
5. Run-length histograms of black and white pixels along horizontal, vertical, main diagonal, and side diagonal directions; each histogram uses eight bins, spaced apart as powers of 2, i.e. counting runs of length $\leq 1,3,7,15,31,63,127$ and ≥ 128 (RL{B,W}{X,Y,M,S}H)
6. The vector formed by the total number, mean, and variance of the runs of black and white pixels along the horizontal, vertical, main diagonal, and side diagonal directions as used in (Wang et al., 2006) (RL{B,W}{X,Y,M,S}V)
7. Histograms (as in 5) of the widths and heights of connected components (CCXH, CCYH)
8. The joint distribution of the widths and heights of connected components as a 2-dimensional 64-bin histogram (CCXYH)
9. The histogram of the distances between a connected component and its nearest neighbor component (CCNNH)

4 CLASSIFICATION

To evaluate the various features, we use a simple nearest neighbor classifier, that is, a test sample is classified into the class the closest training sample belongs to. The distance measures used are the Jensen-Shannon divergence for histograms and the Euclidean distance for all other features (Deselaers et al., 2004). If different feature sets are combined, the overall distance is calculated as the weighted sum of the individual normalized distances. The weights are proportional to the inverse of the error rate of a particular feature. No tuning with respect to these weights or with respect to the distance measures has been performed. Although a *k*-nearest-neighbor approach gives better results in many cases we only evaluated the 1-nearest-neighbor classifier. The nearest neighbor error rates are determined using leave-one-out cross-validation.

The nearest neighbor classifier serves as a good baseline classifier, although in many cases we can find a more suitable classifier for a given task. As we concentrate on features in this paper, we did not test any other classifiers. However, an important shortcoming of the nearest neighbor classifier is its requirement on computational resources. Both memory and run-time can be prohibitive for some applications. To explore a very fast approach with minimum requirements on computational resources, we also trained a log-linear classifier using the maximum entropy criterion (Keysers et al., 2002). The classification using this classifier can be obtained by computing a dot product of the feature vector with a weight vector for each class and choosing the maximum, and is thus very fast. As only these weight vectors need to be stored, the memory requirement is also minimal. Furthermore, the maximum entropy approach yields a probabilistic model, such that we obtain an estimate of the posterior probability for each class. The maximum entropy approach was evaluated on a regular 50/50 split of the data into training and test set and thus only uses half the amount of training data. The histograms were not normalized for the maximum-entropy approach, but the absolute numbers were used instead to allow the classifier to utilize this additional information.

5 DATA SET

To evaluate our approach for document zone classification, we use the University of Washington III (UW-III) database (Guyon et al., 1997). The database consists of 1600 English document images with bounding boxes of 24177 homogeneous page segments or blocks, which are manually labeled into different classes depending on their contents, making the data very suitable for evaluating a block classification system, e.g. (Inglis and Witten, 1995; Wang et al., 2006).

The documents in the UW-III dataset are categorized based on their degradation type as follows:

1. Direct scans of original English journals
2. Scans of first generation English journal photocopies
3. Scans of second or later generation English journal photocopies

Table 2: Leave-one-out nearest neighbor error rates and extraction run-times for each feature and for combinations.

feature	# features	extr.-time [s]	error [%]
TTFH	512	5.51	**3.4**
RIFH	512	12.59	7.8
DSI	1024	0.01	8.1
FR	1	0.02	27.3
RLBXH	8	0.01	7.9
RLWXH	8	0.01	5.1
RLBYH	8	0.01	8.2
RLWYH	8	0.01	5.6
RLBMH	8	0.01	11.8
RLWMH	8	0.01	6.6
RLBSH	8	0.01	10.5
RLWSH	8	0.01	6.2
RLBXV	3	0.01	12.9
RLWXV	3	0.01	9.7
RLBYV	3	0.01	14.6
RLWYV	3	0.01	12.1
RLBMV	3	0.01	17.2
RLWMV	3	0.01	12.6
RLBSV	3	0.01	16.7
RLWSV	3	0.01	12.2
CCXH	8	0.04	14.5
CCYH	8	0.04	14.9
CCXYH	64	0.04	6.2
CCNNH	8	0.05	19.0
RL**V, constant weight			4.1
RL**H, constant weight			1.8
RL*, CC*, 1/error weight			**1.5**
FR, RL*, CC*, 1/error weight			1.5
TTFH, FR, RL*, CC*, 1/error weight			1.5
RL*, CC*, *logistic, 50/50 data split*			2.1

