Page Frame Detection for Double Page Document Images

N. Stamatopoulos
Computational Intelligence Laboratory,
Institute of Informatics and Telecommunications,
National Research Center "Demokritos"
153 10 Athens, Greece
nstam@iit.demokritos.gr

B. Gatos
Computational Intelligence Laboratory,
Institute of Informatics and Telecommunications,
National Research Center "Demokritos"
153 10 Athens, Greece
bgat@iit.demokritos.gr

T. Georgiou
Department of Informatics and Telecommunications
University of Athens, Greece
t.georgiou@di.uoa.gr

ABSTRACT
Scanning two book pages at the same time helps to accelerate the scanning process but on the other hand introduces several difficulties if the user needs to have one page per image. A major difficulty is the appearance of noisy black borders around text areas as well as of noisy black stripes between the two pages. In this paper, we propose a novel algorithm for detecting the page frames on double page document images. Our aim is to split the image into the two pages as well as to remove noisy borders. First we apply a pre-processing which includes binarization, noise removal and image smoothing. Then, we detect the vertical zones of the two pages. In this stage, we introduce the vertical white run projections which have been proved efficient for detecting vertical zones of text areas. Finally, the horizontal zones of the two pages are detected based on horizontal white run projections. The experimental results on several double page document images from fifteen different books demonstrate the effectiveness of the proposed technique.

Categories and Subject Descriptors
I.7.5 [**DOCUMENT AND TEXT PROCESSING**]: Document Capture – *Document analysis, Scanning*; I.4.3 [**IMAGE PROCESSING AND COMPUTER VISION**]: Enhancement

General Terms
Algorithms, Design, Experimentation

Keywords
Document Image Enhancement, Border Removal, Page Splitting

DAS '10, June 9-11, 2010, Boston, MA, USA

1. INTRODUCTION
Document images are usually produced by scanning books or periodicals. Scanning two pages at the same time is a very common practice as it helps to accelerate the scanning process. However, it may affect the performance of subsequent processing such as document analysis and optical character recognition (OCR) since the majority of approaches are able to process only single page images. Furthermore, another drawback of scanning two pages at the same time is the appearance of noisy black borders around text areas as well as of noisy black stripes between the two pages (see Fig.1).

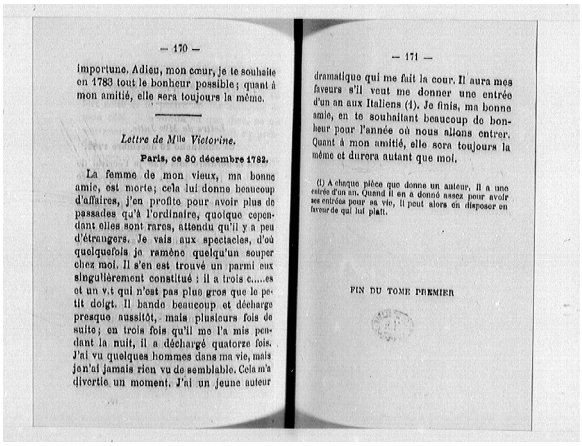


Figure 1. An example of a double page document image.

There are only few techniques in the literature that address the problem of page splitting. According to these approaches, double page documents are split in their middle after border removal or by defining the coordinates of the print space. Yacoub et al. [13] involve a splitting process as a preprocessing stage for the conversion of a large collection of complex documents and deployment for online web access to its information rich content. First, the double page images are cropped to the border of the page, using several criteria, which include searching for horizontal and vertical projections of straight page edges and checking the computed dimension versus a database of magazine sizes sampled throughout history. Once the double-pages are cropped, the centerfold of the double-page is identified and two individual pages are generated using a splitting operation.

Furthermore, a page splitting approach is proposed in the frame of the METADATA ENGINE Project [5]. This method is based on the assumption that books are printed according to a clearly defined printing space. Only a limited number of special elements may appear in the surrounding margins. Therefore, the method determines the coordinates of the print space used in the given document and then applies this zone to the actual page image. Next a virtual margin is added around the printing space and cut the pages. If a document contains supplements that do not conform to the default print space, such as maps, tables, graphs, and pictures, this variation is detected automatically. Finally, there are also commercial packages that support the process of detection and splitting double page document images such as WiseBook [12] and Scan Tailor [9].

Concerning only border detection and removal many different techniques have been proposed. Fan et al. [3] proposes a scheme to remove the black borders of scanned documents by reducing the resolution of the document image. Le and Thoma [7] propose a method for border removal which is based on classification of blank/textual/non-textual rows and columns, location of border objects, and an analysis of projection profiles and crossing counts of textual squares. In [1], Avila and Lins propose the invading and non-invading border algorithms which work as "flood–fill" algorithms. Moreover, Stamatopoulos et al. [11] rely on projection profiles combined with a connected component labelling process. Signal cross-correlation is also used in order to verify the detected noisy text areas. Finally, Shafait et al. [10] try to find the actual page contents area, ignoring marginal noise along the page border. They rely on geometric matching algorithms to find the optimal page frame of structured documents by exploiting their text alignment property. A common feature of all these techniques is that they process only single page images.

In this paper we propose a novel methodology that detects the optimal page frames of double page document images that is based on the vertical and horizontal white run projections. Our aim is to split the image into the two pages as well as to remove noisy borders. The remainder of the paper is organized as follows. In Section 2 the proposed methodology is detailed while experimental results are discussed in Section 3. Finally, conclusions are drawn in Section 4.

