Example-Based Logical Labeling of Document Title Page Images

Joost van Beusekom[1], Daniel Keysers[2], Faisal Shafait[2], Thomas M. Breuel[1]

Image Understanding and Pattern Recognition (IUPR) Research Group
[1]Technical University of Kaiserslautern, Germany
[2]German Research Center for Artificial Intelligence (DFKI), Kaiserslautern, Germany
{joost, tmb}@iupr.net, {faisal.shafait, daniel.keysers}@dfki.de

Abstract

This paper presents a flexible and effective example-based approach for labeling title pages which can be used for automated extraction of bibliographic data. The labels of interest are "Title", "Author", "Abstract" and "Affiliation". The method takes a set of labeled document layouts and a single unlabeled document layout as input and finds the best matching layout in the set. The labels of this layout are used to label the new layout. The similarity measure for layouts combines structural layout similarity and textural similarity on the block-level. Experimental results yield accuracy rates from 94.8% *to* 99.6% *obtained on the publicly available MARG dataset. This shows that our lightweight method has equivalent and partially better performance when compared to other more complex labeling methods known from the literature.*

1 Introduction

Plain-text search in large databases of scientific papers is often a time consuming procedure. This process can be sped up by adding meta information as e.g. defining which part of the text contains the title and which one the abstract. This allows to focus the search on these fields. Extracting this information from document images is often done manually. This becomes more and more a problem as the databases of papers are growing, which makes the automated extraction of this meta-data a relevant problem.

Automated extraction of bibliographic information from paper title pages requires methods for automated recognition of the "function" of these blocks. This process is often referred to as "logical labeling". The problem to be solved is to assign correct labels to blocks, given an image and a segmentation. Let $L = \{l_1, ..., l_m\}$ be the set of all possible labels and $B = \{b_1, ..., b_n\}$ the blocks of the page. Then logical labeling consists of applying a suitable function $f : B \mapsto L : f(b) = l$, the logical labeling function that assigns a label to each block of the page.

For each document type a different label set is needed. For example, for business letters the labels "Sender", "Subject" and "Logo" may be used. Frequent labels on title pages of scientific papers are "Title", "Author", "Abstract" and "Affiliation", which we focus on in this paper.

Section 2 gives a short overview over related methods. Section 3 presents our approach for logical labeling. Evaluation and error measures are to be found in Section 4 and Section 5. Section 6 concludes this paper.

2 Related Work

Many different methods have been proposed for a broad field of labeling applications. Due to the lack of a general approach for solving the logical labeling problem for a broad class of documents types, many different approaches have been developed for many special classes of documents. A good overview can be found in the survey paper by Mao et al. [9]. The wide diversity of different methods, labels, and test sets makes it difficult to compare the approaches.

The subproblem of labeling title pages of scientific papers has been treated by several authors. Kim et al. [4] propose a rule based system using optical character recognition (OCR) to obtain plain text as well as document image analysis methods for extracting block features. They obtain an overall accuracy of 96.7% for labeling the four labels for three different layout types on the MARG[1] database. Mao et al. [8] extend this rule-based approach to be usable for a broader class of layout types, without the need of defining new rules for each new layout type. They evaluate their approach on a small part of the MARG dataset and claim to

This work has been partially supported by the Rhineland-Palatinate cluster of excellence "Dependable adaptive systems and mathematical modeling" and by the BMBF (German Federal Ministry of Education and Research), project IPeT (01 IW D03).

[1]http://marg.nlm.nih.gov/index2.asp

have better results than the initial system. In [10] Mao et al. present a labeling method based on Hidden Markov Models. These are used to represent the layout using projection profiles. Their approach was tested on 69 pages and compared to other models. Results are given in graphical form. The accuracy ranges between 70% for the baseline heuristic model and 91% for their Hidden Semi-Markov model.

Liang et al. [7] present a method capable of adaptively learning the layout type (or layout model). The block-based layouts are represented as graphs which express the relative position of the blocks to each other. The model graph is matched to the unlabeled graph. The result of the graph matching combined with the model labels are used for labeling the new graph. Unfortunately no quantitative evaluation of this interesting method is given. Instead of matching layouts using NP-hard graph matching, our approach reduces the problem to solving the assignment problem, which can be solved optimally in $O(n^3)$. Furthermore, the features and the distance measure differ.

