A Knowledge-Based Approach to the Layout Analysis

Floriana Esposito, Donato Malerba and Giovanni Semeraro

Dipartimento di Informatica - Università degli Studi - via Orabona, 4 - 70126 Bari, Italy
{esposito / malerbad / semeraro}@vm.csata.it

Abstract

In this paper, we present a hybrid approach to the problem of the document analysis in which the document image is segmented by means of a top-down technique and then basic blocks are grouped bottom-up in order to form complex layout components. In this latter process, called layout analysis, only generic knowledge on typesetting conventions is exploited. Such a knowledge is independent of the particular class of processed documents and turns out valuable for a wide range of documents. Preliminary results of the layout analysis system LEX (Layout EXpert) show the methodological validity of this approach.

1: Introduction

The problem of transforming data present on paper into a computer-revisable form is becoming more and more important, since it seems one of the main obstacles to the realization of challenging projects, such as building distributed digital libraries [7]. This transformation requires a solution to several problems, namely the separation of text from graphics (*document analysis*), the classification of the whole document (*document classification*), the identification of some relevant components of the page layout (*document understanding*) and the transformation of portions of the document bitmap image into sequences of characters.

All such document processing tasks are realized by PLRS (Page Layout Recognition System), a system mostly developed at the University of Bari [4]. In particular, the document classification and understanding phases exploit machine learning techniques in order to automatically tailor the system on the exigencies of different users. In this paper, we show how other AI techniques can be used in the preliminary phase of document analysis and, in particular, in the layout analysis process. Next Section is devoted to a short presentation of PLRS, while in Section 3 we explain why the layout analysis process should be considered an intelligent activity. Then, in Section 4 the main functions performed by LEX (Layout EXpert) are described, and in Section 5 some preliminary results are presented.

2: An overview of the system PLRS

The functions performed by different modules of PLRS and the intermediate results produced at each step are reported in Figure 1. Initially the document is scanned with a resolution of 300 dpi and thresholded into a binary image. Then, a possible skew is detected and corrected by analysing the horizontal projection profile [1]. Also, this analysis allows PLRS to estimate the complexity of the document as the ratio of the mean distance between peaks and the peak width on the horizontal histogram. The complexity ratio, which is greater than/lower than 1.0 for simple/complex documents, is used to evaluate one smoothing parameter, C_a, of the run length smoothing algorithm (RLSA) [17] applied in the segmentation phase. Indeed, the RLSA first performs a horizontal smearing of the page with a smoothing parameter C_h, then a vertical smearing with a smoothing factor C_v and finally an additional horizontal smoothing, guided by a parameter C_a, on the AND of the two bitmaps obtained in the two previous smearing steps. In PLRS, C_h is kept constant, while C_v is inversely proportional to the mean width of peaks in the horizontal histogram and C_a is directly proportional to the complexity factor. Therefore, documents with many rows, possibly organized in several columns, are more finely segmented, while documents with few spaced rows are subjected to a coarser segmentation.

In order to speed up the segmentation process, the RLSA does not operate on the original bitmap, but on a reduced document image with a resolution of 75 dpi.

The result of the segmentation process is a list of rectangular *basic blocks* enclosing either textual or nontextual content portions. In order to classify the blocks according to their content, a decision tree is used. There are only five classes associated with the leaves of the decision tree, namely text, horizontal line, vertical line, picture and graphics. The classification is based on ten numerical features computed for each block, namely height, length, area, eccentricity, total number of black pixels in the reduced bitmap, total number of black pixels in the segmented block, number of white-black transitions in the reduced bitmap, percentage of black

0-8186-7128-9/95 $4.00 © 1995 IEEE

Proceedings of the Third International Conference on Document Analysis and Recognition (ICDAR '95)
0-8186-7128-9/95 $10.00 © 1995 IEEE

