2011 International Conference on Document Analysis and Recognition

A Novel Skew Detection Technique Based on Vertical Projections

A. Papandreou and B. Gatos

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

Abstract— **Document skew detection is often done by the use of horizontal projections. In this paper, we introduce a new document skew detection approach that is based on vertical projections as well as bounding box minimization criterion. We motivated by the fact that the majority of the Latin characters have vertical strokes. We claim that the proposed approach is more efficient and gives more accurate results compared with the state-of-the-art skew detection algorithm based on horizontal projections since it is noise and warp resistant. For these reasons, it can be efficiently applied to historical machine printed documents. Experimental results on a database with representative historical printed documents prove the efficiency of the proposed approach. Moreover, an analysis of the performance of the main elements of the proposed technique is also done.**

Keywords – Document Skew Detection; Vertical Image Projections; Historical Document Image Processing

I. INTRODUCTION

One of the most important document image analysis step is the detection and correction of its skew. Some degree of skew is unavoidable either a paper is scanned manually or mechanically. This results in a skewed image. A reliable skew estimation and correction technique has to be used in scanned documents as a pre-processing stage in almost all document analysis and recognition systems. In the literature, a variety of skew detection and correction techniques are available. Skew estimation approaches are classified into four main categories according to the basic approach they adopt [1]. It includes Hough transform [2], projection profile [3-7], nearest neighbor clustering [8] and interline cross correlation [10].

The traditional projection profile approach is a simple solution to detect skew angle of a document image. It is initially proposed by Postl [3] and is based on horizontal projection profile. According to this approach, a series of horizontal projection profiles are calculated at a range of angles. The profile with maximum variation refers the best alignment to the text lines. At this stage, projection angle is the actual skew angle of the skewed document. In order to reduce high computational costs, several variations of this basic method have been proposed. Baird [4] proposes a technique for selecting the points to be projected: for each connected component the midpoint of the bottom side of the bounding box is projected. The objective function is to compute the sum of the squares of the profiles. Ciardiello [5] projected selected sub-region (one with high density of black pixels per row) of the document image; the function is to maximize the mean square deviation of the profile. Ishitani [6] uses a profile which is defined in a different style. A cluster of parallel lines on the image is selected and the bins of the profile store the number of black/white transitions along the lines. Bloomberg [7] reduce computational complexity of projection profile based approach by extracting a sample image in the skewed document. Skew angle is calculated by sample image rather than whole document and this results in a faster skew estimation method.

To conclude, projection profile based approaches are computationally expensive as different projections are calculated at different angles in a particular range. Additionally, projection profile methods are limited to estimate skew angle within ±10° to 15° [4] [9]. Moreover, the accuracy depends upon the angular resolution of the projection profile. Finally, projection profile based approaches cannot deal with noisy documents and broken characters [9].

After examining all the skew detection methodologies that are based on projection profiles we observe that only horizontal projections are used. In this work, we are motivated from the fact that the majority of the Latin characters have vertical strokes and we propose a new skew correction approach based on vertical projections that are noise and warp resistant.

II. PROPOSED METHODOLOGY

The horizontal projection profile is based on the histogram of black pixels along horizontal scan-lines. For a script with horizontal text lines, the horizontal projection profile will have peaks at text line positions and troughs at positions in between successive text lines [10]. This concludes to the fact that any noise and warp will ruin those peaks and troughs of the horizontal projection histogram and the efficiency of this technique. The proposed skew estimation methodology is based on vertical projections, which we claim to be noise and warp resistant. In the vertical approach, in a script with horizontal orientation, we don't expect any troughs but we seek for higher density of black pixels, concentrated in certain columns when the text is properly aligned. This is due to the nature of the Latin alphabet, with the majority of its characters (33/52) having at least one major vertical oriented stroke (see Fig.1).

1520-5363/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDAR.2011.85

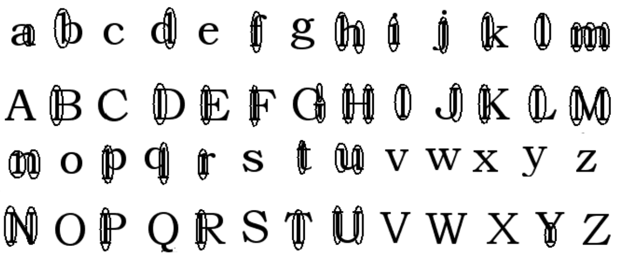


Figure 1. The latin alphabet with the major vertical strokes circled. There are secondary vertical strokes which also contribute in the performance of the proposed method.

