Extraction of Text Areas in Printed Document Images

Jean Duong[1,2], Myriam Côté[1], Hubert Emptoz[2], Ching Y. Suen[3]

[1]Laboratoire d'Imagerie, Vision et Intelligence Artificielle (LIVIA)
École de Technologie Supérieure (ETS)
1100, rue Notre Dame Ouest, H3C 1K3 Montreal, Quebec, Canada

[2]Laboratoire de Reconnaissance des Formes et Vision (RFV)
Institut National des Sciences Appliquées (INSA) de Lyon
20, avenue Albert Einstein, 69621 Villeurbanne Cedex, France

[3] Center for Pattern Recognition and Machine Intelligence (CENPARMI)
Suite GM-606, Concordia University
1455 de Maisonneuve Blvd. West, H3G 1M8 Montréal, Quebec, Canada

e-mail: duong@livia.etsmtl.ca, cote@gpa.etsmtl.ca, emptoz@rfv.insa-lyon.fr
phone: (1-514) 396-8800 + 7674

Abstract

In this paper, we present a document analysis system which is expected to extract regions of interest in greyscale document images. Collected areas are then clustered in text zones and non-text areas using geometric and texture features. The system works in two steps. Regions of interest are retrieved via cumulative gradient considerations. In classification module, we introduced some entropic heuristic. Experiments are done on the MediaTeam Document Database to show the relevance of this criteria.

Introduction

The word document comes from latin *documentum* which itself is derived from *docere*, that is "teaching". According to the definition found in dictionaries, a document can be viewed as a "proof", an "evidence". More precisely, a document may be "an original or official paper relied on as the basis, proof, or support of something", "something (as a photograph or a recording) that serves as evidence or proof", "a writing conveying information", "a material substance (as a coin or stone) having on it a representation of thoughts by means of some conventional mark or symbol" [1]. For the purpose of this article, we focused on paper documents.

In spite of the wide spread use of computers and other digital facilities, paper document keeps occupying a central place in our everyday life. Conversely to what was expected, the amount of paper produced presently is larger than ever. Important institutions like administrations, libraries, archive services, etc. are heavy paper producers and consumers. To some point of view, paper is one of the most reliable information supports. Unlike numerical records, it is not constrained by format compatibility question, or device needs.

On the other side, document storage for safety or accessibility considerations is a very tricky problem. Research is presently done in such a direction. To digitize a paper document, one must retrieve an image (for instance, via a scanner) and work on it. In this paper, we are interested in the preliminary step of document analysis and recognition. Our present goal is to extract textual areas.

We briefly present previous research in the physical segmentation a document and text detection (in section 1). Then we describe our work (in section 2) which mainly consists of detection of zones of interest (in sections 2.1 and 2.2) and discrimination of text areas (in section 2.3).

1 Previous works

Text detection in document images is part of document analysis and page segmentation. Traditionaly, page segmentation methods are divided in three groups: top-down, bottom-up and hybrid approaches [6, 5, 23, 17, 18].

In top-down techniques, documents are recursively divided from entire images to smaller regions. Theses techniques are often fast, but the efficiency depends on a priori knowledge about the class of documents to be processed. Developments have been produced in early times. The most well known are projection methods [16, 14, 15] (with many variations, like white streams method [21], rectangulation [8], etc.), histogram analysis, form definition languages [29], rule based systems [30, 15], or space transforms (Fourier transform, Hough transform, etc. [31, 19, 12, 7, 6, 20, 22, 11]).

Bottom-up methods start with the tiniest elements (pixels), merging them recursively in connected components or regions, and then in larger strutures. Most popular bottom-up techniques are mathematical morphology [25, 26, 2], run length smoothing algorithm [6, 32, 9], and region growing-based methods [31, 12, 3, 35, 20].

2 Presentation of our system

Our purpose is to retrieve text information in printed document images. Therefore we have chosen to work with greyscale images from the MediaTeam document database.

