A typed and handwritten text block segmentation system for heterogeneous and complex documents

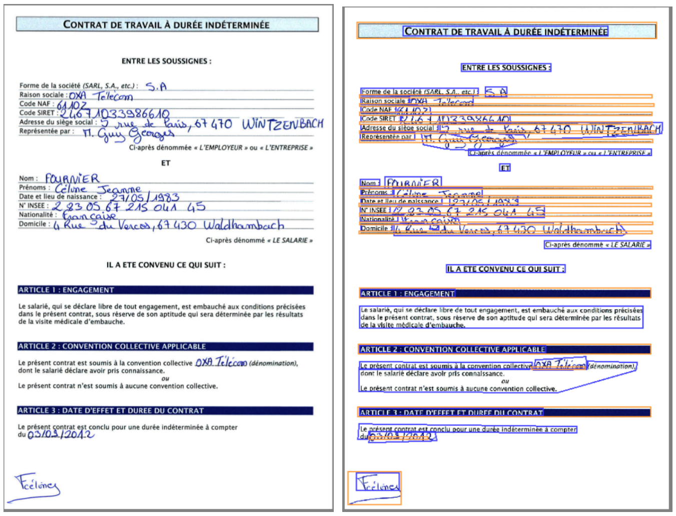
P. Barlas*, S. Adam*, C. Chatelain†, T .Paquet*
*Laboratoire LITIS - EA 4108, Universite de Rouen, FRANCE 76800
Email: philippine.barlas@insa-rouen.fr,{Sebastien.Adam,Thierry.Paquet}@litislab.eu
†Laboratoire LITIS - EA 4108, INSA Rouen, FRANCE 76800
Email: clement.chatelain@insa-rouen.fr

***Abstract*—This paper presents a Document Image Analysis (DIA) system able to extract homogeneous typed and handwritten text regions from complex layout documents of various types. The method is based on two connected component classification stages that successively discriminate text/non text and typed/handwritten shapes, followed by an original block segmentation method based on white rectangles detection. We present the results obtained by the system during the first competition round of the MAURDOR campaign.**

I. Introduction

Document image analysis is an important research issue that has grown over the last twenty years. As a consequence of the important number of contributions in this domain, several significant efforts have been accomplished in evaluating document analysis systems. In 2013, the first MAURDOR campaign [8] was led to evaluate the progress in automatic systems dedicated to document image analysis. This campaign makes an important step beyond existing ones ([1], [19], [20]) through the variability of the documents to be processed. Indeed, the dataset contains heterogeneous documents (blank or completed forms, printed and manually annotated business documents, handwritten correspondence, maps, ID, newspapers articles, blueprints, etc.), with mixed typed and handwritten text, in various language (French, English and Arabic). Moreover, the MAURDOR campaign assesses not only complete processing chains (going from document segmentation to information retrieval, through text recognition) but also each step independently. In order to lead this evaluation of different tasks, document analysis problem have been divided into five subtasks respectively dedicated to segmentation, writing type identification, language identification, text recognition for each type/language, and information retrieval.

In this paper, we present a part of the system proposed by the LITIS for the first module. This module concerns the task of document segmentation into 8 classes of homogeneous areas : text, photographic image, hand drawn line area, graph area, table area, edge line area, separator, and material damage area. In this work, we focus on the subtask addressing the problem of text segmentation, which is the main issue for subsequent text recognition and information retrieval. Within the MAURDOR context, documents can contain both typed and hanwritten text, in different languages and mixed with other graphical information. As a consequence of this heterogeneity, the MAURDOR assessment protocol requires to identify homogeneous blocks in script and language, in such a way that text blocks can be submitted to the corresponding recognition system. Figure 1 illustrates the objective of text detection in such a context.