2. Proposed Methodology

The proposed methodology for page frame detection of double page document images is illustrated in Fig.2. It consists of three distinct steps. At a first step, a pre-processing which includes binarization, noise removal and image smoothing is applied. At a next step, the vertical zones of the two pages are detected. Finally, the frame of both pages is detected after calculating the horizontal zones for each page. Problem definition as well as a detailed description of all steps are given in the following subsections.

2.1 Problem Definition

Consider the input double page gray scale image I_g with dimension of I_xxI_y. A safe criterion to verify that I_g is a double page image is that I_x must be greater than I_y. Our aim is to calculate the frames of the two pages defined by the coordinates (xL_1,yL_1)-(xL_2,yL_2) and (xR_1,yR_1)-(xR_2,yR_2) as demonstrated in Fig.3(a). Using this information we produce two images that contain only page information (see Fig.3(b)-(c)).

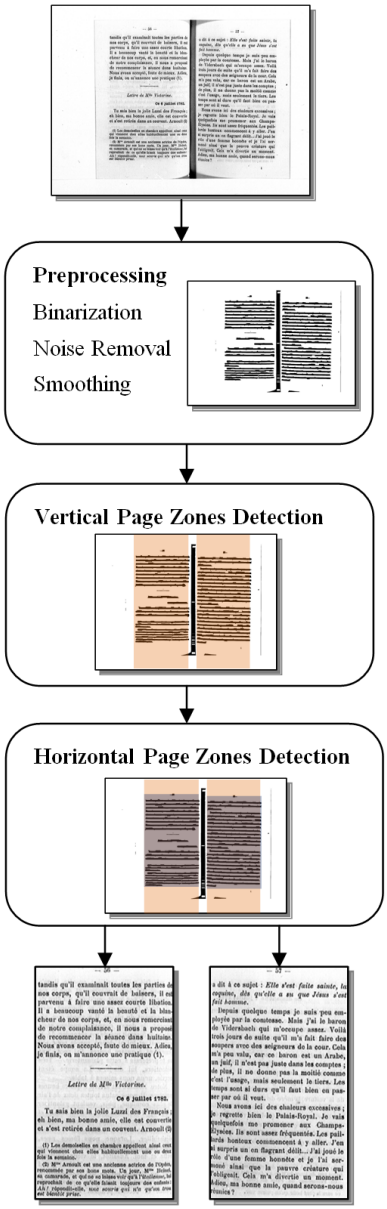
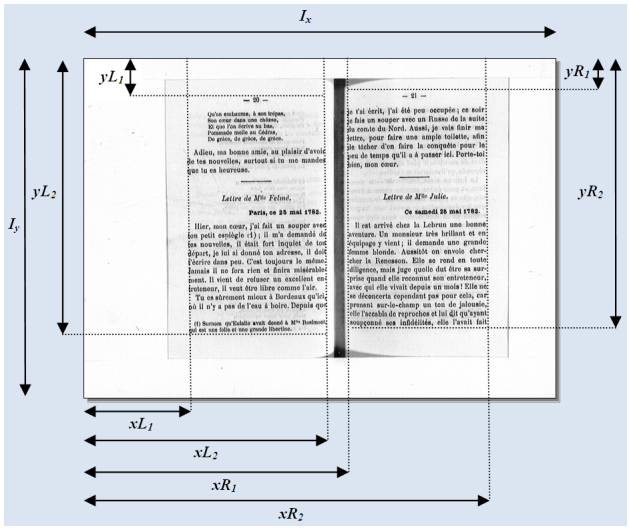


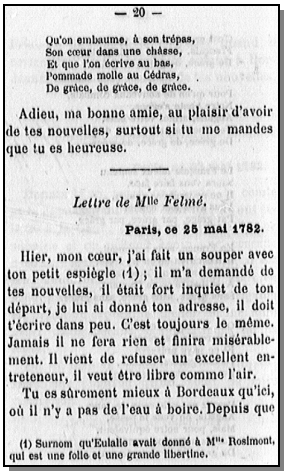
Figure 2. Block diagram of the proposed methodology.

2.2 Pre-processing

In this stage, we first proceed to image binarization using the efficient technique proposed in [4]. Then, we calculate the average character height C_h of the image based on the connected component histogram [6]. At a next step, we remove noisy small components having height or width less than C_h/10 and we proceed to image smoothing based on ARLSA [8]. Using this algorithm we connect pixels at text line level without merging text components with neighboring black border stripes. Fig.4 demonstrates the pre-processing steps.



(a)



— 20 —

Qu'on embaume, à son trépas,
Son cœur dans une châsse,
Et que l'on écrive au bas,
Pommade molle au Cédrat,
De grâce, de grâce, de grâce.

Adieu, ma bonne amie, au plaisir d'avoir de tes nouvelles, surtout si tu me mandes que tu es heureuse.

Lettre de Mlle Felmé.

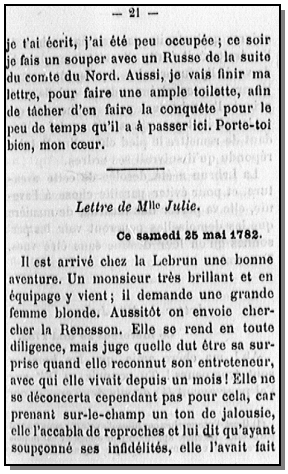
Paris, ce 25 mai 1782.