2.1 Areas detection

We assume that text elements are written in Latin alphanumeric symbols and set in horizontal lines. According to this hypothesis, regions with important gray level variations in horizontal direction are likely to be text areas, as stated by F. Lebourgeois and S. Souafi ([27, 28]).

Let us consider a given document image as a matrix M. A gradient image G is computed by applying the 2×1 mask $(-1, 1)$ to M([13]). A third matrix S is obtained from G by the relation

$$S(i,j) = \sum_{k=j-\frac{w}{2}}^{k=j+\frac{w}{2}} |G(i,k)|$$

As we said, text elements produce high horizontal gradient values. A value histogram H is computed for S. To find possible text regions, we have to determine a threshold θ. As shown in Fig. 1, H is varies considerably, depending on the class of document. Thus, it sems difficult to process it to get θ.

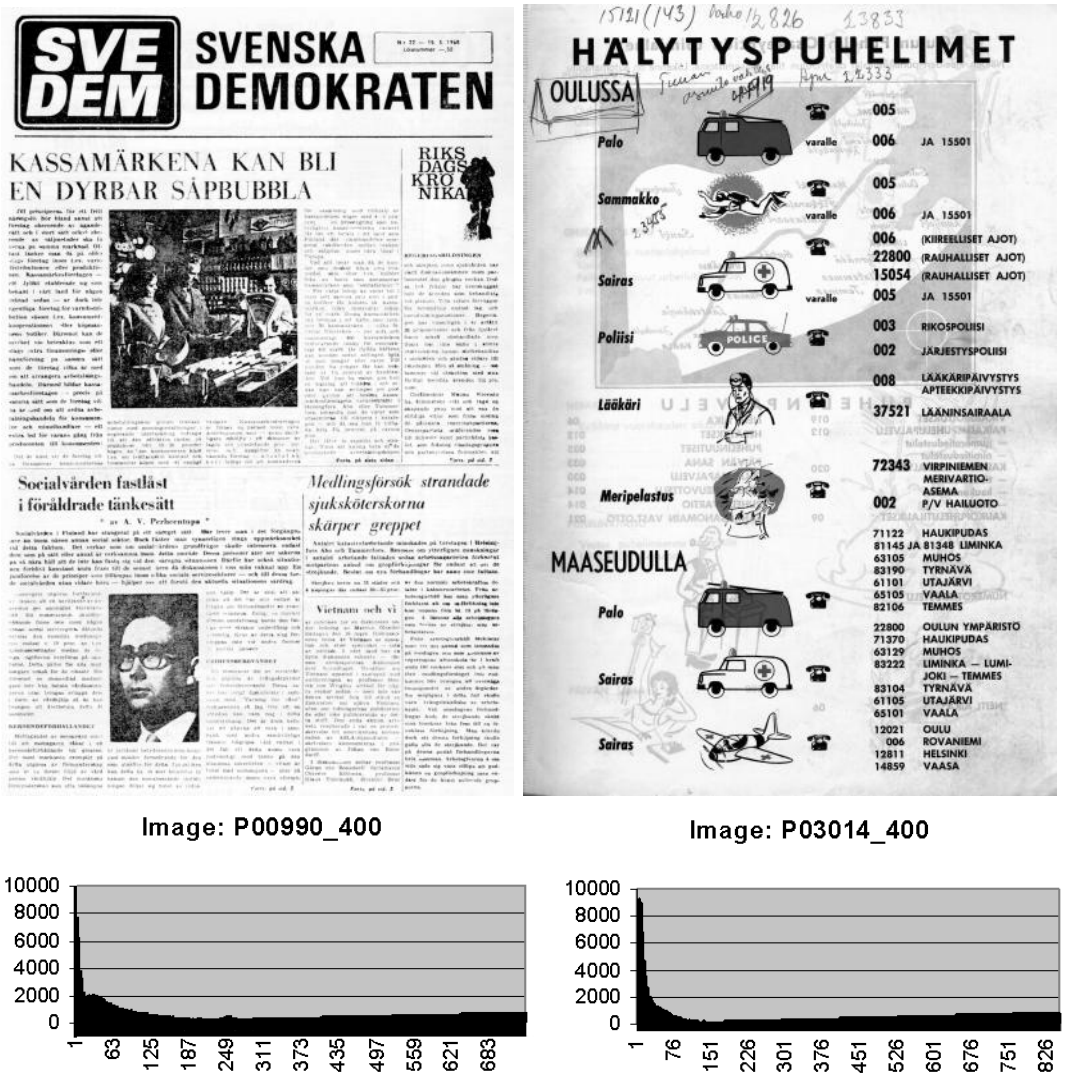

Figure 1: Examples of cumulative gradient histograms for two different document images from MediaTeam Document Database. P00990 comes from the class ARTICLE, and P03014 from PHONEBOOK