show results for combined feature sets.

We can observe the following results:

- The Tamura texture feature is the single best feature but is more than 100 times slower to compute than most other features.
- The use of histograms as descriptors of the run-lengths distribution leads to much lower error rates than the use of number, mean, and variance. The combination of these histograms alone leads to a very good error rate of 1.8%.
- Interestingly, the use of the white (background) runs for the computation of features consistently leads to better results than the use of black (foreground) runs.
- Among the run-lengths based features, those based on the horizontal runs lead to the best error rates.
- The fill ratio as a single feature does not lead to good results, which is not surprising as it consists only of a single number. However, it is very useful to distinguish drawings from text. This is however also achieved by using the distribution of the white run lengths, such that the FR feature is not part of the best observed feature set.
- By using a logistic classifier trained with the maximum entropy criterion (training time a few minutes, time for one classification in the order of a few microseconds) on a feature set that is very fast to extract, we can construct a zone type classifier that can classify more than five zones per second even without performance tuning. At the same time, the error rate is at 2.1% only slightly higher than that of the best observed classifier.

Table 3 shows the frequency of misclassifications between different classes of the best classifier. We can observe that high recognition accuracy was achieved for the text, ruling, speckles, math, halftone, and drawing classes. However, our system failed to recognize logos correctly, and most of the logos were misclassified as either text, or halftone/drawing. Note that the accuracy rate for type ‘logo’ in (Wang et al., 2006) is even lower at 0.0%. The reason for this effect is the very small number of samples for this class, which on the other hand implies that it has only a very small influence on the overall system error rate. Similarly, the table detection accuracy was not high, and about 21% of the tables were misclassified as text.

To visualize the errors made, we looked at the nearest-neighbor images for each misclassified block. Figure 3 shows some typical examples. It can be seen that some of these images cannot be simply classified correctly by using the block content alone, and even humans are likely to make errors if they are asked to classify these images.

7 CONCLUSION

From the analysis of the obtained results we can conclude that we can construct a very accurate classifier based on run-lengths histograms alone. These features are very easy to implement and fast to extract and thus should be part of any practical baseline system. Interestingly, the distribution of the background runs is more important for document zone classification than the distribution of the foreground runs. Including a few more features based on run-length and connected component measurements we achieved a very competitive[1] error rate of below 1.5% on zones extracted form the UW-III database without

[1] For a comparison to our results also note that at most 0.2% (53/24177) of the error rate Wang et al. present is

Table 3: Contingency table showing the distribution of the classification of zones of a particular type in percent. (The total number of errors equals 201 within 13811 tests.) The labels M, L, T, A, D, H, R, S correspond to the types math, logo, text, table, drawing, halftone, ruling, and speckles, respectively.

	M	L	T	A	D	H	R	S	error [%]	# samples
M	90.8	0.0	8.6	0.0	0.0	0.6	0.0	0.0	9.2	476
L	9.1	27.3	36.4	0.0	9.1	9.1	0.0	9.1	72.7	11
T	0.1	0.0	99.8	0.0	0.0	0.0	0.0	0.0	0.2	10450
A	0.8	0.0	20.7	68.6	9.9	0.8	0.0	0.0	31.4	121
D	1.5	0.3	3.0	5.5	86.0	3.5	0.0	0.3	14.0	401
H	0.0	0.9	0.0	0.0	9.7	86.7	0.9	1.8	13.3	113
R	0.4	0.0	1.3	0.0	0.4	0.0	96.1	2.2	3.9	232
S	0.1	0.0	0.5	0.0	0.1	0.1	0.0	99.4	0.6	2007