CONTRAT DE TRAVAIL À DURÉE INDÉTERMINÉE

ENTRE LES SOUSSIGNES :

Forme de la société *(SARL, S.A., etc.)* : S.A
Raison sociale : OXA Télécom
Code NAF : 6110Z
Code SIRET : 24671033986610
Adresse du siège social : 5 rue de Paris, 67 470 WINTZENBACH
Représentée par : M. Guy Georges

Ci-après dénommée *« L'EMPLOYEUR »* ou *« L'ENTREPRISE »*

ET

Nom : FOURNIER
Prénoms : Céline Jeanne
Date et lieu de naissance : 27/05/1983
N° INSEE : 2 83 05 67 215 041 45
Nationalité : Française
Domicile : 4 Rue du Vercos, 67 430 Waldhambach

Ci-après dénommé *« LE SALARIE »*

IL A ETE CONVENU CE QUI SUIT :

ARTICLE 1 : ENGAGEMENT

Le salarié, qui se déclare libre de tout engagement, est embauché aux conditions précisées dans le présent contrat, sous réserve de son aptitude qui sera déterminée par les résultats de la visite médicale d'embauche.

ARTICLE 2 : CONVENTION COLLECTIVE APPLICABLE

Le présent contrat est soumis à la convention collective OXA Télécom *(dénomination)*, dont le salarié déclare avoir pris connaissance.

ou

Le présent contrat n'est soumis à aucune convention collective.

ARTICLE 3 : DATE D'EFFET ET DUREE DU CONTRAT

Le présent contrat est conclu pour une durée indéterminée à compter du 03/09/2012

Fig. 1. Example of a document and its text zones ground truth

The problem of text detection has been the subject of many contributions in the literature [14]. One can oppose region-based methods and texture based approaches. Region-based methods analyse the properties of either the connected components or the edges of the images in order to locate the text [13], [15], [16]. The texture-based approaches [10], [12], [18] try to identify the textural specificities of the text entities. However, whatever the kind of approaches, existing works are generally dedicated to a particular class of document (forms, handbook, handwritten mail document, graphic documents, . . .), containing a given script (typed or handwritten), in a given language. Therefore, existing approaches are not subject to the homogeneous script and language constraint.

In this work, we propose a connected component oriented approach for text identification and segmentation. The particularity of the approach relies on the fact that the system can handle heterogeneous and complex documents thanks to a learning based approach. The results obtained during the MAURDOR campaign using this system are presented, discussed and compared with 3 other participating systems. In section 2, an overview of the proposed method is given. Then, details concerning the three main steps are given in section 3, 4 and 5. Section 6 presents experimental results obtained during

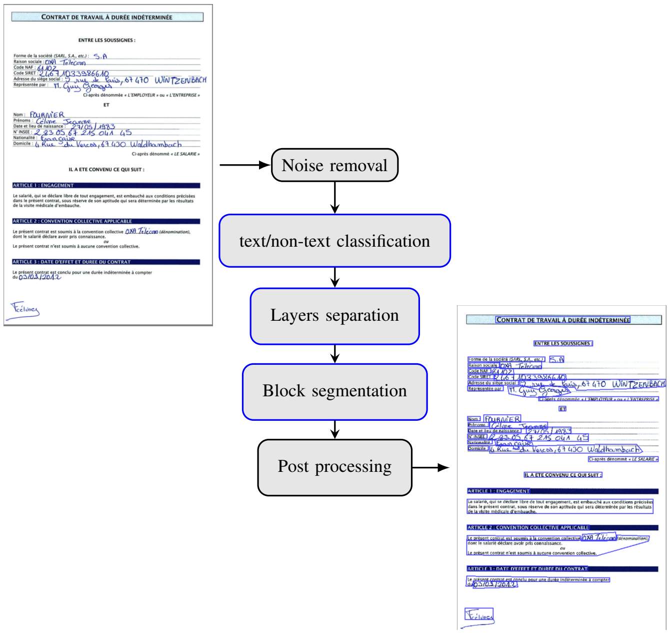


Fig. 2. Overview of the proposed system

the first MAURDOR campaign. Finally, the paper concludes with a brief summary and a discussion of future work.

II. Overview of the proposed system

The system proposed by the LITIS to handle the task of area segmentation is composed of several detectors (text, tables, images...) that work in parallel. The different detectors work at different levels of the document: image of the document, connected components, lines, and blocks, depending on the nature of the objects to be detected. As an example, the table detector [7] uses a line based representation of the document that feed a classifier to locate tables. The text detector, which is the main purpose of this paper, is based on both connected component information and an original document segmentation method based on white zones.