We choose to separate cumulative gradient values into two classes using a k-means procedure. A binary image is built. Pixels with high (resp. low) values for cumulative gradient are set to black (resp. white). A connected components search is then performed.

As expected, with suitable values of w, potential text regions are detected (example in Fig. 2). But one may quote that image areas are sometimes cut, mostly when containing huge homogenous parts (2.3).

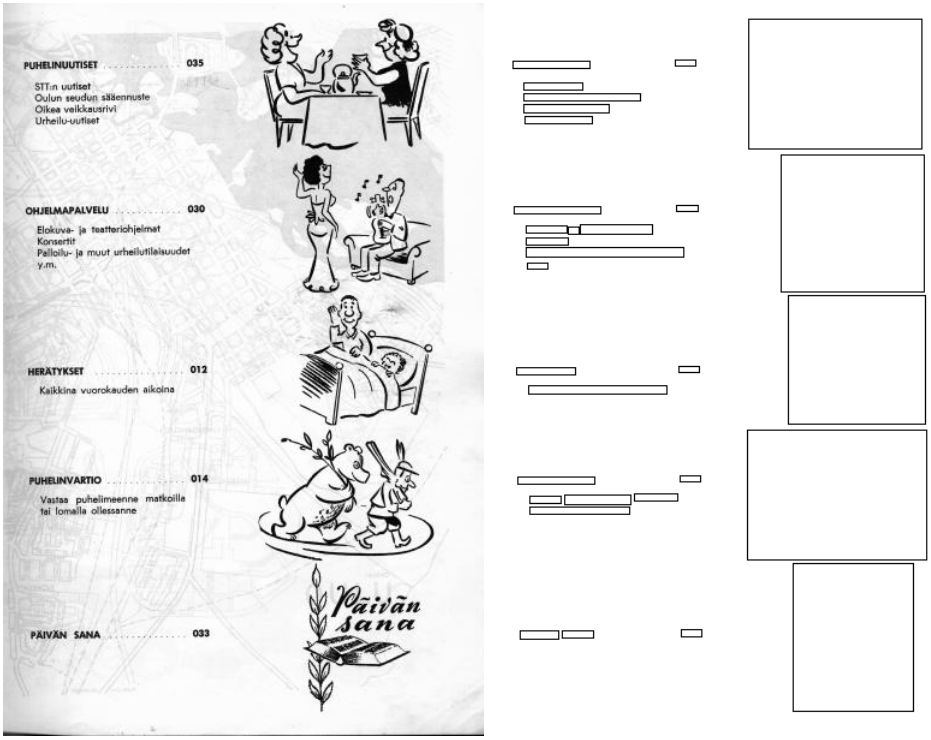

Figure 2: Example of regions of interest extracted using the cumulative horizontal gradient method with mask width $w = \frac{W}{50}$. (W is the width of the image.)

2.2 Binarization

Document images have to be binarized for further processing. Instead of performing a global thresholding for the whole image, we prefer to consider every chip separately, as Manmatha et al. did ([33, 34]).

For each area found in the previous step, we compute the intensity histogram. A binarization threshold is calculated via a simple k-means process, separating high and low gray level values.