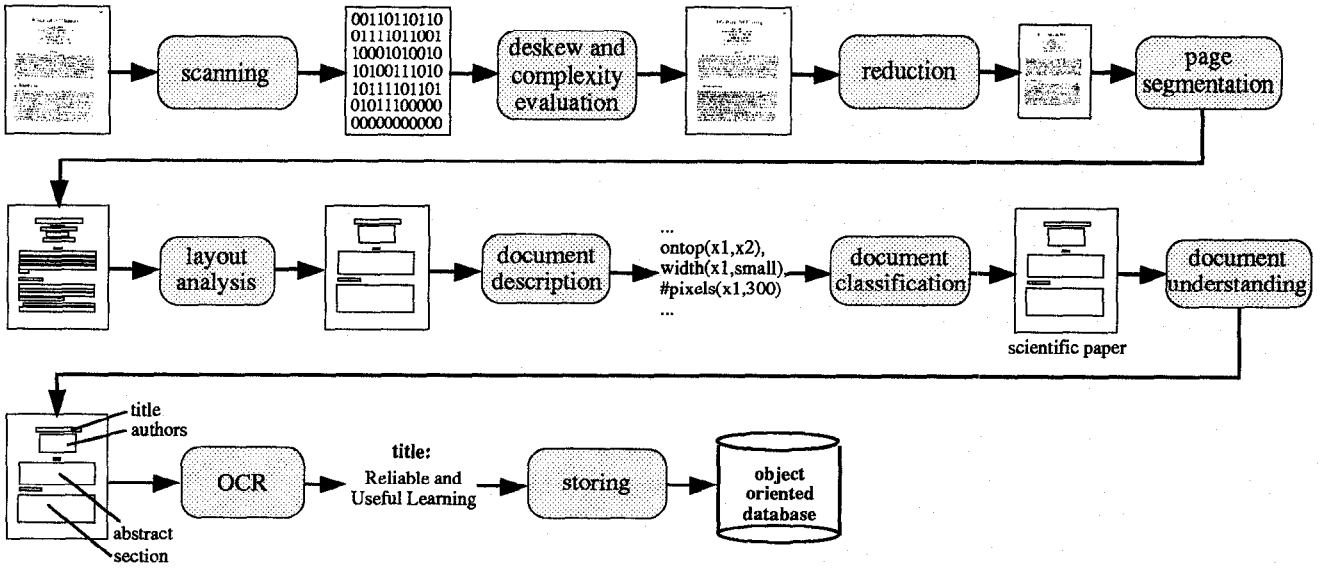


Figure 1. Blocks diagram of the Page Layout Recognition System (PLRS).

pixels in the reduced bitmap, percentage of black pixels in the segmented block, mean horizontal length of the black runs of the reduced bitmap. These features are almost the same used by Wong, Casey and Wahl [17] in their approach to text discrimination, but differently from them we use a decision tree instead of a linear pattern classifier. The decision tree has been induced from a set of 5473 examples of pre-classified blocks obtained from 53 documents.

The result of the segmentation process is an over-detailed description of the page layout. Actually, we do not need so much information for the subsequent phases of document classification and understanding. Therefore an intermediate step of *layout analysis*, which aims at grouping together blocks into composite layout components, is necessary. Ideally, the output of the layout analysis process should be a set of layout objects each of which can be associated with a distinct object of the logical structure, such as title and authors of a scientific paper. In practice, however, a sub-optimal layout structure in which it is still possible to distinguish different logical components should be considered a good output of a layout analyser. This is even more sensible when the document analysis is separated from the document classification and understanding processes, as in PLRS.

When the user is able to define a set of classes of documents that are relevant for the particular application domain, then it is possible to train the system to recognize the membership class of a document on the ground of a symbolic description of the page layout. The induction of recognition rules is based on machine learning techniques.

A similar approach has also been adopted for the document understanding phase [3]. In this case, layout-logical relationships between one or more elements of the layout hierarchy and one element of the logical hierarchy are exploited in order to identify some logical components of a document without reading its content by means of an OCR. Once the document has been understood, the OCR is applied to only those logical components deemed useful for storing and retrieval purposes.

3: Knowledge-based layout analysis

As already pointed out, in PLRS the high level processes of document classification and understanding are based only on geometrical characteristics of the page layout. This means that the detection of a "good" page layout is crucial for the success of the application. In fact, when the page layout groups together too many blocks, it may become hard to understand the document, since one layout component may correspond to more than one logical component. On the contrary, an extreme fragmentation of the page layout makes the subsequent learning process much slower because of the over-detailed description of the training documents.