Aiello and al. [1] present a complete document understanding system also capable of performing logical labeling. Features including aspect-ratio, font style and number of lines are used with a decision tree as classifier. Performance measure on the UW II database yielded up to 98% precision.

3 Logical Labeling by Example

Our method is an example-based approach to logical labeling. Given the segmentation of a new document image, represented by a set of blocks, it finds the best matching document in the database and then transfers the labels. The first step is to find the best match for a given document image and its segmentation. The second step consists of copying the labels from the best matching known document to the new unlabeled document. An illustration for the method can be found in Figure 1.

For the first step we use an enhancement of the layout distance measure presented in [14]. In our previous work we presented a block-based layout distance measure used for document image retrieval by layout dissimilarity. The distance measure takes a set of blocks (the new unknown layout, the query layout) and a set of sets of blocks as the known reference layouts. It then computes the best matching layout for the query layout and returns this document. To compute the best match, a distance measure for block sets representing layouts is used.

This distance measure is computed in two steps: given two layouts to be compared, L_1 and L_2. Each layout consists of a set of blocks $B_1 = \{\mathbf{b}_{11}, ..., \mathbf{b}_{1m}\}$ and $B_2 = \{\mathbf{b}_{21}, ..., \mathbf{b}_{2n}\}$. The distance between these two layouts is then computed as follows: first for every pair of blocks b_{1i} and b_{2j} the normalized overlapping area is computed

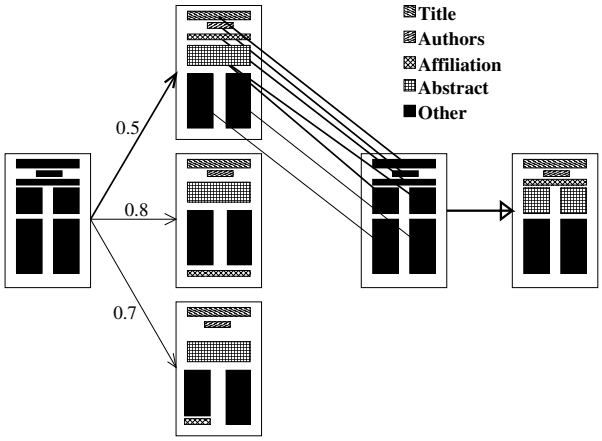

Figure 1. Illustration of the method: in the first step the distance between the unlabeled document and a dataset of labeled documents is computed using geometric and textural features. The best matching layout is taken as reference and then the labels are transferred using the matching information.

(0 stands for a perfect overlap, 1 for no overlap at all).

$$D_{\mathrm{ov}}(\mathbf{b}_{1i}, \mathbf{b}_{2j}) = 1 - \frac{2 \times \mathrm{Ov}(\mathbf{b}_{1i}, \mathbf{b}_{2j})}{\mathrm{area}(\mathbf{b}_{1i}) + \mathrm{area}(\mathbf{b}_{2i})}$$

where $\mathrm{Ov}(\mathbf{b}_i, \mathbf{b}_j)$ is the number of overlapping pixels of both blocks and $area(b)$ is the number of pixels of block b.

This results in an $m \times n$ cost matrix, which is used for the matching part. The matching is done by solving the minimum weight edge cover problem, which is a bipartite graph matching problem. It matches each block at least once while minimizing the total cost. The total cost is defined as the sum of the costs of all matched blocks.

Geometrical features do not contain enough information for this task. For logical labeling, texture information is needed. Therefore, textural features describing the content of a block have to be added. The distance between these feature vectors is combined with the overlapping distance to one new block distance. More details about the features used for this purpose are given in Section 3.1. The new block distance is then defined as follows:

$$D_{\mathrm{all}}(\mathbf{b}_{1i}, \mathbf{b}_{2j}) = D_{\mathrm{ov}}(\mathbf{b}_{1i}, \mathbf{b}_{2j}) \times D_{feat}(F_{1i}, F_{2j})$$

where $D_{feat}(F_i, F_j)$ is the distance between the two feature vectors F_{1i} and F_{2j} extracted from the blocks b_{1i} and b_{2j}. The matching of the blocks from layout L_1 to the ones from layout L_2 is again done by solving the minimum weight edge cover problem.

After having found the best matching document, the copying of the labels is straightforward: the minimum

weight edge cover returns the assignment matrix containing the information which blocks are matched to get the minimum distance. This information is used to copy the labels from the known layout, which are already labeled, to the blocks of the new layout. It may happen that a block gets more than one label assigned. Then some fall-back solution has to be found. One could be to take the label of the block with the minimum distance.