2.3 Text separation

To separate text from non-text elements, we cluster the set of areas found via gradient discrimination. We use geometric features combined with some texture primitives.

As we considered horizontal cumulative gradient to detect regions of interest, pieces of text lines are expected to be retrieved. Thus, we keep regions as potential text areas or we exclude them according to their height. Average height μ_{height} is computed over the whole set of regions of interest. Chips taller than $3 \times \mu_{height}$ are classified non text regions. (For most of the printed documents, page setup satisfies this constraint.)

For each possible text chip C, a horizontal projection histogram $H_{P(C)}$ can be designed after binarization. Once this is done, entropy E_C for $H_{P(C)}$ is computed as following:

$$E_C = -\sum_{i \in I_{P(C)}} p_i ln\,(p_i)$$

where $I_P(C)$ is the set of data index for H_P and $p_i = \frac{H_{P(C)}[i]}{\sum_{k \in I} H_{P(C)}[k]}$. We can use this value as a criteria to classify C as a text or a non text region. Projection histogram entropy has been used previously in many occasions, as for example in handwritten document processing to detect word baselines direction ([10]). For us, text elements are supposed to be gathered in horizontal lines. Then, by definition of $H_{P(C)}$ and E_C,text areas should have lower entropy values than common graphic chips.

3 Experiments

We worked on the MediaTeam document database. From here, regions of interest, text chips, etc. are approximated by the rectangular areas defined by their bounding boxes. We define the performance of our system for one document image as the ratio

$$\frac{|t|}{|T|}$$

with

- T the set of regions of interest found by our system and included in text areas according to the database information. If T is void, the image is not considered for the compilation of statistics.
- t the set of regions of interest found by our system, labelled as text chips, and included in text areas according to the database information.

We have tested our system on the entire database with mask width parameter $w = \frac{W}{50}$ for cumulative gradient computations where W is the width of document images. This represents approximately the width of two printed characters for a text with font type roman and size 12pt. A too narrow mask could cause regions of interest to be splitted in numerous small chips, and a too wide one could result in irrelevant large composite areas. In a first set of experiments, the heuristic based on entropy is used exclusively. In practice, we compute entropy values E_C for the set of all regions of interest. This enable us to estimate a mean value $\mu_{entropy}$ and standard deviation $\sigma_{entropy}$. If one of the E_C is greater than $\mu_{entropy} + 3 \times \sigma_{entropy}$, we perform an automatic clustering of E_Cs set in two subsets for high and low values respectively. Chips with entropy in high values subset are eliminated. Results are reported in Fig. 4. Only relevant samples (images for which at least one region of interest is found in a text zone) are counted.