Many approaches have been proposed for the extraction of the layout structure from the digital image. They have been classified as *top-down* or *bottom-up* [15]. In top-down methods, the page is repeatedly split into smaller and smaller blocks. Typical examples are the RLSA and the projection profile cuts [11]. In bottom-up methods, basic layout components are extracted from the bitmap and then grouped together into larger blocks on the ground of their characteristics. Neighbourhood line density [9] and connected component analysis [6] are two examples. *Hybrid* approaches are also known; for instance, the document analysis process performed by PLRS combines a top-down technique for

Proceedings of the Third International Conference on Document Analysis and Recognition (ICDAR '95)
0-8186-7128-9/95 $10.00 © 1995 IEEE

segmenting a page image and a bottom-up layout analysis method for assembling basic blocks into larger and larger layout components (*frames*).

Nevertheless, a different dimension of classification of document analysis systems is given by the *amount of explicit knowledge* they are provided with. Here, by knowledge we intend the body of facts and principles used by humans when they are asked to analyse a document. In this sense, we can say that the RLSA or the projection profile cuts need a limited amount of knowledge expressed in terms of few numerical parameters. This view differs from that given by Tang, Yan and Suen [16], according to which top-down is considered a synonym of knowledge-based while bottom-up stands for data-driven. The processes of either repeatedly subdividing a document image into smaller parts or grouping together blocks into larger layout components can be *both* accomplished with either little or much knowledge.

It is possible to provide several examples of both top-down and bottom-up approaches based on an extensive use of knowledge. For instance, in Gobbledoc [13], a document is segmented and labelled by using a publication-specific grammar, which describes all legal page formats allowed for a given publication. When a parse of the page is possible, logical labels are assigned to the various layout components. Dengel and Barth [2] define a hierarchical document layout model, called *geometric tree*, which contains knowledge about different possible document layouts. The internal nodes of the geometric tree represent different document classes, while the leaves correspond to possible final interpretations. Therefore, the problem of analysing and understanding a document is cast as a search through the geometric tree for the most plausible interpretation. Another method is exploits a Form Definition Language (FDL) in order to express models of layouts [8]. In this case, a Form Dividing Engine (FDE) parses a document image by means of the layout rules and produces the layout structure. In all these examples of top-down approaches, declarative knowledge on the possible layouts, expressed as grammar rules or geometric trees or FDL rules, is always required to segment a document image.

It is even less difficult to find examples of bottom-up approaches which are strongly dependent on a knowledge-base. For instance, in the document spectrum or *docstrum* method [14], the structural block determination is based on some criteria which can be expressed as rules of a production system. Rules are also used by Fisher, Hinds and D'Amato [5] in their bottom-up approach to document image segmentation.

Nagy, Kanai and Krishnamoorthy [12] distinguish three levels of knowledge on the layout structure of a document:

- *Generic* knowledge (e.g. type base lines in a word are collinear).
- *Class-specific* knowledge (e.g. no text line lateral to a graphical object).
- *Publication-specific* knowledge (e.g. maximum type size is 22 points).

Moreover, they observe that the knowledge base for bottom-up processing is necessarily different from that used for top-down processing: it is much less document specific. We agree with them and we observe that in general, knowledge used in top-down approaches is derived from the relations between the geometric and logical structures of specific classes of documents. Indeed, page grammars and geometric trees are used in order to segment the document image and simultaneously label some layout components with logical classes. However, hand-coding publication-specific knowledge is a demanding task: Dengel and Barth examined about two-hundred different business letters to define a geometric tree.

As already pointed out, in PLRS document-specific knowledge for classification and understanding is automatically learned from examples of documents. Therefore, this problem is overcome. Nevertheless, it is necessary to exploit much knowledge in the layout analysis process. Such a knowledge is independent of the particular class of processed documents, thus it turns out valuable for a wide range of documents.