According to relative frequency of Latin alphabet characters in several languages (English [11], Spanish [12], French [13] and German [14], see Table I), we jumped into the conclusion that over 61% of the characters in a random text have at least one major vertical stroke.

TABLE I. RELATIVE FREQUENCY OF LATIN ALPHABET CHARACTERS IN THE MOST COMMON LANGUAGES

Characters of Latin Alphabet	Common Languages with Latin Alphabet			
	English	*Spanish*	*French*	*German*
A	8.167%	7.636%	6.51%	12.53%
B	1.492%	0.901%	1.89%	1.42%
C	2.782%	3.260%	3.06%	4.68%
D	4.253%	3.669%	5.08%	5.86%
E	12.702%	14.715%	17.40%	13.68%
F	2.228%	1.066%	1.66%	0.69%
G	2.015%	0.866%	3.01%	1.01%
H	6.094%	0.737%	4.76%	0.70%
I	6.966%	7.529%	7.55%	6.25%
J	0.153%	0.545%	0.27%	0.44%
K	0.772%	0.049%	1.21%	0.01%
L	4.025%	5.456%	3.44%	4.97%
M	2.406%	2.968%	2.53%	3.15%
N	6.749%	7.095%	9.78%	6.71%
O	7.507%	5.378%	2.51%	8.68%
P	1.929%	3.021%	0.79%	2.51%
Q	0.095%	1.362%	0.02%	0.88%
R	5.987%	6.553%	7.00%	6.87%
S	6.327%	7.948%	7.27%	7.98%
T	9.056%	7.244%	6.15%	4.63%
U	2.758%	6.311%	4.35%	3.93%
V	0.978%	1.628%	0.67%	0.90%
W	2.360%	0.114%	1.89%	0.02%
X	0.150%	0.387%	0.03%	0.22%
Y	1.974%	0.308%	0.04%	0.90%
Z	0.074%	0.136%	1.13%	0.52%
Characters with Major Vertical Strokes	**62.42%**	**61.51%**	**62.4%**	**61.02%**

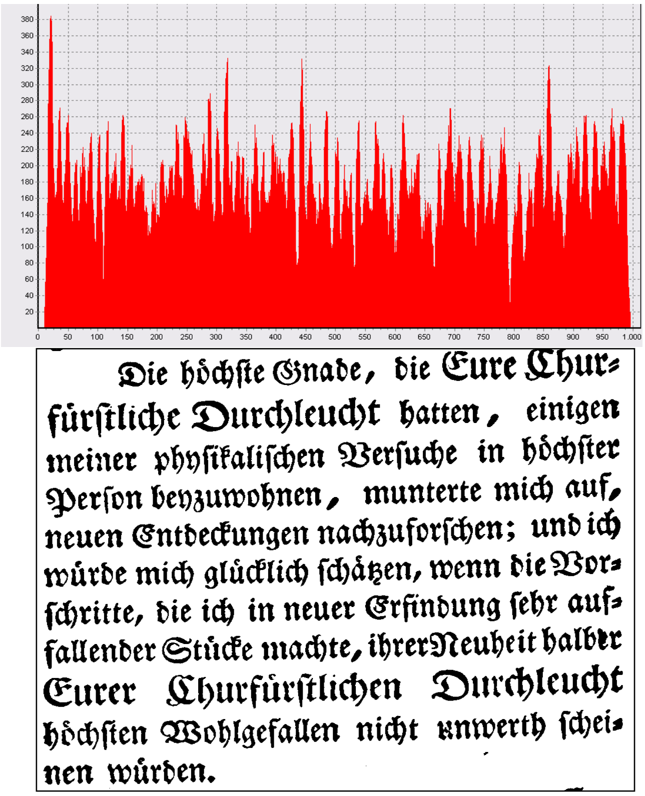
In this way it can be observed that by summing the black pixels of each column and computing the energy at a number of angles that we rotate the document, the higher energy will be matched with the correct skew angle in which the vertical strokes of all rows are aligned and so we have bigger concentration of black pixels in certain columns. This concludes in the estimation of the skew angle. In order to compute the highest concentration of black pixels we detect the angle in which the energy function $A(\theta)$ is maximized:

$$A(\theta) = \sum_{i=1}^{m} c_i^2(\theta) \qquad (1)$$

where m is the width of the image and $c_i(\theta)$ is the number of black pixels counted in the i-th column while the document was rotated by angle θ.