Hier, mon cœur, j'ai fait un souper avec ton petit espiègle (1) ; il m'a demandé de tes nouvelles, il était fort inquiet de ton départ, je lui ai donné ton adresse, il doit t'écrire dans peu. C'est toujours le même. Jamais il ne fera rien et finira misérablement. Il vient de refuser un excellent entreteneur, il veut être libre comme l'air.

Tu es sûrement mieux à Bordeaux qu'ici, où il n'y a pas de l'eau à boire. Depuis que

(1) Surnom qu'Eulalie avait donné à Mlle Rosimont, qui est une folle et une grande libertine.

(b)



— 21 —

je t'ai écrit, j'ai été peu occupée ; ce soir je fais un souper avec un Russe de la suite du comte du Nord. Aussi, je vais finir ma lettre, pour faire une ample toilette, afin de tâcher d'en faire la conquête pour le peu de temps qu'il a à passer ici. Porte-toi bien, mon cœur.

Lettre de Mlle Julie.

Ce samedi 25 mai 1782.

Il est arrivé chez la Lebrun une bonne aventure. Un monsieur très brillant et en équipage y vient ; il demande une grande femme blonde. Aussitôt on envoie chercher la Renesson. Elle se rend en toute diligence, mais juge quelle dut être sa surprise quand elle reconnut son entreteneur, avec qui elle vivait depuis un mois ! Elle ne se déconcerta cependant pas pour cela, car prenant sur-le-champ un ton de jalousie, elle l'accabla de reproches et lui dit qu'ayant soupçonné ses infidélités, elle l'avait fait

(c)

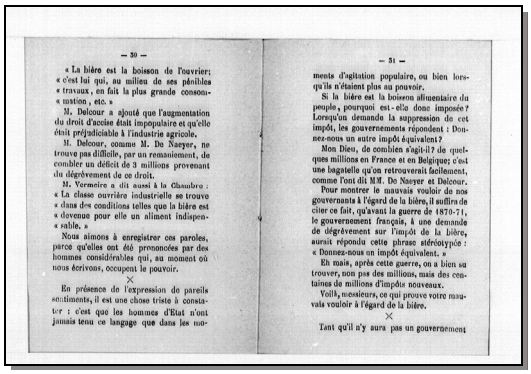
Figure 3. Problem definition: (a) input double page image; (b)-(c) output images after page frame detection.

2.3 Detection of Vertical Page Zones

In this stage, we aim to define the vertical page zones that are defined by offsets xL_1, xL_2 for the left page and xR_1, xR_2 for the right page, respectively (see Fig.3(a)). For this purpose, we use vertical projections only in the interval of (x,I_y+a*I_y)-(x,I_y-a*I_y) in order to avoid taking into account undesirable noisy areas at the top or the bottom of the image. For our experiments, we consider a=1/8. We focus on the white pixels of the image and introduce the vertical white run projections $HV()$ which have been proved efficient for detecting vertical zones of text areas. The motivation for proposing these projections is the need to stress the existence of long vertical white runs in the image. The vertical white run projections $HV()$ are defined as follows:

$$HV(x) = \frac{\sum_{j=1}^{wv_i} (y_{j2} - y_{j1})^2}{(1-2*a)*Iy} \quad (1)$$

where wv_i is the number of white runs (i,y_{j1})-(i,y_{j2}) for row $x=i$ in the range of $y=a*I_y\ldots$ $(1-a)*I_y$ and $HV(x) \in [0\ldots\ (1-2*a)*I_y]$. An example of the vertical white run projections compared to classical vertical white pixel projections is given in Fig. 5. In this figure it is demonstrated that using white run projections we can better discriminate text from non-text vertical zones.



— 20 —

« La bière est la boisson de l'ouvrier; « c'est lui qui, au milieu de ses pénibles « travaux, en fait la plus grande consommation, etc. »

M. Delcour a ajouté que l'augmentation du droit d'accise était impopulaire et qu'elle était préjudiciable à l'industrie agricole.

M. Delcour, comme M. De Naeyer, ne trouve pas difficile, par un remaniement, de combler un déficit de 3 millions provenant du dégrèvement de ce droit.

M. Vermeire a dit aussi à la Chambre : « La classe ouvrière industrielle se trouve « dans des conditions telles que la bière est « devenue pour elle un aliment indispensable. »

Nous aimons à enregistrer ces paroles, parce qu'elles ont été prononcées par des hommes considérables qui, au moment où nous écrivons, occupent le pouvoir.

×

En présence de l'expression de pareils sentiments, il est une chose triste à constater : c'est que les hommes d'Etat n'ont jamais tenu ce langage que dans les mo-

— 21 —

ments d'agitation populaire, ou bien lorsqu'ils n'étaient plus au pouvoir.

Si la bière est la boisson alimentaire du peuple, pourquoi est-elle donc imposée? Lorsqu'on demande la suppression de cet impôt, les gouvernements répondent : Donnez-nous un autre impôt équivalent?

Mon Dieu, de combien s'agit-il? de quelques millions en France et en Belgique; c'est une bagatelle qu'on retrouverait facilement, comme l'ont dit MM. De Naeyer et Delcour.