LEX is the knowledge-based system devoted to the layout analysis process. Given the result of the segmentation process, it exploits generic knowledge and rules on typesetting conventions in order to group basic blocks together into frames. LEX has been implemented in Prolog because of the simplicity of expressing and manipulating declarative knowledge in logic programming [10].

4: The Layout EXpert

The two main functions performed by LEX are:

- A *global analysis* of the document in order to determine possible areas containing paragraphs, sections, columns, figures and tables.
- A *local analysis* of the document aiming at grouping together blocks which possibly fall within the same area.

Initially, LEX acquires information on the basic blocks detected in the page layout by the RLSA. In particular, for each block the following information is read:

- A progressive index distinguishing various blocks.
- The coordinates of the top-left and bottom-right hand corner of the block.
- The *type* of block: text, horizontal line, picture, vertical line, and graphic.
- The total number of black pixels in the reduced image.
- The total number of black pixels in the reduced image after the smoothing process accomplished by the RLSA.
- The number of horizontal white-black transitions in the reduced image.

LEX first performs a global analysis and then a local analysis.

Proceedings of the Third International Conference on Document Analysis and Recognition (ICDAR '95)
0-8186-7128-9/95 $10.00 © 1995 IEEE

4.1 Global analysis

In the phase of global analysis of a document, properties shared by a number of layout components are identified with the aim at determining the characteristics concerning the general structure of the document. For instance, all the text blocks having almost the same length and indentation are taken into account in order to detect the presence of columns.

The main criteria considered in this phase aim at finding several types of areas, such as columns, sections/paragraphs, and figures.

The steps of column and section/paragraph detection are tightly correlated to each other. Indeed, once a column has been detected and sections and paragraphs in it have been identified, LEX analyses each area found in the previous step and evaluates the possibility of further splitting it into columns. This recursive process stops when either the size of the area is too small (lower than 1/56 of the total size of the document) or no further sub-area can be detected.

The step of column identification relies on the analysis of the vertical histograms computed on the basic blocks of type text. A column is made up of at least two successive bars in the histogram that lay between the right border of the previous column and the first point with a zero height bar.

When columns are detected, the horizontal histograms for all columns are computed in order to detect possible sections/paragraphs inside each column. In particular, all spacings between text blocks are calculated, and the most frequent of them is selected. Such a spacing is very useful in formatted documents, where most of the lines in a paragraph/section are equally spaced. Therefore, a cut point for a column can be established whenever the distance between two bars in the horizontal histogram is significantly greater than the most frequent spacing. The coordinates of the first and last bar between two successive cut points determine an area inside the column. Each area is subsequently analysed in order to detect possible internal columns according to the criteria explained above.

When all possible areas for the blocks of type text have been found, the system focuses its attention on the graphic areas which contain graphic blocks. A search for such areas is performed within those zones of the document which are complementary to the text areas found in the previous step. If one of such zones contains blocks of either type picture or type graphic that cover more than 50% of the whole zone, then it is considered a graphic area.

4.2 Local analysis

In this phase, LEX analyses common properties of two or more layout components in order to establish whether the following criteria are satisfied:

- *Proximity*: adjacent components belonging to the same column/area and equally spaced.
- *Continuity*: overlapping components.
- *Similarity*: components of the same type, with an almost equal height.

Pairs of layout components that satisfy some of these properties may be grouped together.

All these criteria are followed in the very first step of the local analysis, when basic blocks of type text are grouped together into text lines. This is a necessary step of layout analysis, since the high complexity of the document may force the segmentation process to adopt small values for the smoothing factors C_v and C_h, and consequently to over-segment the reduced bitmap. This means that each line is split into several basic blocks, or even that some words are split into characters. In order to group together all words or characters in a line, LEX determines the list of all text blocks in an area and for each of them checks whether there is another block in the list that satisfies an alignment condition. In particular, two blocks are merged if their projections on the vertical axis overlap. This rule reflects the convention that type base lines in a word are collinear. In this first step, fragments of horizontal/vertical lines are also grouped together according to the continuity, proximity and similarity criteria. Specifically, if two horizontal (vertical) lines are not too far along the vertical (horizontal) axis and are almost aligned, then they are merged together.