As one can see on Fig 2, the proposed approach relies on three main steps preceded by a noise removal preprocessing stage. The preprocessing consists in filtering small connected components (CC) as well as large CC's close to the borders of the document. The first main step is a CC text/non-text classification. It consists in extracting simple shape features to classify CC's into text or non-text components. The second main step is a layer separation that consists in separating textual CC's into typed components and handwritten components. Finally, the third step consists in a block segmentation based on the search of empty rectangles applied on the three layers (non-text, typed and handwritten) previously obtained. Finally, a post processing stage combines blocks between handwritten and typed layers in order to reduce segmentation errors by removing small handwritten blocks included in a typed block and vice versa. The following sections focus on the three steps dedicated to text identification and segmentation and identified by blue boxes in Fig 2.

III. Text/non-text detection

A key step in our system is the discrimination of each connected component into text or non-text component. We use a learning based approach consisting in extracting simple features representing the shape of the connected components and its neighborhood, that feed a MLP classifier.

Feature extraction : Both the shapes of a connected component and its context contain discriminative information for the text/non-text separation since textual connected components have regular shapes (regular height, width ...) while the shape of graphical connected components includes a lot of variability. Therefore, for each connected component, a set of simple features inspired by [17] is extracted: aspect ratio, area ratio, density, compactness, eccentricity, number of connected components included in the current connected component, and number of connected components overlapped with the current connected component.

$$f_1 = \frac{min(W_c,H_c)}{max(W_c,H_c)},\ f_2 = \frac{A_c}{A_p},\ f_3 = \frac{\#BlackPixelsCC}{A_c},$$

$$f_4 = \frac{PerimeterCC}{\#BlackPixelsCC},\ f_5 = \frac{(\mu_{20}-\mu_{02})^2-4\mu_{11}}{(\mu_{20}+\mu_{02})^2},$$

$$f_6 = \#CC\_included,\ f_7 = \#CC\_overlapped,$$

where W_c, H_c, A_c represent the width, height, area of the connected component and W_p, H_p, A_p represent the width, height, area of the picture.

The context around the connected component is also informative to discriminate horizontally aligned textual connected components from irregular graphical connected components. Therefore, we add three window-based features obtained extracting the variance of the neighborhood of the connected component. The context features are obtained calculating the variance of the width, the height and the y-position of the centroid of the connected components included in a window centered around the current connected component. The window used is a rectangular window expressing the horizontal alignment of textual components. Its size is adapted to the size of the connected component so that the width of the window is $9 \times W_c$ and its height is H_c.

Classification : A MLP is trained on a set of 2000 document images from the MAURDOR training dataset [8] containing both text and graphic components (forms, catalogue pages, bills, administrative documents, maps, handwritten letters, etc.). The MAURDOR dataset is labeled at the block level so that we defined a connected component label regarding the label of the block in which it is included. The training dataset is composed of 100000 textual connected components (both typed and handwritten) and 100000 non-textual connected components (logos, signatures, drawings, form-fields, images, technical scheme ...), randomly chosen from the documents. We have tested the system with a cross validation approach on 4-folds. The average precision and recall values obtained for text are 82.9% and 82.7% respectively and for non-text the precision and recall are 82.6% and 83.0%. The Figure 3 shows an example of result obtained after the step of text/non-text detection.

IV. Layers separation

One of the major difficulties of the segmentation task in the MAURDOR campaign is to produce homogeneous

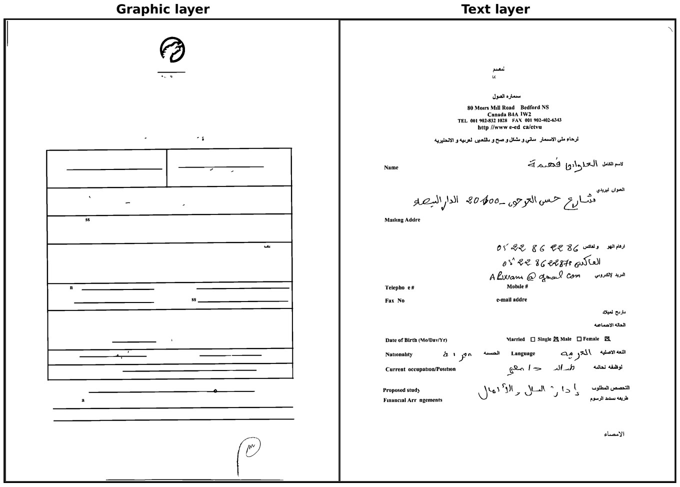


Fig. 3. Example of a text/non text separation result

semantic areas. Indeed, this requires the separation of text blocks into typed blocks and handwritten blocks. Moreover, the dataset includes three main languages : French, English and Arabic. This multilingual context further complicates the distinction between typed and handwritten since unlike French and English, the Arabic typed writing is cursive.