Pour montrer le mauvais vouloir de nos gouvernants à l'égard de la bière, il suffira de citer ce fait, qu'avant la guerre de 1870-71, le gouvernement français, à une demande de dégrèvement sur l'impôt de la bière, aurait répondu cette phrase stéréotypée : « Donnez-nous un impôt équivalent. »

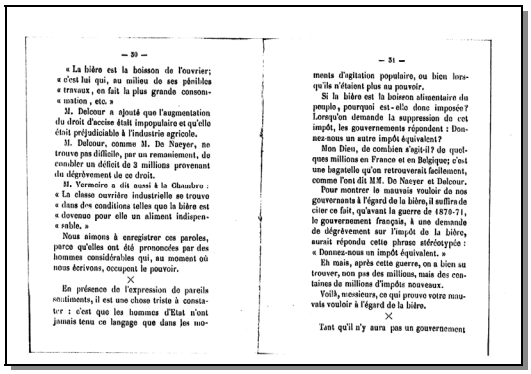
Eh mais, après cette guerre, on a bien su trouver, non pas des millions, mais des centaines de millions d'impôts nouveaux.

Voilà, messieurs, ce qui prouve votre mauvais vouloir à l'égard de la bière.

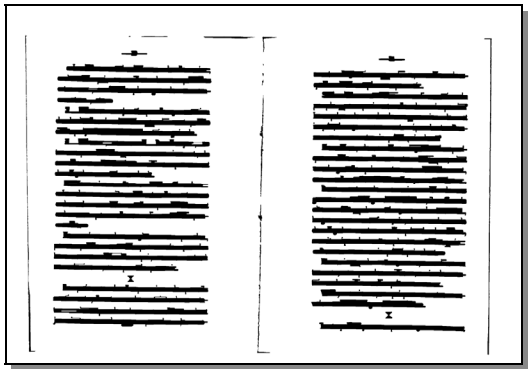
×

Tant qu'il n'y aura pas un gouvernement

(a)

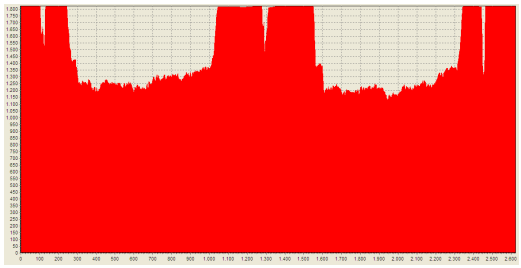


(b)

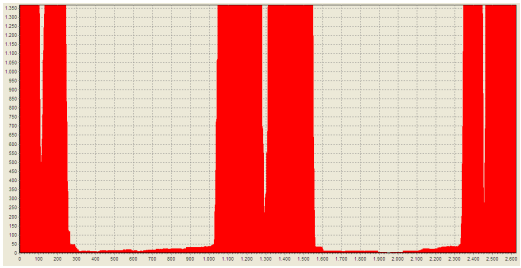


(c)

Figure 4. Pre-processing steps: (a) original image; (b) binary image; (c) smoothed image.



(a)



(b)

Figure 5. (a) Vertical white pixel projections and (b) vertical white run projections of image in Fig.4(c).

We can safely consider that if

$$HV(x)<(1-2*a)I_y/2 \quad (2)$$

then the corresponding image column may contain page information. Also, if we have consecutive n image columns that satisfy eq. (2) with $n>b*I_x$ then a vertical page zone is detected. For our experiments, we consider b=1/6. In Fig.6 we present an image example where two page zones have been detected. In this example, the left vertical zone has $n=n_1>b*I_x$ and the right vertical zone $n=n_2>b*I_x$.

Although detecting two vertical page zones is the most common case, we also have the cases where more than two, just one or no vertical page zones are detected. In the following, we examine all these cases.

(i) When we detect two vertical page zones, the x-offsets of the left one are assigned to *xL1*, *xL2* and the x-offsets of the right one to *xR1*, *xR2*.

(ii) We detect just one vertical page zone when one of the two pages is empty or has very little information. In this case, we approximate the other region based on the coordinates of the detected region. If we have detected one vertical zone defined by x-offsets *x1* and *x2* then:

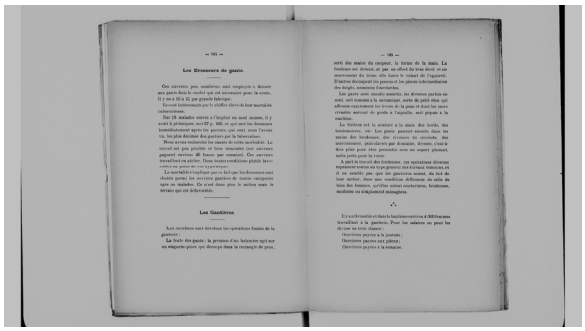
$$xL_1=\begin{cases} x_1\ , & \text{if } x_1<\frac{I_x}{2} \\ I_x-x_2\ , & \text{otherwise} \end{cases},\quad xL_2=\begin{cases} x_2\ , & \text{if } x_1<\frac{I_x}{2} \\ I_x-x_1\ , & \text{otherwise} \end{cases}$$

$$xR_1=\begin{cases} x_1\ , & \text{if } x_1>\frac{I_x}{2} \\ I_x-x_2\ , & \text{otherwise} \end{cases},\quad xR_2=\begin{cases} x_2\ , & \text{if } x_1>\frac{I_x}{2} \\ I_x-x_1\ , & \text{otherwise} \end{cases} \quad (3)$$