For evaluation purposes, only the assignments have been used: if the two blocks representing a two-columns abstract are matched against a one-column abstract, two correct matches are counted. If one block is assigned to the abstract and one to the author, one error and one correct assignment are counted. The motivation for this assignment-based evaluation is that the actual matching process is the one we want to analyze, and not the fall-back solution.

3.1 Features for Textural Descriptors

In the domain of image retrieval many features have been developed for textural descriptors. Many of them have found application in block classification systems. Keysers et al. have shown in [3] that the following features perform well on block classification.

- Run-length histograms: run-length histograms of black and white pixel sequences in four directions: horizontal, vertical, main diagonal (upper left to lower right corner) and side diagonal (lower left to upper right corner). Each histogram uses 8 bins, with exponentially increasing bin size: $\{1\}$, $\{2,3\}$, $\{4,...,7\}$, $\{8,...,15\}$, $\{16,...,31\}$, $\{32,...,63\}$, $\{64,...,127\}$, $\{128,...\}$. These values are then normalized by the total number of run-length counts in all direction to obtain values between 0 and 1. In total 64 run-length features are computed.

- Connected component sizes histograms: width and height of connected components are used to compute two 8-bin histograms. An 8×8 histogram is computed for the combination of width and height of the connected components. These values are also normalized by the total number of components in the block. In total 80 connected components features are used.

As a measure for similarity of block content is wanted, features performing well for block classification systems are the first candidates for this task.

3.2 Distance Metric for Feature Vectors

To compute the distance between two feature vectors, many different methods are available. As the features are represented by histograms, dissimilarity measures for histograms are used. Puzicha et al. show in [12] that the following two distance measures for histograms are a reasonable choice for textural feature histograms: Kullback-Leibler divergence [6] ("KLD") and Jensen-Shannon divergence [13] ("JSD"). Given two probability distributions P and Q of a discrete variable, the KLD is defined as follows:

$$D_{KLD}(P||Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}$$

In our case, P and Q are two feature histograms.

The JSD is defined as follows:

$$D_{JSD}(P||Q) = \frac{1}{2} D_{KLD}(P||M) + \frac{1}{2} D_{KLD}(Q||M)$$

where $M = \frac{1}{2}(P+Q)$.

4 Evaluation

Evaluation has been done on the MARG database (1553 images) and on the UW II database (624 images) which is a subset of UW III database [11]. The MARG database contains binarized, skew corrected images of title pages of medical journal articles and the corresponding ground-truth. UW II contains binarized, skew corrected images of scientific paper pages and the corresponding ground-truth. To obtain reproducible results, the ground truth blocks of the datasets where used. In a real world application, a standard page segmentation algorithm may be used.

For testing the proposed approach on the MARG dataset, the feature vector for each ground-truth block was computed. Our approach was tested using leaving-one-out cross validation. In this evaluation method each document image is once used as "unlabeled" document and the remaining document images are used as labeled dataset. To test the robustness of the method (how well does it adapt to layouts of which no exact example is present) and to avoid splitting the dataset arbitrarily into a test- and a training-set, the following two tests have been done: cross validation with leaving-journal-out: in this case all the document images form the same journal where removed from the dataset before finding the best match. Analogously, leaving-type-out cross validation removes all the document images of the same layout type from the dataset before finding the best match. Definitions of the layout-types can be found in [2].

For testing the performance on the UW II dataset in a way that is comparable to Aiello et al. [1] we first extracted the UW II images from the UW III dataset. These are all the images with filenames starting with "W0" and "W1". It is interesting to notice, that the UW III dataset does not contain the image "W0H1". So our dataset of UW II only contains 623 images instead of 624. We split the 623 images into a test set and a training set of approximately the same size: 311 images for the test-set and 312 images for

the training-set. To guarantee reproducibility, the test set consists of all images with even index ("W000", "W002", ..., "W1U9") and the training set of all images with odd index ("W001", "W003", ..., "W1UA"). Then exactly the same matching and labeling method is applied.

The accuracy is defined as the number of correct label assignments divided by the number of all assignments with that label.

5 Results

Testing the two different distance measures for feature vectors, the Jensen-Shannon-Divergence and the Kullback-Leibler-Divergence, showed that the Jensen-Shannon-Divergence gives better results. All results in this section are obtained using JSD.