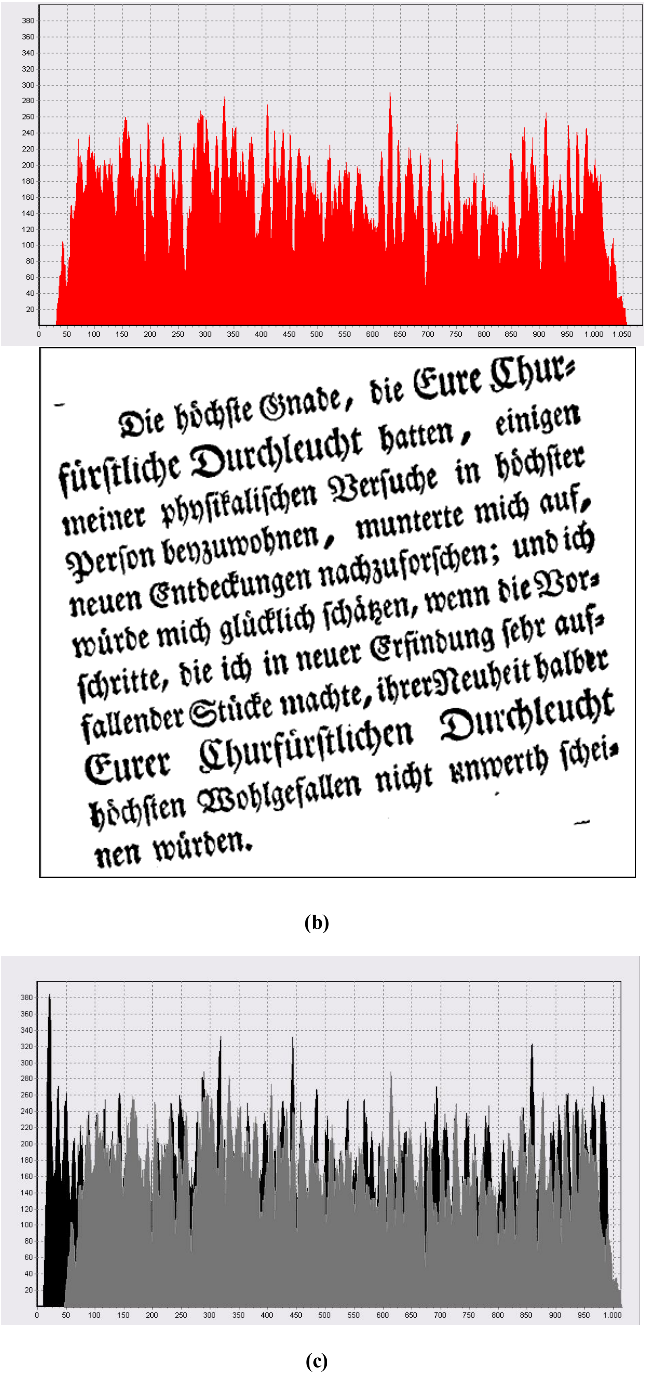
As you can see from the histograms in Fig.2, when the document is properly aligned (see Fig.2a) great peaks are observed in certain columns, which show great concentration of black pixels and consequently vertical-stroke alignment. On the other hand when there is skew in the document all of the peaks will be average (see Fig.2b). It is observed that in documents as the number of the rows, more and more vertical strokes will be aligned in the correct skew angle. That results in even greater difference between the peaks of the histogram. Also the black border of a scanned page might be formed of vertical lines and this will also contribute in deskewing the document when the proposed method is applied (see Fig.3).

Since we are not expecting or depending on any troughs, noise will not affect our results, if it is relatively uniformly distributed, and not in random lines. Moreover warp, which is majorly affecting the troughs in the horizontal projection profiles, has no impact in our method, since the vertical strokes remain almost inviolated.



Die höchſte Gnade, die Eure Churfürſtliche Durchleucht hatten, einigen meiner phyſikaliſchen Verſuche in höchſter Perſon beyzuwohnen, munterte mich auf, neuen Entdeckungen nachzuforſchen; und ich würde mich glücklich ſchätzen, wenn die Vorſchritte, die ich in neuer Erfindung ſehr auffallender Stücke machte, ihrer Neuheit halber Eurer Churfürſtlichen Durchleucht höchſten Wohlgefallen nicht unwerth ſcheinen würden.