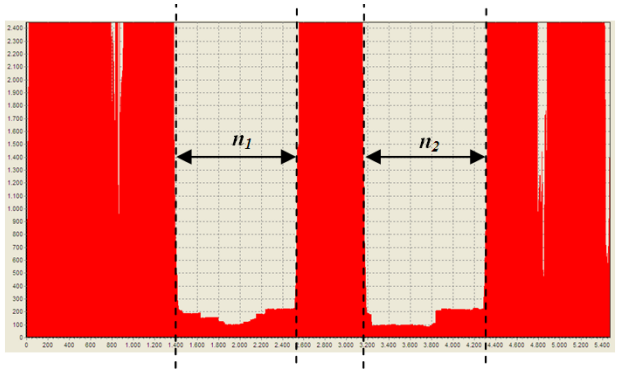
(iii) We detect more than two or no vertical page zones when we have a multi column or complex documents (e.g. magazines). In this case, we search for a vertical split line SL_x near the middle of the page. SL_x is set to the x-offset which corresponds to the longest white vertical stripe near $I_x/2$. SL_x is calculated as follows:

$$SL_x=\underset{x\in[\frac{I_x}{2}-c*I_x,\frac{I_x}{2}+c*I_x]}{arg\,max}HV(x) \quad (4)$$

where parameter c defines the distance from the middle of the page. For our experiments, we consider c=1/6. Then, the x-offsets of the vertical page zones are defined as follows:



(a)



(b)

Figure 6. (a) Original image;(b) Vertical white run projections *HV(x)* and the two detected vertical page zones.

$$xL_1=max(\underset{x\in[0,d*I_x]}{arg\,max}HV(x)),$$

$$xL_2=min(\underset{x\in[SL_x-e*I_x,SL_x]}{arg\,max}HV(x)),$$

$$xR_1=max(\underset{x\in[SL_x,SL_x+e*I_x]}{arg\,max}HV(x)), \quad (5)$$

$$xR_2=min(\underset{x\in[I_x-d*I_x,I_x]}{arg\,max}HV(x))$$

where parameter d defines the maximum distance from image boundaries that we expect that page starts and parameter e the maximum distance from the page middle. For our experiments, we consider $d=e=1/6$.

For the cases (i) and (ii) we finally proceed to a better adjustment of x-coordinates in order to preserve text regions that may fall outside the detected vertical zones. This is accomplished by using the following formulas:

$$xL_1 = max(\underset{x\in[\,xL_1-d*I_x\,,xL_1\,]}{arg\,max}\ HV(x))\,,$$

$$xL_2 = min(\underset{x\in[\,xL_2\,,xL_2+e*I_x\,]}{arg\,max}\ HV(x))\,,$$

$$xR_1 = max(\underset{x\in[\,xR_1-d*I_x\,,xR_1\,]}{arg\,max}\ HV(x))\,, \qquad (6)$$

$$xR_2 = min(\underset{x\in[\,xR_2\,,xR_2+e*I_x\,]}{arg\,max}\ HV(x))$$

An example of better adjusting xR_2 value is demonstrated in Fig. 7.

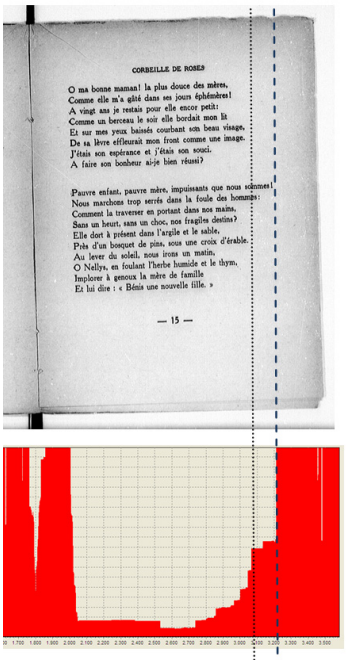


Figure 7. An example of xR_2 value before (vertical line "…") and after (vertical line "---") adjustment.

2.4 Detection of Horizontal Page Zones

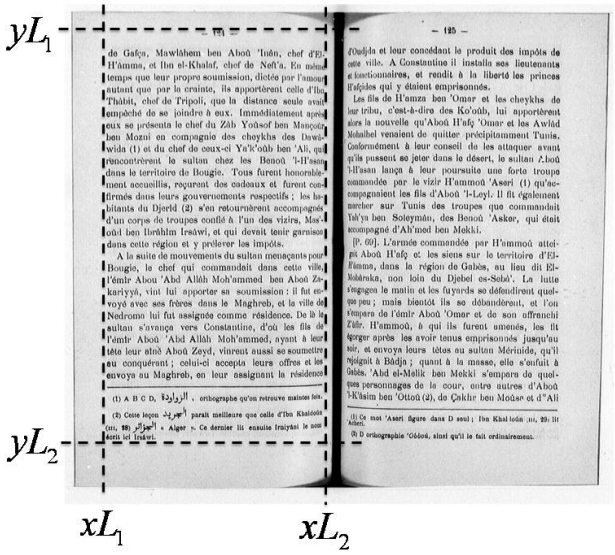
In this stage, we aim to define the horizontal page zones that are defined by offsets yL_1, yL_2 for the left page and yR_1, yR_2 for the right page, respectively (see Fig.3(a)).

First, we calculate the horizontal white run projections $HH_1(y)$ and $HH_2(y)$ in the same way we calculated the vertical white run projections. $HH_1(y)$ corresponds to the left vertical zone page and is calculated in the interval of xL_1 and xL_2 while $HH_2(y)$ corresponds to the right vertical zone page and is calculated in the interval of xR_1 and xR_2.