The classification between handwritten and typed components rely on a codebook based approach, inspired from the method described in [5]. Hence, the learning stage relies on two steps. First, a codebook is built. It contains a collection of contour fragments extracted from a first connected components learning dataset. Then, a MLP classifier is learnt using as features the histogram of occurrences of the fragments of this codebook in the connected components of a second learning dataset.

A. Codebook construction

Fragment extraction and representation: An efficient way to discriminate writing type is to extract fragments of external contour of connected components. A fragment is defined by a fixed length l and an overlapping area of fixed size s, moving along the external contour of the connected component as illustrated on Fig 4. The overlapping area represents the number of pixels shared by the fragment i and the fragment $i + 1$. Fragments are extracted over the whole contour of the connected component. We choose to represent fragments using the chaincode histogram (CCH) described in [6] which is a translation and scale invariant shape descriptor.

Codebook generation: The codebook generation step aims at finding a collection of similar contour fragments in a first learning dataset. In the proposed system, this stage is realized through a 2D Self-Organizing Map (SOM) [4] trained on CCH feature vectors. In order to tackle the difficulty of discriminating typed and handwritten text in the presence of both latin and arabic script, the dataset has to contain four kinds of text : Latin typed, Arabic typed, Latin handwritten and Arabic handwritten. As a consequence, for the codebook construction, many dataset have been merged to be representative of the different writings. A set of Arabic handwritten fragments has been extracted on the IFNENIT database [9]. Latin handwritten fragments are taken from the RIMES database [1] and typed fragments in both Arabic and Latin come from automatically generated images. Different sizes of fragments were tested as well as various sizes of overlap. The size of a fragment has been experimentally fixed to $l = 20$ pixels and $s = 10$ pixels. The number of fragments extracted for each class is about 130000 fragments. Several codebooks of different sizes were tested and we chose empirically to use a 20×20 codebook.

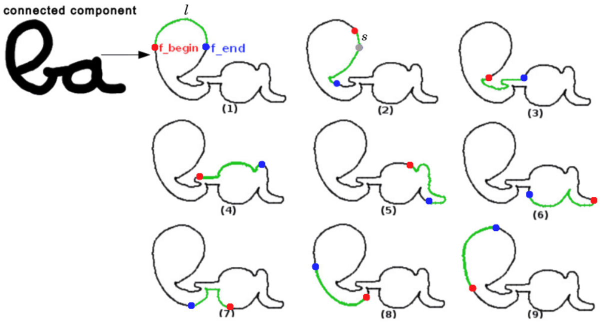


Fig. 4. Fragments extraction on a connected component

B. Connected component classification

Once the codebook built, a feature vector is extracted from each connected component of the MAURDOR dataset in Arabic and Latin for both typed and handwritten. For each connected component of this dataset, fragments are extracted and for each fragment of the connected component, the nearest fragment in the codebook is identified using an euclidean distance. Then, the number of occurrence of codebook fragments in the external contour of the component is computed. This leads to a 400 features vector.

A MLP is thus trained on this dataset containing approximately 25000 samples of each class (Arabic typed, Latin typed, Arabic handwritten and Latin handwritten). The script decision is taken at the connected component level and the result is mapped into two classes : typed and handwritten. The performance of the typed/handwritten classification task is presented in table I.

TABLE I. PERFORMANCE OF THE TYPED/HANDWRITTEN CLASSIFICATION TASK USING A 4-FOLDS CROSS VALIDATION

%	**Recall**	**Precision**
typed	79.0	83.6
handwritten	80.7	75.7

We are now able to discriminate the connected component into three layers (a graphic layer, a typed text layer and a handwritten text layer). We now present the proposed segmentation algorithm that gather the connected components into homogeneous blocks that belong to one of the three types.