(a)



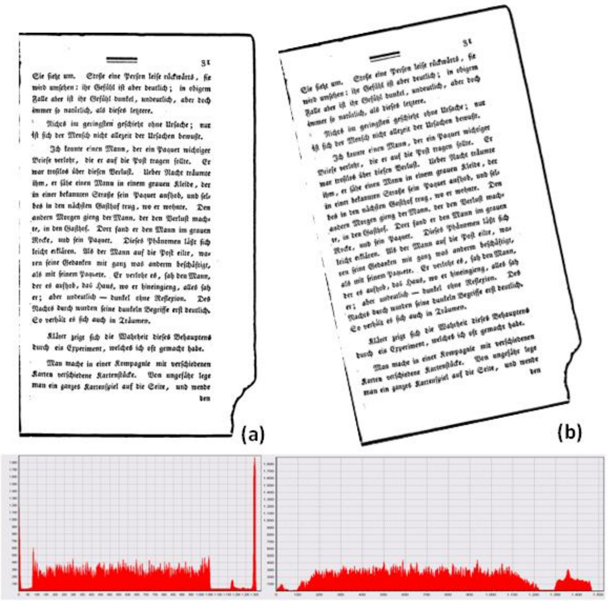
Die höchſte Gnade, die Eure Churfürſtliche Durchleucht hatten, einigen meiner phyſikaliſchen Verſuche in höchſter Perſon beyzuwohnen, munterte mich auf, neuen Entdeckungen nachzuforſchen; und ich würde mich glücklich ſchätzen, wenn die Vorſchritte, die ich in neuer Erfindung ſehr auffallender Stücke machte, ihrer Neuheit halber Eurer Churfürſtlichen Durchleucht höchſten Wohlgefallen nicht unwerth ſcheinen würden.

(b)

(c)

Figure 2. Sample of german historical document with significant warping (a) vertical histogramm / document with no skew, (b) vertical histogram / document with 5 degrees skew, (c) comparison of histograms of a and b.

Additionally, we improved our vertical projection algorithm by combing it with a bounding box approach. Techniques based on the same idea are common is skew detection [15]. We estimate the area of the rectangle that includes the 4 extreme points of the text and then we use it to divide the energy that we calculate from eq.1.



31

Sie ſiehe um. Streife eine Perſon leiſe rückwärts, ſie wird umſehen: ihr Gefühl iſt aber deutlich; in obigem Falle aber iſt ihr Gefühl dunkel, undeutlich, aber doch immer ſo natürlich, als dieſes leztere.

Nichts im geringſten geſchieht ohne Urſache; nur iſt ſich der Menſch nicht allezeit der Urſachen bewußt.

Ich kannte einen Mann, der ein Paquet wichtiger Briefe verlohr, die er auf die Poſt tragen ſollte. Er war troſtlos über dieſen Verluſt. Ueber Nacht träumte ihm, er ſähe einen Mann in einem grauen Kleide, der in einer bekannten Straße ſein Paquet aufhob, und ſelbes in den nächſten Gaſthof trug, wo er wohnte. Den andern Morgen gieng der Mann, der den Verluſt machte, in den Gaſthof. Dort fand er den Mann im grauen Rocke, und ſein Paquet. Dieſes Phänomen läßt ſich leicht erklären. Als der Mann auf die Poſt eilte, waren ſeine Gedanken mit ganz was anderm beſchäftigt, als mit ſeinem Paquete. Er verlohr es, ſah den Mann, der es aufhob, das Haus, wo er hineingieng, alles ſah er; aber undeutlich — dunkel ohne Reflexion. Des Nachts durch wurden ſeine dunkeln Begriffe erſt deutlich. So verhält es ſich auch in Träumen.

Klärer zeigt ſich die Wahrheit dieſes Behauptens durch ein Experiment, welches ich oft gemacht habe.