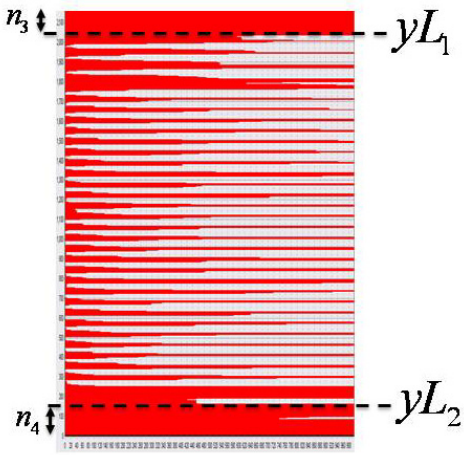
We can safely consider that if

$$HH_1(y) > f*(xL_2-xL_1), \text{ for the left page}$$
$$HH_2(y) > f*(xR_2-xR_1), \text{ for the right page} \qquad (7)$$

then the corresponding image row may belong to the border areas around text areas. We search into the intervals $[0,\ g*I_y]$ and $[(1-g)*I_y,\ I_y]$ in order to define the top and bottom border areas, respectively. If we have consecutive n image rows that satisfy eq. (7) with $n>h*I_y$ then a top or bottom border area is detected. The y-offsets of the top border areas are assigned to yL_1 and yR_1 and the y-offsets of the bottom border areas are assigned to yL_2 and yR_2. yL_1 and yR_1 are set to 0 if no top border areas are detected. Similarly, yL_2 and yR_2 are set to I_y if no bottom border areas are detected. For our experiments, we consider f=2/3, g=1/4 and h=1/25. In Fig.8 we present an image example where the top and bottom border areas of the left vertical zone have been detected. In this example, the top border area has $n=n_3>h*I_y$ and the bottom border area $n= n_4>h*I_y$.



(a)



(b)

Figure 8. Detection of horizontal page zones (a) Original image and its left vertical and horizontal zones; (b) Horizontal white run projections $HH_1(y)$ of the left vertical zone and the two detected border areas.

3. EXPERIMENTAL RESULTS

The performance evaluation method used is based on a pixel based approach and counts the pixels at the correct page frames and the detected page frames. For this purpose, we manually mark the correct page frames in the original double page document image in order to create the ground truth set. Fig. 9 shows an example of a manually marked document image.

Let G be the set of all pixels inside the correct page frame in ground truth, R the set of all pixels inside the result page frame and $T(s)$ a function that counts the elements of set s. We calculate the Precision and Recall as follows:

$$Precision = \frac{T(G \cap R)}{T(R)} \tag{8}$$

and

$$Recall = \frac{T(G \cap R)}{T(G)} \tag{9}$$

A performance metric FM can be extracted if we combine the values of precision and recall:

$$FM = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{10}$$

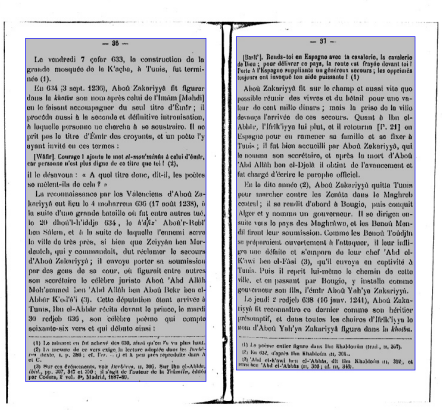


Figure 9. Manually marked page frames (ground truth).

To verify the validity of the proposed methodology we used 458 double page document images from 15 different historical books that are owned by the Bibliothèque Nationale de France (BNF) [2]. In Table 1 all the historical books that are used in our experiments are presented. In order to calculate the Precision, the Recall and the FM of a double page image we calculate the values for each individual page frame and then we extract the average values of them.

Tables 2 illustrates the evaluation results of each book after applying only the first step of the proposed methodology in which the vertical zones of the two pages are detected. As horizontal zones of the page frames we used the ground truth horizontal zones. Table 3 illustrates the overall evaluation results after applying both steps of the proposed methodology.

Table 1. Historical books that are used in our experiments

Book ID	Title
1	Almohades & Des Hafçides
2	Corbeille de Roses
3	Correspondance D' Eulalie
4	Dr Jullien. L Industrie des gants, étude d hygiène professionnelle et de médecine légale
5	Fantasio
6	L Art (Paris. 1865)
7	L' Expérience Italienne
8	La Bière De L' Avenir
9	La Nouvelle Revue
10	La Préehistoire Du Christianisme
11	La Sicile, souvenirs, récits et légendes
12	Le Chauffeur Est A Vos Ordres
13	Pseavmes De David
14	Société royale des sciences, lettres et arts de Nancy
15	Figures des Histoires de la Saincte Bible

Table 2. Evaluation results after detecting only the vertical zones of the two pages

Book ID	#Images	Precision (%)	Recall (%)	FM (%)
1	53	99.45	99.50	99.47
2	22	99.97	99.83	99.90
3	15	94.92	100.0	97.39
4	10	99.31	98.92	99.11
5	27	97.50	92.95	95.17
6	47	100.0	98.99	99.49
7	25	100.0	100.0	100.0
8	19	100.0	100.0	100.0
9	47	99.91	98.92	99.41
10	30	100.0	100.0	100.0
11	21	99.84	99.82	99.83
12	28	99.74	100.0	99.87
13	28	96.68	99.73	98.18
14	62	100.0	100.0	100.0
15	24	100.0	100.0	100.0
TOTAL	458	99.15	99.24	99.20