V. BLOCK SEGMENTATION

This section describes the approach for producing the homogeneous areas using the connected components and their classification. The method is illustrated in Fig 5. First an aggregating method is used to group the connected components into bigger entities using RLSA; and then a detection of vertical and horizontal white spaces allows to produce a mask

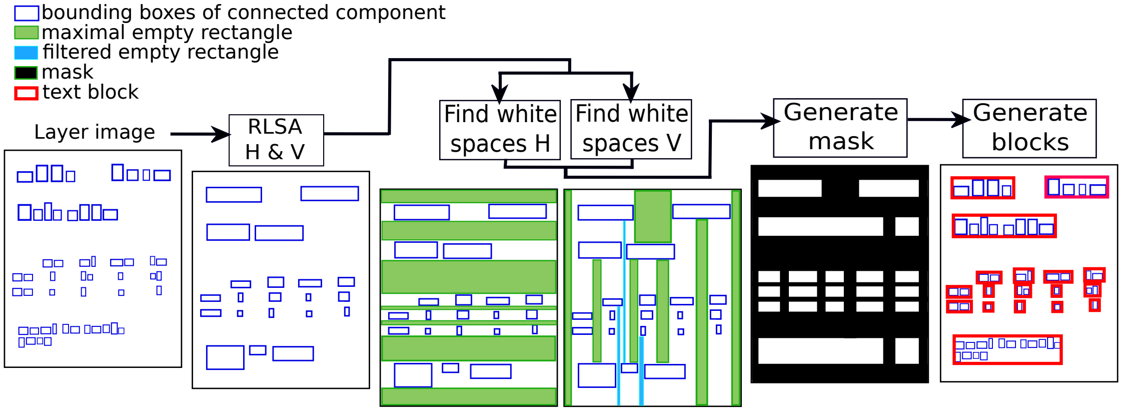


Fig. 5. The proposed method for block segmentation

that segments the document into areas. These two main steps are now described.

Aggregating connected components: The first step of the approach consists in applying the Run Length Smoothing Algorithm (RLSA) on the image of the layer. The horizontal and vertical RLSA is used to connect close connected components. The thresholds used for the horizontal and vertical RLSA depend on the mean width and mean height of the connected components in the layer. This way, the thresholds will be higher for handwritten text where spaces are more important between connected components. Conversely the thresholds will be smaller for typed text where letters and words are closer from each other.

Segmenting the document using white spaces: The second step consists in looking for maximal white spaces using the method proposed by Thomas Breuel in [3] and implemented in the Leptonica library [2]. This approach finds a cover of the background whitespace of a document in terms of maximal empty rectangles. Rectangles whose size is too small (in light blue in Fig 5) are filtered out in order to keep only the significant ones. White empty rectangles (in green in Fig 5) are used to generate a binary mask for block segmentation (0 corresponding to a pixel covered by a white rectangle and 1 corresponding to a potential text pixel). We extract text blocks by applying the mask on the connected components of the layer.

VI. Experimental results

The system was evaluated during the first MAURDOR campaign in march 2013. In this section, the MAURDOR dataset is presented, the metric are described, and the results are exposed.

A. *The MAURDOR dataset*

The MAURDOR dataset is composed of heterogeneous documents distributed according the following categories :
C1 (12%) : Blank or completed (by hand) forms;
C2 (40%) : Printed, but also manually annotated business documents (invoice, bill, receipt, catalogue page, newspaper article, contract, legal or administrative document, check, map, plan, drawing, reservation confirmation etc.);
C3 (25%) : Private handwritten correspondence, sometimes with printed letterheads (invitation letter, post-it, block-note page, etc.);
C4 (20%) : Printed, but also manually annotated business correspondence (handwritten mail, medical receipt, scanned mail, fax header,etc.);
C5 (3%) : Other documents such as plans, schemes, drawings, alphanumeric tables, etc.

Fonts and handwriting are different across documents and documents are digitized according to different methods. The documents are either in French, Arabic or English but they can occasionally contain text in other languages. The Fig 6 contains some examples of documents.