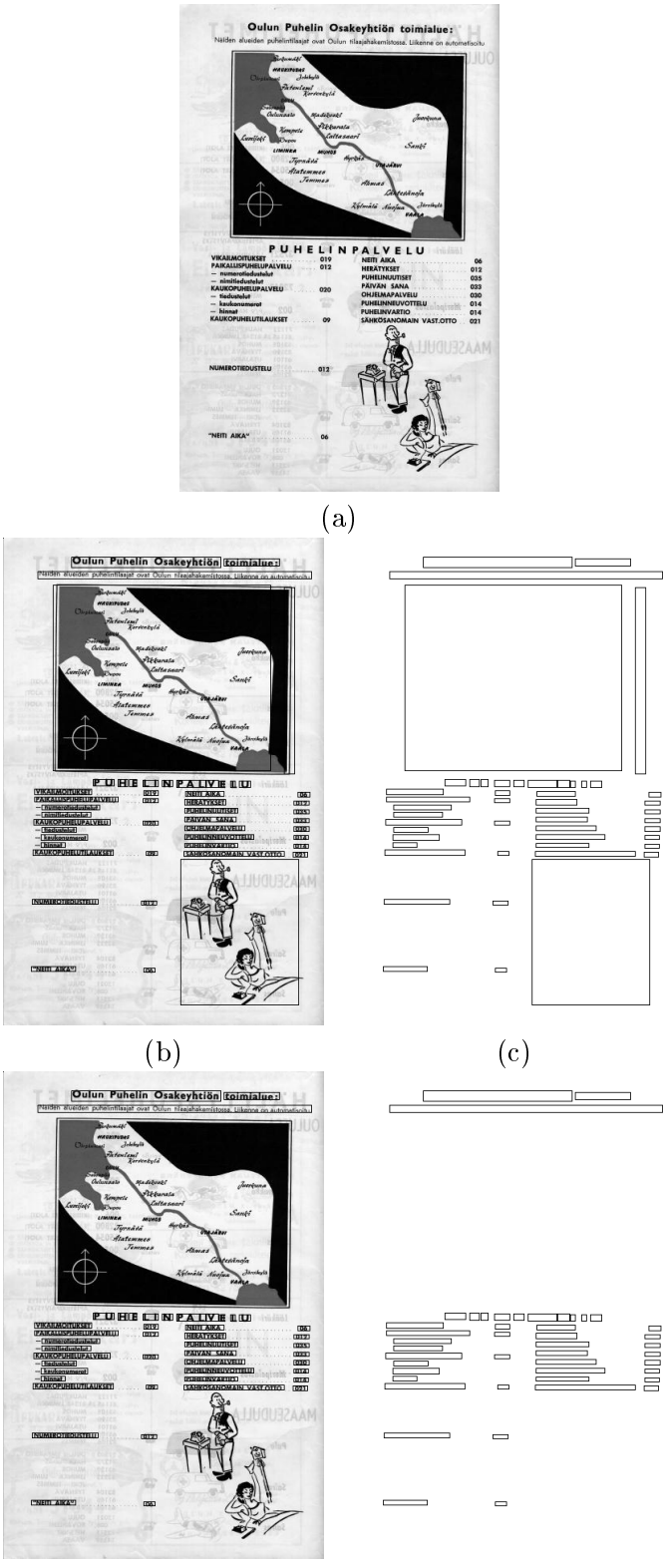

Figure 3: Example of a document image. Original image (a), regions of interest with bounding boxes ((b) and (c)) and text areas with bounding boxes ((d) and (e)) found by entropy discrimination.

Document class (MediaTeam labels)	Number of samples	Average performance	Standard deviation
ADDRESS_LIST	6	0.747939	0.361794
ADVERTISEMENT	24	0.955644	0.056020
ARTICLE	233	0.747378	0.283192
BUSINESS_CARDS	11	0.960173	0.074453
CHECK	3	0.931235	0.060796
COLOR_SEGMENTATION_IMAGES	0	0.000000	0.000000
CORRESPONDENCE	24	0.820787	0.268626
DICTIONARY	12	0.968567	0.057541
FORM	23	0.861363	0.219767
LINE_DRAWING	7	0.950642	0.072432
MANUAL	35	0.883139	0.222080
MATH	17	0.667742	0.400412
MUSIC	9	0.838010	0.324545
NEWSLETTER	42	0.855444	0.243351
OUTLINE	19	0.844259	0.244266
PHONEBOOK	7	0.880145	0.102296
PROGRAM_LISTING	12	0.923048	0.218133
STREET_MAP	3	1.000000	0.000000
TERRAIN_MAP	5	0.933333	0.149071

Figure 4: Experimental results with mask width parameter $w = \frac{W}{50}$ for cumulative gradient computations. (W is the with of the image.) Values reported for each class are the number of significant sample images (with at least one text region), performance mean and standard deviation over the subset of significant samples. In COLOR_SEGMENTATION_IMAGES, no text zone is specified for none of the images. This class is not interesting for our purposes.

Since the database is divided into nineteen parts, we tried to merge some of them together in order to get larger test sets for our system. Using results produced by our first experiment, we gathered MediaTeam samples into six classes named $\mathcal{C}_0, \mathcal{C}_1, \ldots, \mathcal{C}_5$ (Fig. 5).