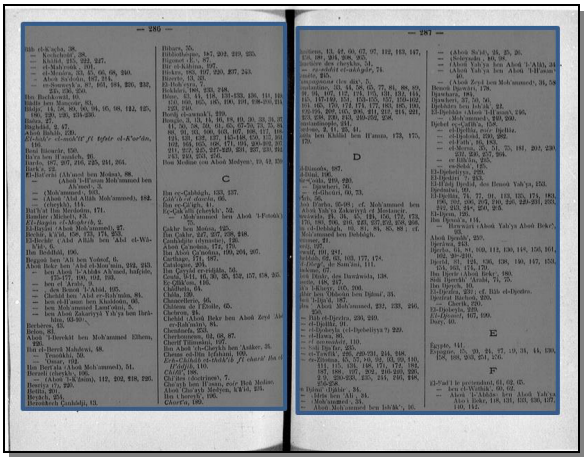
Table 3. Overall Evaluation Results of the proposed methodology

Book ID	#Images	Precision (%)	Recall (%)	FM (%)
1	53	98.96	99.47	99.21
2	22	99.82	99.80	99.81
3	15	94.78	99.50	97.08
4	10	99.35	98.63	98.99
5	27	97.51	92.32	94.84
6	47	99.98	98.84	99.41
7	25	99.52	99.95	99.73
8	19	99.50	99.97	99.73
9	47	99.88	98.75	99.31
10	30	99.43	99.93	99.68
11	21	99.84	98.74	99.29
12	28	99.34	99.44	99.39
13	28	96.63	99.69	99.14
14	62	100.0	99.87	99.93
15	24	100.0	99.97	99.98
TOTAL	458	98.97	98.99	98.98

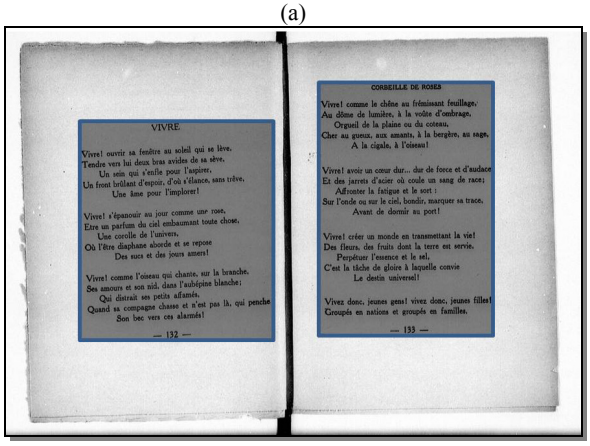
The evaluation results of both steps demonstrate the success of the proposed methodology which detects the page frames with great accuracy. Precision and recall are about 99%, which indicates that the proposed methodology removes the noisy black borders of the document images without crop page information. Some representative results are shown in Fig. 10. We want to stress that although a lot of parameters are involved in our methodology (a=1/8, b=c=d=1/6, f=2/3, g=1/4 and h=1/25) we have safely assigned values to them since the methodology has been proved efficient for a great variety of books.

As it can be observed, Book 5 ("Fantasio") achieves the lowest recall value (92.32%). Document images of this book have complex layout with multiply columns and images. Fig. 10(c) presents a representative double page document image of Book 5 in which the left page frame has been cropped.

For comparisons purposes, we also applied the commercial package Scan Tailor [9] at the same dataset. However, the pixels of the original image do not remain the same so the pixel based evaluation approach cannot be applied. For this reason, we only present some representative results using Scan Tailor (see Fig. 11 and 12). As it is observed, although Scan Tailor splits the image into two pages successfully in several examples; however in many cases the noisy black borders cannot be efficiently removed.

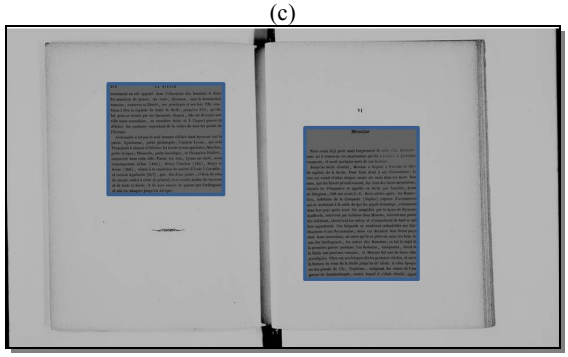


(a)



(b)

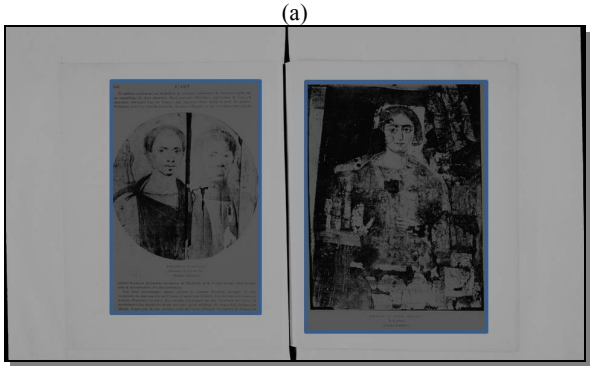
(c)



(d)