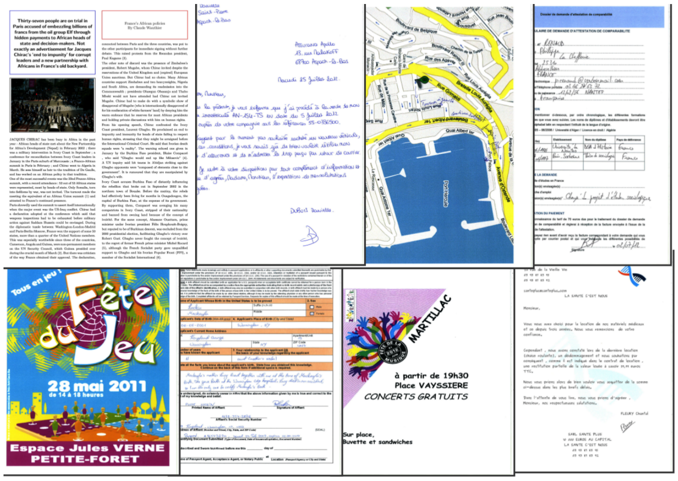


Fig. 6. Example of documents used in the MAURDOR campaign

For the first campaign the corpus was composed of 4000 documents for training and 1000 documents dedicated to the evaluation.

B. *The ZoneMap metric*

The official metric for the MAURDOR campaign is The ZoneMap metric, which has been proposed by The French National Metrology and Testing Laboratory (LNE) to evaluate the task of zone segmentation and classification. After grouping reference and hypothesis zones according to the maximization of the coverage rate between them, five configurations are distinguished : Match, Miss, FalseAlarm, Merge and Split. Then, a global error rate is calculated measuring the surface error as well as the zone classification error on the basis of the different configurations. This computation takes into account pixels values in order to stronger penalize errors in informative zones than in background regions. The ZoneMap metric is fully detailed in the evaluation plan available on the website of the campaign [8].

C. The first MAURDOR campaign results

In this section, the results of the proposed system obtained during the first MAURDOR campaign are given and compared with the results of the three other participants. These results are given using both the ZoneMap metric mentioned above (the smaller the better) and the standard Jaccard index (the higher the better), defined as the ratio $(R \cap H)/(R \cup H)$ weighted by the pixel values. In both cases, we consider only the text regions of the documents, what can be done using the tool provided by the LNE by setting the weight of the text to '1' and the others to '0'.

TABLE II. RESULTS OF THE FIRST CAMPAIGN FOR TEXT DETECTION AND SEGMENTATION

Participants	**ZoneMap**	**Jaccard (%)**
LITIS (this work)	**26.35**	57.8
participant_1	44.79	**65.3**
participant_2	37.23	56.0
participant_3	34.13	46.6

As one can see in this table, the LITIS system obtains the best results for text detection from the ZoneMap metric point of view, and the second position when considering the Jaccard index. This difference can be explained through the definition of the ZoneMap metric that penalizes region splitting and merging. Hence, it seems that the proposed system performs well in the segmenting subtask and that further improvements are needed for the detection subtask. By qualitatively analyzing the results, we have highlighted that one of the major drawback of our system is the detection of the text in photographic image and graphical parts of documents. This statement is confirmed by table III which provides the ZoneMap values according to document categories. Indeed, this table shows that the system is less efficient to extract text in the C5 category, that contains many graphical parts.

TABLE III. ZONEMAP SCORES FOR TEXT DETECTION AND SEGMENTATION ACCORDING TO THE CATEGORY OF DOCUMENT (THE SMALLER THE BETTER)

Participants	**C1**	**C2**	**C3**	**C4**	**C5**
LITIS (this work)	**23.88**	**27.88**	**22.51**	**17.41**	**38.27**
participant_1	37.26	47.72	42.91	26.72	61.00
participant_2	39.21	38.44	25.89	30.48	45.56
participant_3	34.33	32.06	32.12	54.52	44.36

VII. DISCUSSION AND FUTURE WORK

This paper has presented a system for text detection and segmentation in heterogeneous documents in terms of layout (forms, newspapers, bill, private correspondence ...), script (typed and handwritten) and language (French, English and Arabic). This system relies on a learning based approach that combines the connected components information for the text detection and the white rectangles analysis for the segmentation. It achieves the best performances during the first MAURDOR campaign for the text detection and segmentation evaluated with the ZoneMap metric. Although efficient, the system could be improved by avoiding the detection of text blocks in the graphical parts of the documents. These false alarms are mainly due to the presence of graphical artifacts in the text layer (misclassification of small graphical connected components). Our current works concern the detection of the graphical parts of the document, in order to consider a different approach for text detection in these zones.