The results for the leaving-n-out cross-validation on the MARG dataset can be found in Table 1.

For the leaving-one-out cross-validation the number of wrongly labeled blocks is quite low. The few errors are mostly due to mis-labelings of "Affiliation" and "Abstract" blocks. Visual inspection of the mis-labeled blocks showed, that these errors result mainly from the fact that some journals have single title pages in the database where no "Abstract" block is present. In consequence, "Abstract" blocks in the best matching document will be matched to another type of block and this produces errors. Furthermore, "Title" and "Author" are in some rare cases mis-labeled. This is mainly due to some special cases of articles from the same journal where subtitles are present. These subtitles are labeled as "Title" in the ground-truth. Size, position and texture of these subtitle blocks are very similar to "Author" blocks, so they are very likely to be mis-labeled.

In Table 1 the number of assignments gives the absolute number of assignments between blocks for the different leaving-n-out tests. The reason for these numbers to vary slightly from test to test is explained on an example: consider a document containing a one-column abstract, represented by one block, that is matched to a document containing a two-column abstract, represented by two blocks. This gives two correct assignments. If in the next test the first document is matched to a document having a one-column abstract (one block), it will return only one correct match. If one block is matched to the abstract, the other one to something else, one correct and one wrong assignment will be counted.

Table 1. Accuracy (in [%]) and number of assignments for the leaving-one-out, the leaving-journal-out and the leaving-type-out evaluation.

Label	lv-one-out		lv-jour-out		lv-type-out	
Title	99.6	1565	99.4	1565	99.2	1565
Author	99.9	1558	99.1	1557	95.7	1561
Affiliation	99.2	1578	97.8	1580	91.0	1607
Abstract	99.7	2069	99.0	2094	93.9	2198
Overall	99.6	6770	98.9	6796	94.8	6913

For the leaving-journal-out test, in mean 1547 labeled layouts remain in the dataset when all title pages from the same journal are removed before finding the best match. The main source of error for this test, apart from the problems stated above, results from mis-labelings between the two classes "Affiliation" and "Abstract". Visual inspection of the errors leads to the conclusion that in most of these cases, the best matching document does not fit to the given query layout and that in these cases wrong block assignments are more frequent. Another source of error are "Affiliation" and "Abstract" blocks which have quite often similar size and position and do not enormously differ in texture. This explains the relatively high number of mis-labelings between these two classes.

For the leaving-type-out evaluation, all documents of the same type are removed before searching the best match. There are in total 9 different layout types in the MARG database. They are called "typeA" to "typeH" and the ninth type is "typeO", standing for "other types". The distribution of the documents per type can be found in Table 3. There is a clear trend that "Affiliation" and "Abstract" mis-labelings increase when the quality of the best match decreases. The performance for "Title" remains nearly constant, which is due to the unique position of the title. It is almost always above all the other blocks.

Results for the UW II test are presented in Table 2. Accuracy for Aiello et al. is computed using the confusion matrix presented in [1]. The results show that both methods work in total similarly well, with a slight advantage for our method. Differences can be found in the performance of the different labels. A general problem when testing on the UW II dataset is that it contains a high number of duplicate images, differing only in some noise components. This makes it difficult to give an objective evaluation on this dataset. Although the results are not obtained on exactly the same data, due to probably different test- and training-sets, one can conclude that our method is also applicable to different label sets without the need of retraining (apart from giving some labeled layouts as examples) or reconfiguring any parameters.

Comparing the results of our approach to the results obtained by the rule-based system proposed by Kim et al. in [4] is difficult. Neither is the subset of documents used for their test defined in detail, nor the training-set on the basis of which their rules have been created. Furthermore, OCR data has not been used in our approach. Considering

Table 2. Accuracy and Error rates for the test on the UW II dataset (in [%]).

Label	Num. Assg.	Error	Acc.	Acc. [1]
title	34	14.7	85.3	90.6
text-body	2228	1.1	98.9	97.7
page-num	309	0.0	100.0	100.0
caption	272	11.0	89.0	86.8
Overall	2843	2.1	**97.9**	**96.8**

these differences, our lightweight approach works at least equivalently well with error rates between 0.4% and 5.2% depending on the test mode.