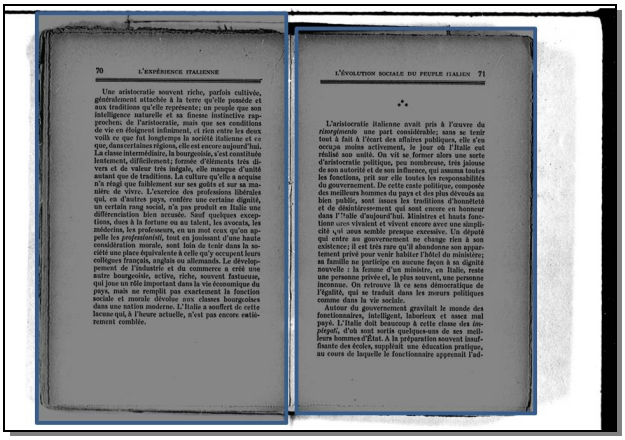
Figure 10. Example images from Books (a) ID=1 (b) ID=2 (c) ID=5 (d) ID=11 showing the page frames detection using the proposed methodology.

(a)

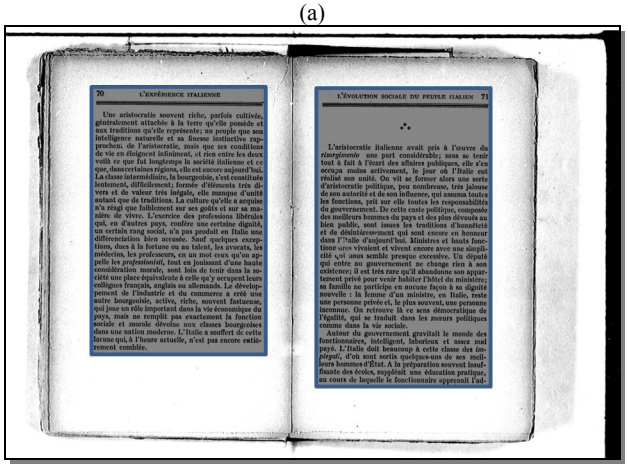


(b)

Figure 11. An example image from Book ID=6 showing the page frames detection using (a) Scan Tailor and (b) the proposed methodology.



(a)



(b)

Figure 12. An example image from Book ID=7 showing the page frames detection using (a) Scan Tailor in which noise black borders cannot be removed and (b) the proposed methodology.

4. CONCLUSION

A novel methodology has been proposed in order to detect the optimal page frames of double page document images. The proposed methodology is based on the vertical and horizontal white run projections which have been proved efficient for detecting zones of text areas. Our aim is to split the image into the two pages as well as to remove noisy borders. Experimental results on several double page document images from fifteen different books demonstrate the effectiveness of the proposed technique. Precision and recall are about 99%, which indicates that the proposed methodology removes the noisy black borders of the document images without crop page information.

5. ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Community's Seventh Framework Programme under grant agreement n° 215064 (project IMPACT).

6. REFERENCES

[1] Avila, B.T. and Lins, R.D., 2004, A New Algorithm for Removing Noisy Borders from Monochromatic Documents, Proc. of ACM-SAC'2004, Cyprus, ACM Press, 1219-1225.

[2] Bibliothèque nationale de France: http://www.bnf.fr

[3] Fan, K.C., Wang, Y.K., Lay, T.R., 2002. Marginal Noise Removal of Document Images, Pattern Recognition, 35(11), 2593-2611.

[4] Gatos, B., Pratikakis, I., and Perantonis, S.J., 2006. Adaptive Degraded Document Image Binarization. Pattern Recognition, 39, 317-327.

[5] Highly automated digitisation of books and journals: The METADATA ENGINE Project! http://meta-e.aib.uni-linz.ac.at/

[6] Konidaris, T., Gatos, B., Ntzios, K., Pratikakis, I., Theodoridis, S. and Perantonis, S.J., 2007. Keyword-Guided Word Spotting in Historical Printed Documents Using Synthetic Data and User feedback. International Journal on Document Analysis and Recognition (IJDAR), special issue on historical documents 9, 2-4, 167-177.

[7] Le, D.X., Thoma, G.R., 1996. Automated Borders Detection and Adaptive Segmentation for Binary Document Images, International Conference on Pattern Recognition, pp. III: 737-741.

[8] Nikolaou, N., Makridis, M., Gatos, B., Stamatopoulos, N., and Papamarkos, N., 2010. Segmentation of historical machine-printed documents using Adaptive Run Length Smoothing and skeleton segmentation paths. Image and Vision Computing, vol. 28, no. 4, 590-604.

[9] Scan Tailor: http://scantailor.sourceforge.net/

[10] Shafait, F., Beusekom, J., Keysers, D., Breuel, T.M., 2008. Document cleanup using page frame detection, Int. Jour. on Document Analysis and Recognition, vol. 11, no. 2, 81–96.

[11] Stamatopoulos, N., Gatos, B., and Kesidis, A., 2007. Automatic Borders Detection of Camera Document Images. In 2nd Int. Workshop on Camera-Based Document Analysis and Recognition, Curitiba, Brazil, 71-78.

[12] WiseBook: http://www.cadcam.org/wise-book.php

[13] Yacoub, S., Burns, J., Faraboschi, P., Ortega D., Peiro, J.A., Saxena, V., 2005. Document Digitization Lifecycle for Complex Magazine Collection, Proceedings of the ACM symposium on Document engineering, 197-206.