REFERENCES

[1] E. Augustin, M. Carr, E. Grosicki, J. M. Brodin, E. Geoffrois and F. Preteux, *RIMES evaluation campaign for handwritten mail processing*, In Proceedings of the Workshop on Frontiers in Handwriting Recognition, 2006, 231–235

[2] D. Bloomberg, *Leptonica: An open source C library for efficient image processing, analysis and operation*, http://www.leptonica.com/, 2007

[3] T.M. Breuel, *Two Geometric Algorithms for Layout Analysis*,In Workshop on Document Analysis Systems,2002,188–199

[4] M. Bulacu and L. Schomaker, *A comparison of clustering methods for writer identification and verification*, In Proceedings of the Eighth International Conference on Document Analysis and Recognition,2005, 1275–1279

[5] G. Ghiasi and R.W. Daly, *An efficient method for offline text independent writer identification*, Pattern Recognition (ICPR),IEEE,2010,1245–1248

[6] J. Iivarinen and A. Visa, *Shape Recognition of Irregular Objects*, Intelligent Robots and Computer Vision XV: Algorithms, Techniques, Active Vision, and Materials Handling, Proc. SPIE 2904,1996,25–32

[7] T. Kasar, P. Barlas, S. Adam, C. Chatelain and T. Paquet, *Learning to Detect Tables in Scanned Document Images using Line Information*,In Proceedings of the Twelfth International Conference on Document Analysis and Recognition, 2013

[8] MAURDOR campaign website, *http://www.maurdor-campaign.org/*

[9] M. Pechwitz, S. Snoussi Maddouri, V. Mrgner, N. Ellouze and H. Amiri, *IFN/ENIT-DATABASE OF HANDWRITTEN ARABIC WORDS* , CIFED, 2002

[10] A.K. Jain and S. Bhattacharjee, *Text segmentation using gabor filters for automatic document processing*, Machine Vision and Applications, Volume 5, 1992, 169–184

[11] F.M. Wahl, K.Y. Wong and R.G. Casey, *Block segmentation and text extraction in mixed text/image documents*, Computer Graphics and Image Processing, Volume 20, 1982, 375–390

[12] S. Kumar,R. Gupta, N. Khanna, S. Chaudhury and S.D. Joshi, *Text Extraction and Document Image Segmentation Using Matched Wavelets and MRF Model*, Image Processing, IEEE Transactions on , Volume 16, 2007, 2117–2128

[13] Y. Zhong, K. Karu and A.K. Jain, *Locating text in complex color image*, Pattern Recognition, 1995, 1523–1535

[14] K. Jung, K. Kim and A.K. Jain, *Text information extraction in images and video: a survey*, Pattern Recognition, Volume 37, 2004, 977–997

[15] J. Ohya, A. Shio and S. Akamatsu, *Recognizing characters in scene images*, IEEE Trans. Pattern Anal. Mach. Intell., 1994 , 214–224

[16] H. Hase, T. Shinokawa, M. Yoneda, C.Y. Suen, *Character string extraction from color documents*,Pattern Recognition,Volume 34, 2001, 1349–1365

[17] C. Chatelain, L. Heutte and T. Paquet, *A syntax-directed method for numerical field extraction using classifier combination*, IWFHR, 2004, 93–98

[18] V. Wu, R. Manmatha, E.M. Riseman, *TextFinderan automatic system to detect and recognize text in images*, IEEE Trans. Pattern Anal. Mach. Intell., 1999, 1224–1229

[19] B. Gatos, N. Stamatopoulos, G. Louloudis, *ICDAR 2009 Handwriting Segmentation Contest*, ICDAR, 2009, 1393–1397

[20] E. Grosicki, H. El Abed, *ICDAR 2011 - French Handwriting Recognition Competition*, ICDAR, 2011: 1459–1463