For each pair of classes in the original database, we performed a variance comparison with Fisher-Snedecor test. If two classes behave similarly in variance, we carry on with comparison of means via the Student test. For all these inferences ([4, 24]), we used a level of significance $\alpha = 0.05$.

Once statistical equivalences between classes are established, we build an adjacency graph. In this structure, vertices are the classes of the database, and edges exist between classes with similar statistical behavior. Merging classes is equivalent to maximal cliques[1] research in the adjacency graph.

[1] A clique is a complete sub-graph

$$
\begin{aligned}
\mathcal{C}_0 &= ADDRESS\_LIST \cup ARTICLE \cup CORRESPONDENCE \cup MATH \cup MUSIC \\
\mathcal{C}_1 &= ADVERTISEMENT \cup BUSINESS\_CARDS \cup CHECK \cup DICTIONARY \cup \\
&\quad LINE\_DRAWING \cup MANUAL \cup NEWSLETTER \cup PROGRAM\_LISTING \cup \\
&\quad TERRAIN\_MAP \\
\mathcal{C}_2 &= COLOR\_SEGMENTATION\_IMAGES \\
\mathcal{C}_3 &= FORM \cup OUTLINE \\
\mathcal{C}_4 &= PHONEBOOK \\
\mathcal{C}_5 &= STREET\_MAP
\end{aligned}
$$

Figure 5: New classes built via statistical inference.

We ran our system on the new classes, with the same parameters as in previous experiment (Fig. 6).

Class	Average performance	Standard deviation
$\mathcal{C}_0$	0.798545	0.273572
$\mathcal{C}_1$	0.928004	0.172281
$\mathcal{C}_3$	0.889247	0.201742
$\mathcal{C}_4$	0.919551	0.084348
$\mathcal{C}_5$	1.000000	0.000000

Figure 6: Experiments for entropy filtering with new classes. Class $\mathcal{C}_2$ has been excluded for our experiments, since there is no text in its document images.

We tried to improve our system, introducing geometric considerations. As text lines are expected, we decided to eliminate regions with bounding boxes abnormally tall (taller than three times the mean height over all the regions of interest) before performing entropy discrimination. Some other filters can be set after entropy selection. We have tested eccentricity (ratio width versus height) and surface dicrimination: Zones with eccentricity under 0.25 are deleted. Among the remaining zones, areas with surface exceeding ten times the mean surface value are dropped. Results are reported in Fig. 7

Experiment	2	3	4
$\mathcal{C}_0$	0.756768	0.755442	0.754259
$\mathcal{C}_1$	0.902964	0.891879	0.888798
$\mathcal{C}_3$	0.888098	0.863941	0.861395
$\mathcal{C}_4$	0.880383	0.848546	0.848546
$\mathcal{C}_5$	1.000000	1.000000	1.000000

Figure 7: Average performances for experiments with height and entropy filtering (2); height, entropy and eccentricity filtering (3); height, entropy, eccentricity and surface filtering (4)

We quote a significant decrease of our system's performances due to height filter. This can be explained by important font sizes variability in documents we had to deal with (class $\mathcal{C}_3$ is the less affected, since forms are rather homogeneously printed). Similarly, eccentricity filter

may cause the loss of some isolated characters. We could lower the threshold, but setting too low a value would be equivalent to remove the filter.

Finally, adding heuristics is not of much help while working on the entire database. System improvement must be done for each class separatly, with some a priori knowledge. This will require a larger and more homogeneous database. Entropy remains the most important criteria for our system.

Conclusion

We have presented a two step text detection system. It firstly extracts regions of interest in a greyscale document image, using cumulative gradient considerations. Then it classifies them in text and non text zones on the basis of an entropy criteria. We tested it on an entire database, which is not commonly done in our research area. We decided to design a system which will support more generality than others.

We tried to add geometric heuristics to improve our system's performances. More experiments have to be achived to validate them. Heuristics combinations, permutation (that is, the order for the different filters) and constraints levels (threshold design) are to be more precisely examined for zones classification. This will be part of our further work. We guess we will have to bring some restriction to document classes to be processed, or develop specific knowledge-based rules (i.e. give a priori knowledge) for each particular class. A compromise will have to be done between performance and generality.