the need for features based on glyphs or the Fourier transform. By employing a fast logistic (log-linear) classifier trained using the maximum entropy criterion on these features, we arrived at a fast and accurate, yet easy to implement overall classifier with a slightly higher error rate of 2.1%. In our experiments we did not use context information as done in (Wang et al., 2006) and thus could keep the decision rule very simple. However, context models are likely to help in the overall classification and an inclusion of our approach into Wang et al.'s context model is possible. Examining the errors made by the system makes it seem likely that further improvements significantly below the reached error rate may be difficult to achieve without a significantly increased effort, for example by using a dedicated sub-classifier to distinguish between text and table zones.

ACKNOWLEDGEMENTS

We wish to thank Oleg Nagaitsev for help with the implementation and Thomas Deselaers for making available the open source image retrieval system FIRE, which provided us with the implementation of some of the features used. This work was partially funded by the BMBF (German Federal Ministry of Education and Research), project IPeT (01 IW D03).

REFERENCES

Deselaers, T., Keysers, D., and Ney, H. (2004). Features for image retrieval: A quantitative comparison. In *DAGM 2004, Pattern Recognition, 26th DAGM Symposium*, volume 3175 of *Lecture Notes in Computer Science*, pages 228–236, Tübingen, Germany.

Guyon, I., Haralick, R. M., Hull, J. J., and Phillips, I. T. (1997). Data sets for OCR and document image understanding research. In Bunke, H. and Wang, P., editors, *Handbook of character recognition and document image analysis*, pages 779–799. World Scientific, Singapore.

Inglis, S. and Witten, I. (1995). Document zone classification using machine learning. In *Proc Digital Image Computing: Techniques and Applications*, pages 631–636, Brisbane, Australia.

Keysers, D., Och, F.-J., and Ney, H. (2002). Maximum entropy and Gaussian models for image object recognition. In *Pattern Recognition, 24th DAGM Symposium*, volume 2449 of *Lecture Notes in Computer Science*, pages 498–506, Zürich, Switzerland. Springer.

Kise, K., Sato, A., and Iwata, M. (1998). Segmentation of page images using the area Voronoi diagram. *Computer Vision and Image Understanding*, 70(3):370–382.

Liang, J., Phillips, I., Ha, J., and Haralick, R. (1996). Document zone classification using the sizes of connected components. In *Proc. SPIE*, volume 2660, Document Recognition III, pages 150–157, San Jose, CA.

Okun, O., Doermann, D., and Pietikainen, M. (1999). Page Segmentation and Zone Classification: The State of the Art. Technical Report LAMP-TR-036, CAR-TR-927, CS-TR-4079, University of Maryland, College Park.

Sivaramakrishnan, R., Phillips, I. T., Ha, J., Subramanium, S., and Haralick, R. M. (1995). Zone classification in a document using the method of feature vector generation. In *ICDAR '95: Proceedings of the Third International Conference on Document Analysis and Recognition (Volume 2)*, page 541ff.

Wang, Y., Haralick, R., and Phillips, I. (2000). Improvement of zone content classification by using background analysis. In *Fourth IAPR International Workshop on Document Analysis Systems (DAS2000)*.

Wang, Y., Phillips, I. T., and Haralick, R. M. (2006). Document zone content classification and its performance evaluation. *Pattern Recognition*, 39:57–73.

Wong, K. Y., Casey, R. G., and Wahl, F. M. (1982). Document analysis system. *IBM Journal of Research and Development*, 26(6):647–656.

caused by their distinction between the text classes of different font-sizes and the class 'other' with the remaining classes. On the other hand, we add a new class 'speckles', which is related to 0.15% (21/13811) error.