In the second step of local analysis, only two criteria are adopted: similarity and proximity. Among all the lines detected in the previous step, those belonging to the same area determine a set of lines. This step of local analysis aims at grouping together lines of the same type, provided that their projections on the horizontal axis match within a tolerance threshold. Such a condition prevents LEX from merging lines that represent distinct layout objects, such as a line centred with respect to a set of lines with a different alignment, since it might represent the title of a section, or a line which presents an indentation on the left/right margin, since it might be the beginning/ending line of either a section or a paragraph. It is worthwhile to note that this process is extended to any type of line, but always concerns lines of the same type. Thus, lines of different types are never assembled together, even though alignment conditions are met. After having detected the set of text lines, LEX groups together the graphic/picture blocks falling inside a graphic area detected in the local analysis process. The only criterion adopted in this phase is that of proximity.

Results of this step are taken into account by the third phase in order to determine the first frame level. With this purpose, it is useful to consider the columns detected by the process of global analysis. Heuristics exploited in this phase are inspired to the criteria of proximity and continuity. In fact, two sets of lines are merged if their projections on the horizontal axis satisfy some alignment conditions and their distance on the vertical axis is lower than a fixed threshold.

Proceedings of the Third International Conference on Document Analysis and Recognition (ICDAR '95)
0-8186-7128-9/95 $10.00 © 1995 IEEE

The introduction of such a *soft* condition allows LEX to group together sets of lines that were considered distinct by the previous step because of a different alignment. Furthermore, the type of alignment between sets of lines plays a relevant role in this step. For instance, lines representing the beginning (end) of a section/paragraph are grouped to the rest of the section/paragraph because they show an alignment on the right (left) margin and their vertical distance is lower than the fixed threshold. It is interesting to observe the behaviour of this third step of local analysis on a type of documents in which first level frames are intermixed by horizontal lines, such as journal indexes. In that case, LEX avoids merging two frames separated by a horizontal line.

The last step of the local analysis aims at determining the second frame level. As in the previous step, also in this case it is useful to consider the columns detected by the global analysis. In addition, the first level frames are taken into account. Heuristics implemented in this step rely upon criteria of proximity and continuity. For each first level frame occurring into the column that is currently under analysis, the mean and the standard deviation of the distance between two consecutive lines in the frame are computed. Then, a first level frame is grouped with an adjacent first level frame in the column if the following conditions are satisfied: their projections on the horizontal axis overlap, there exists either a left or a right margin alignment, and their distance is lower than the sum of the mean and the standard deviation computed before. Again, the existence of a horizontal line between two frames inhibits the process of merging these two areas.

5: Results

LEX has been tested on several single page documents with different layouts, such as business letters, cover pages sent by fax, business letters received by fax, indexes of several journals or magazines as well as pages of papers appeared on journals and proceedings. In all, we created a base of more than two hundred document images in TIFF format. All images are deskewed, even those documents received by fax for which the rotation of the whole image caused some problems in the heading line (this is a case in which it is necessary to deskew only a part of the image and not all).

We illustrate the performance of the layout analysis process in Figure 2. In this case LEX is able to detect columns inside the table at the top-left hand corner, so that all numbers reported in each column are assembled into one frame. Moreover, by grouping together fragments of the horizontal solid lines it is possible to reconstruct part of the lines lost during the processes of reduction and segmentation. LEX, however, does not detect a distinct column for each sub-title, thus it improperly groups together the text of various sub-titles into three distinct lines. By looking at the basic blocks reported in Figure 2b, it is evident that lines have been segmented into word-sized blocks, which are correctly assembled in the first step of the local analysis. The only problem occurs with the number of page that is grouped together the first part of the running head. In fact, this latter falls partly into the left column and partly into the right column, thus in the first step of the layout analysis not all the basic layout objects composing the running head will be clustered. Figure 2c shows that all text lines equally spaced and aligned by right and left columns are assembled together. Finally, in Figure 2d we see the result of the subsequent step in which LEX takes into account the possible indentation at the beginning of paragraphs as well as the incomplete lines at the end of the paragraphs. The result of this step of the layout analysis is a set of first level frames which coincides with the set of second level frames, for this particular document.