After text separation and physical structure retrieval, we'll still have to complete our system with a logical labeling module.

Acknowlegments

We wish to acnowledge all our colleagues in LIVIA and CENPARMI who helped us in this work; specially Dr. S.H. Kim (invited member of CENPARMI) for stimulating discussions, F. Grandidier and J. Milgram (LIVIA) for their advices and help in solving computing troubles.

References

[1] Webster dictionary.

[2] N. Amamoto, S Torigoe, and Y. Hirogaki. Block segmentation and text area extraction of vertically/horizontally written documents. In *Proceedings of thee second International Conference on Document Analysis and Recognition (ICDAR)*, pages 739–742, Tsukuba, Science City (Japan), 1993.

[3] M. Bahi. *Segmentation de surfaces représentées par des nuages de points non organisés.* PhD thesis, Université Claude Bernard de Lyon, Juillet 1997.

[4] Gérald Baillargeon. *Introduction à l'inférence statistique.* Editions S.M.G., Trois Rivière, Québec (Canada), 1992.

[5] Abdel Belaïd. Analyse et reconnaissance de documents. In *Le traitement électronique du document*, chapter 2, pages 11–47. ADBS Editions, Paris (France), 1994.

[6] Abdel Belaïd and Yolande Belaïd. *Reconnaissance des formes. Méthodes et applications.* Informatique, intelligence artificielle (iia). InterEdition, Paris (France), 1992.

[7] Ph. Bolon, J.-M. Chassery, D. Domigny J.-P. Cocquerez, C. Graffigne, S. Philipp A. Montanvert, R. Zéboudj, and J. Zérubia. *Analyse d'images: filtrage et segmentation.* Masson, Paris, Milan, Barcelone, enseignement de la physique edition, Octobre 1995.

[8] L. Boukined, B. Taconet, A. Zahour, and A Faure. Recherche de la structure physique d'un document imprimé par rectangulation. In 8^e *Congrès Reconnaissace de Formes et Intelligence Artificielle (RFIA)*, volume 3, pages 1027–1031, Lyon-Villeurbanne (France), Novembre 1991.

[9] Philippe Chauvet. Systèmes d'analyse, reconnaissance et description de documents complexes. In 8^e *Congrès Reconnaissance de Formes et Intelligence Artificielle (RFIA)*, volume 3, pages 1033–1044, Lyon-Villeurbanne (France), Novembre 1991.

[10] Myriam côté. *Utilisation d'un modèle d'accès lexical et de concepts perceptifs pour la reconnaissance d'images de mots cursifs.* PhD thesis, Ecole Nationale Supérieure des Télécommonications (ENST) de Paris, 1997.

[11] E. R. Davies. *Machine Vision: Theory, Algorithms, Practicalities.* Harcourt Brace Jovanovich, London, San Diego, New York, Boston, Sydney, Tokyo, academic press edition, 1990.

[12] Anil K. Jain. *Fundamentals of Digital Image Processing.* Thomas Kailath, Prentice Hall, Englewoods Cliffs, New Jersey, USA, prentice hall information ans system sciences series edition, 1989.

[13] Ramesh Jain, Rangachar Kasturi, and Brian G. Schunck. *Machine Vision.* McGraw-Hill Inc., mcgraw-hill series in computer science edition, 1995.

[14] M. Krishnamoorthy, G. Nagy, S. Seth, and M. Viswanathan. Syntactic segmentation and labeling of digitalized pages from technical journals. *IEEE Computer Vision, Graphics and Image Processing*, 47:327–352, 1993.

[15] Kyong-Ho Lee, Yoon-Chul Choy, and Sung-Bae Cho. Geometric structure analysis of document images: A knowledge-based approach. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 22(11):1224–1240, November 2000.