Quantitative comparison of our method to the one presented by Mao et al. in [8], which is based on the work of Kim et al. [4] is also not directly possible. On the one hand Mao et al. give performance measures for their whole system where labeling is only one subtask. On the other hand they use a relatively small subset of the MARG dataset (200 images) to evaluate their approach.

In general, our method has several advantages: as it is rule-free, no time-consuming adaption of the main system has to be done. To label new types of layouts at least one labeled example has to be provided. Then a standard segmentation algorithm e.g. Voronoi [5] can be taken to get segmentations of the unlabeled documents. The output of this segmentation then can be fed to the method which then will label the blocks. Another advantage of this approach is its flexibility concerning the labels: the labels are defined by the labeled examples that are presented to the system. So the system can also easily be adapted to other labeling tasks, e.g. business letters.

6 Conclusion

In this paper we presented a method allowing to do example-based logical labeling. We combined geometrical and textural block-based features to one block distance measure. This is then combined with the minimum weight edge cover matching to find the best matching labeled example document in the database, for a given unlabeled segmented document. Then the labels for the unlabeled blocks are set according to the labels of the blocks they have been matched to. Thorough evaluation has been done on the MARG database. Accuracy rates range from 94.8% to 99.6%, depending on the number and the quality of the remaining layouts in the dataset for the leaving-n-out cross validation test. A rough comparison of these results to previous works has been done. Furthermore the flexibility and accuracy of the method have been shown while comparing it to another labeling method on the UW II dataset.

Table 3. Distribution of layout types (typeA to typeH and typeO) in the MARG dataset in [%].

A	B	C	D	E	F	G	H	O
12.6	15.3	14.3	14.0	27.8	1.2	3.8	1.3	9.5

References

[1] M. Aiello, C. Monz, L. Todoran, and M. Worring. Document understanding for a broad class of documents. *IJDAR*, 5(1):1–16, 2002.

[2] G. Ford and D. Thoma. Ground truth data for document image analysis. In *Proc. of Symposium on Document Image Understanding and Technology*, pages 199–205, Apr. 2003.

[3] D. Keysers, F. Shafait, and T. M. Breuel. Document image zone classification - a simple high-performance approach. In *VISAPP 2007*, pages 44–51, Barcelona, Spain, Mar. 2007.

[4] J. Kim, D. X. Le, and G. R. Thoma. Automated labeling in document images. In *Proc. SPIE, Document Recognition and Retrieval VIII*, volume 4307, pages 111–122, Jan. 2001.

[5] K. Kise, A. Sato, and M. Iwata. Segmentation of page images using the area voronoi diagram. *Comput. Vis. Image Underst.*, 70(3):370–382, 1998.

[6] S. Kullback and R. A. Leibler. On information and sufficiency. *Ann. Math. Stat.*, 22:79–86, 1951.

[7] J. Liang and D. Doermann. Logical Labeling of Document Images Using Layout Graph Matching with Adaptive Learning. In *IAPR Conference on Document Analysis System*, pages 212–223, 2002.

[8] S. Mao, J. W. Kim, and G. R. Thoma. Style-independent document labeling: Design and performance evaluation. In *Document Recognition and Retrieval XI*, volume 5296, pages 14–22, Dec. 2003.

[9] S. Mao, A. Rosenfeld, and T. Kanungo. Document structure analysis algorithms: a literature survey. *Proc. SPIE Electronic Imaging*, 5010:197–207, Jan. 2003.

[10] S. Mao and G. R. Thoma. Bayesian Learning of 2D Document Layout Models for Automated Preservation Metadata Extraction. In *Proc. of the 4th IASTED Int. Conf. on Visualization, Imaging, and Image Processing (VIIP 2004)*, pages 329–334, Sept. 2004.

[11] I. T. Phillips. User's reference manual for the UW english/technical document image database III. Technical report, Seattle University, Washington, 1996.

[12] J. Puzicha, Y. Rubner, C. Tomasi, and J. Buhmann. Empirical evaluation of dissimilarity measures for color and texture. In *Int. Conf. on Computer Vision*, volume 2, pages 1165–1173, Corfu, Greece, Sept. 1999.

[13] C. E. Shannon. A mathematical theory of communication. *Bell Sys. Tech. Journal*, 27, 1948.

[14] J. van Beusekom, D. Keysers, F. Shafait, and T. M. Breuel. Distance measures for layout-based document image retrieval. In *Proc. of the 2nd Int. Conf. on Document Image Analysis for Libraries*, pages 232–242, Apr. 2006.