6: Conclusions

In this paper, a Prolog system for the layout analysis of single page documents has been presented. Its novelty is the explicit use of typesetting knowledge in order to group basic blocks produced by the segmentation process. We observed, indeed, that humans are able to accomplish this task very well, even when they cannot read the content of the blocks or do not know which kind of document they are analysing.

The distinct tasks performed by LEX have been briefly described. The most important are:

- The global analysis aiming at the detection of text and graphic areas.
- The local analysis aiming at merging layout components inside each area.

As future work, we plan to extend the local analysis with knowledge on other typesetting conventions embedded in the styles of many desktop publishers.

Acknowledgments

Thanks to Ernesto Bellisari and Vincenzo Bonetti for their precious collaboration on conducting the experiments.

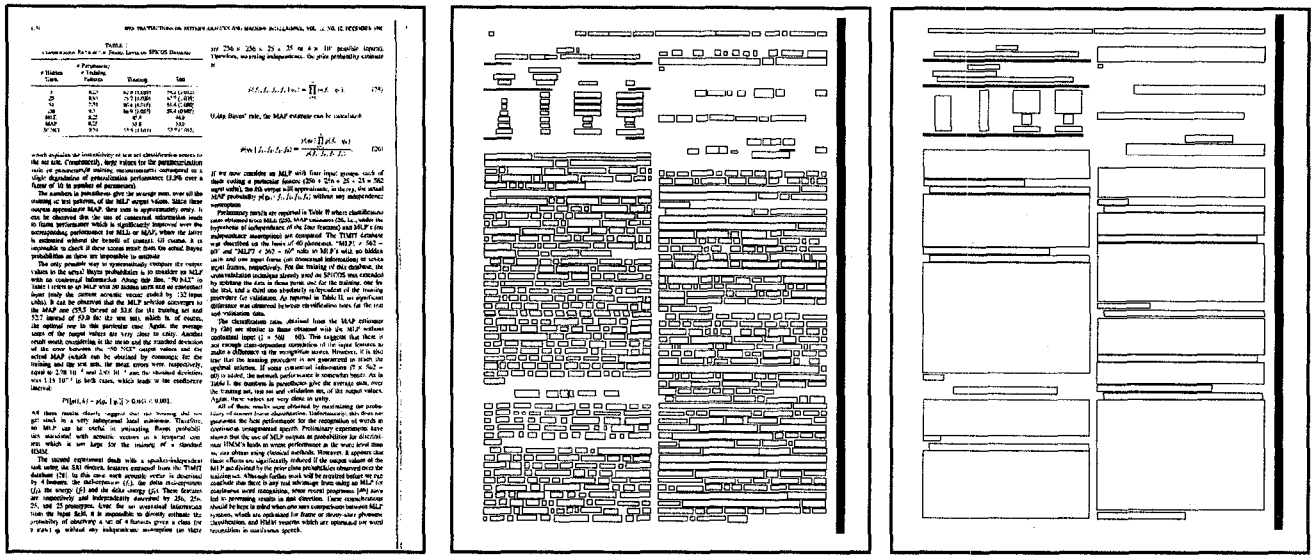
References

[1] Ciardiello, G., G. Scafuro, M. T. Degrandi, M. R. Spada, and M. P. Roccotelli (1988). An experimental system for office document handling and text recognition. *Proc. 9th Int. Conf. on Pattern Recognition*, 739-743. Los Alamitos: IEEE Computer Society.

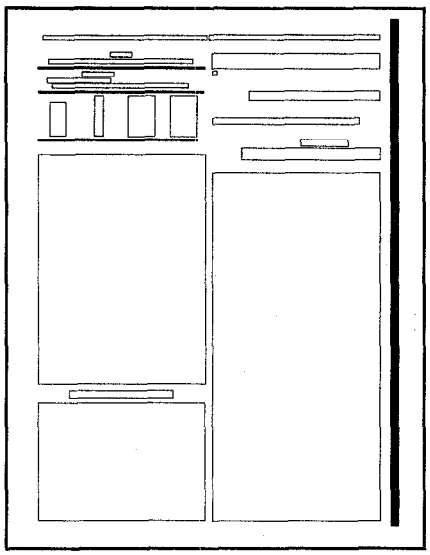
[2] Dengel, A., and G. Barth (1988). High level document analysis guided by geometric aspects. *Int. J. Pattern Recognition and Artificial Intelligence*, 2:641-655.