[16] G. Nagy and S. Seth. Hierarchical representation of optically scanned documents. In 7^{th} *International Conference on Pattern Recognition (ICPR)*, pages 347–349, Montreal (Canada), 1984. IEEE Computer Society Press.

[17] Lawrence O'Gorman and Rangachar Kasturi. *Document Image Analysis.* IEEE Computer Society Executive Briefing. IEEE Computer Society, Los Alamitos (California, USA), 1997.

[18] Oleg Okun, David Dœrmann, and Matti Pietikäinen. Page segmentation and zone classification: The state of the art, November 1999.

[19] J. R. Parker. *Algorithms for Image Processing and Computer Vision.* John Wiley and Sons, Chichester, New York, Brisbane, Toronto, Singapore, Weinheim, design and measurement in electronic engineering edition, 1997.

[20] T. Pavlidis. *Structural Pattern Recognition.* Springer-Verlag, Berlin, Heidelberg, New York, springer series in electrophysics edition, 1977.

[21] T. Pavlidis and J. Zhou. Segmentation by white streams. In *International Conference on Document Analysis and Recognition (ICDAR)*, pages 945–953, St-Malo (France), 1991.

[22] William K. Pratt. *Digital Image Processing.* John Wiley and Sons, New York, Chichester, Brisbane, Toronto, Singapore, wiley-interscience edition, 1991.

[23] Vincent Quint. Edition de documents structurés. In *Le traitement électronique du document*, chapter 1, pages 11–47. ADBS Editions, Paris (France), 1994.

[24] William C. Schefler. *Statistics. Concepts and Applications.* The Benjamin/Cummings Publishing Company, Inc., Menlo Park, California (USA), 1988.

[25] J. Serra. *Image Analysis and Mathematical Morphology (vol.1).* Academic Press, New York, 1982.

[26] J. Serra. *Image Analysis and Mathematical Morphology (vol.2).* Academic Press, New York, 1988.

[27] Souad Souafi-Bensafi, Frank Lebourgeois, and Hubert Emptoz. Modélisation et reconnaissance des structures de documents: application aux sommaires de revues. In *Actes du deuxième Colloque International Francophone sur l'Ecrit et le Document (CIFED)*, Lyon (France), July 3-5 2000.

[28] Souad Souafi-Bensafi, Frank Lebourgeois, Marc Parizeau, and Hubert Emptoz. Contribution à la reconnaissance des structures logiques hiérarchiques dans les documents papier. Technical report, Université Laval (Québec), 2000.

[29] Y. Y. Tang, C.D. Yan, M. Cheriet, and C.Y. Suen. Automatic analysis and understanding of documents. In .H. Chen Patrick S.P. Wang and L.F. Pau, editors, *Handbook of Pattern Recognition and Computer Vision.* The World Scientific Publishing Co. Pte, Ltd, Singapore, 1993.

[30] Souad Tayeb-Bey. *Analyse et conversion de documents: du pixel au langage HTML.* PhD thesis, Institut National des Sciences Appliquées (INSA) de Lyon, 1998.

[31] Ferdinand van der Heijden. *Image Based Measurement Sytems.* John Wiley and Sons, Chichester, New York, Brisbane, Toronto, Singapore, design and measurement in electronic engineering edition, 1994.

[32] Kwan Y. Wong, Richard G. Casey, and Friedrich M. Wahl. Document analysis system. *IBM Journal of Research and Developpment*, 26(6):647–656, November 1982.

[33] Victor Wu and R. Manmatha. Document image clean-up and binarization. Technical report, Computer Science Department, University of Massachusetts, Amherst (Massachussetts, USA), December 1997.

[34] Victor Wu, R. Manmatha, and Edward M. Riseman. Textfinder: An automatic system to detect and recognize text in images. Technical report, Computer Science Department, University of Massachusetts, Amherst (Massachussetts, USA), November 1997.

[35] Steven W. Zucker. Survey: Region growing: Childhood and adolescence. In *Computer Vision, Graphics and Image Processing*, volume 5, pages 382–399. Academic Press, 1976.