[3] Esposito, F., Malerba, D., and G. Semeraro (1993). Automated acquisition of rules for document understanding. *Proc. 2nd*

Proceedings of the Third International Conference on Document Analysis and Recognition (ICDAR '95)
0-8186-7128-9/95 $10.00 © 1995 IEEE



a) b) c)



d)

Figure 2. a) Original document. b) Result of the segmentation process. c) Set of lines detected in the document. d) First level frames detected by LEX.

Int. Conf. on Document Analysis and Recognition, 650-654: Los Alamitos: IEEE Computer Society.

[4] Esposito, F., Malerba, D., and G. Semeraro (1994). Multistrategy learning for document recognition. *Applied Artificial Intelligence*, 8:33-84.

[5] Fisher J.L., Hinds, S.C., and D.P. D'Amato (1990). A rule-based system for document image segmentation. *Proc. 10th Int. Conf. on Pattern Recognition*, 567-572. Los Alamitos: IEEE Computer Society.

[6] Fletcher, L.A., and R. Kasturi (1988). A robust algorithm for text string separation from mixed text/graphics images. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, PAMI-10(6):910-918.

[7] Fox, E.A. (1994). How to make intelligent digital libraries. In Z.W. Ras and M. Zemankova (Eds.), *Methodologies for intelligent systems*, 27-38, LNAI 863, Berlin:Springer-Verlag.

[8] Higashino, J., Fujisawa, H., Nakano, Y., and M. Ejiri (1986). A knowledge-based segmentation method for document understanding. *Proc. 8th Int. Conf. on Pattern Recognition*, 745-748. Los Alamitos: IEEE Computer Society.

[9] Kubota, K., Iwaki, O., and H. Arakawa (1984). Document understanding system. *Proc. 7th Int. Conf. on Pattern Recognition*, 612-614. Los Alamitos: IEEE Computer Society.

[10] Malerba, D., Semeraro, G., and E. Bellisari (1995). LEX: A knowledge-based system for the layout analysis. *Proc. 3rd Int. Conf. on the Practical Application of Prolog*, 429-443, Paris, France.

[11] Nagy, G., Seth, S.C., and S.D. Stoddard (1986). Document analysis with an expert system. In E.S. Gelsema and L.N. Kanal (Eds.), *Pattern recognition in practice II*. North Holland: Elsevier Science Publishers.

[12] Nagy, G., Kanai, J., and M. Krishnamoorthy (1988). Two complementary techniques for digitized document analysis. *ACM Conference on Document Processing Systems*, Santa Fe, N. Mexico.

[13] Nagy, G., Seth, S.C., and S.D. Stoddard (1992). A prototype document image analysis system for technical journals. *IEEE Computer*, 25(7):10-22.

[14] O'Gorman, L. (1993). The Document Spectrum for Page Layout Analysis. *IEEE Trans. on Pattern Analysis and Machine Intelligence*. PAMI-15(11):1162-1173.

[15] Srihari, S.N., and G.W. Zack(1986). Document image analysis. *Proc. of the 8th Int. Conf. on Pattern Recognition*, 434-436. Los Alamitos: IEEE Computer Society.

[16] Tang, Y.Y., Yan, C.D., and C.Y. Suen (1994). Document processing for automatic knowledge acquisition. *IEEE Trans. on Knowledge and Data Engineering*, 6(1):3-21.

[17] Wong, K. Y., Casey, R.G., and F.M. Wahl (1982). Document analysis system. *IBM J. Res. Develop.*, 26(6):647-656.

Proceedings of the Third International Conference on Document Analysis and Recognition (ICDAR '95)
0-8186-7128-9/95 $10.00 © 1995 IEEE