Man mache in einer Kompagnie mit verſchiedenen Karten verſchiedene Kartenſtöcke. Von ungefähr lege man ein ganzes Kartenſpiel auf die Seite, und wende den

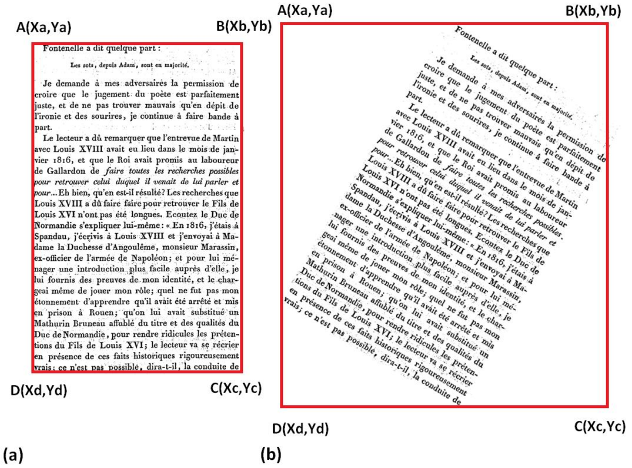
(a)

(b)

Figure 3. The histograms of a deskewd (a) and a skewed (b) image show the contribution of the black borders of the page when it's aligned.

$$T(\theta) = \frac{A(\theta)}{0.3(X_a - X_b)(Y_a - Y_d)} \qquad (2)$$

where X_a, X_b, Y_a and Y_d are the coordinates of the extreme points of our bounding box (see Fig.4), while 0.3 is the specified weight that we use to adjust the contribution of each technique.



Fontenelle a dit quelque part :

Les sots, depuis Adam, sont en majorité.

Je demande à mes adversaires la permission de croire que le jugement du poète est parfaitement juste, et de ne pas trouver mauvais qu'en dépit de l'ironie et des sourires, je continue à faire bande à part.

Le lecteur a dû remarquer que l'entrevue de Martin avec Louis XVIII avait eu lieu dans le mois de janvier 1816, et que le Roi avait promis au laboureur de Gallardon de *faire toutes les recherches possibles pour retrouver celui duquel il venait de lui parler et pour*... Eh bien, qu'en est-il résulté? Les recherches que Louis XVIII a dû faire faire pour retrouver le Fils de Louis XVI n'ont pas été longues. Ecoutez le Duc de Normandie s'expliquer lui-même : « En 1816, j'étais à Spandau, j'écrivis à Louis XVIII et j'envoyai à Madame la Duchesse d'Angoulême, monsieur Marassin, ex-officier de l'armée de Napoléon; et pour lui ménager une introduction plus facile auprès d'elle, je lui fournis des preuves de mon identité, et le chargeai même de jouer mon rôle; quel ne fut pas mon étonnement d'apprendre qu'il avait été arrêté et mis en prison à Rouen; qu'on lui avait substitué un Mathurin Bruneau affublé du titre et des qualités du Duc de Normandie, pour rendre ridicules les prétentions du Fils de Louis XVI; le lecteur va se récrier en présence de ces faits historiques rigoureusement vrais; ce n'est pas possible, dira-t-il, la conduite de

Figure 4. The area of the bounding box ((*Xa*-*Xb*)*(*Ya*-*Yd*)) that includes the text is minimum when the image is aligned.

Since the area of the boundary box is minimum for a text with no skew and the energy maximum, when we compute their ratio, $T(\theta)$ maximizes.

III. Experimental Results

In order to test the proposed methodology two historical books were used. One from Eckartshausen, which was published in 1788 and is owned by the Bavarian State Library [16], and a French historical book, which was published in 1838, and is owned by the Bibliothèque nationale de France [17]. Some of the images of the book suffer from several problems such as warp, noise and degradations (see Fig. 5 and 6).

We rotated the scanned printed document images in different angles and we estimated the skew with 4 different algorithms. A horizontal projections profile methodology based on [3], the horizontal projection profile methodology combined with the minimization of the bounding box technique, the proposed vertical projection profiles approach and the proposed vertical projections profile algorithm combined with minimization of the bounding box.

A. *Testing in Fullpage Images*

Our test-set was formed by 144 simply scanned and binarized document images, with the page borders, noise and blank spots included. Neither segmentation, nor denoising or dewarping or any other pre-process took place. We rotated each image for 11 different angles, from -5° to 5° with 1° step, while the accuracy that the algorithms presumed adequate was 0.2°, as their step in the successive rotations in order to estimate the skew was also 0.2°. The results presented are from a 1584 different estimations for each of the algorithms (see Table II).

TABLE II. Results From Degraded Full-Page Images

Different Skew Estimation Techniques	***Error Deviation in Degrees***	***Percent of Correct Estimation***
Simple Vertical Projection Profile Technique	**0.17**	**54.42%**
Simple Horizontal Projection Profile Technique	0.26	36.87%
Combined Vertical Projretion Profile Technique	**0.14**	**58.46%**
Combined Horizontal Projretion Profile Technique	0.23	40.65%

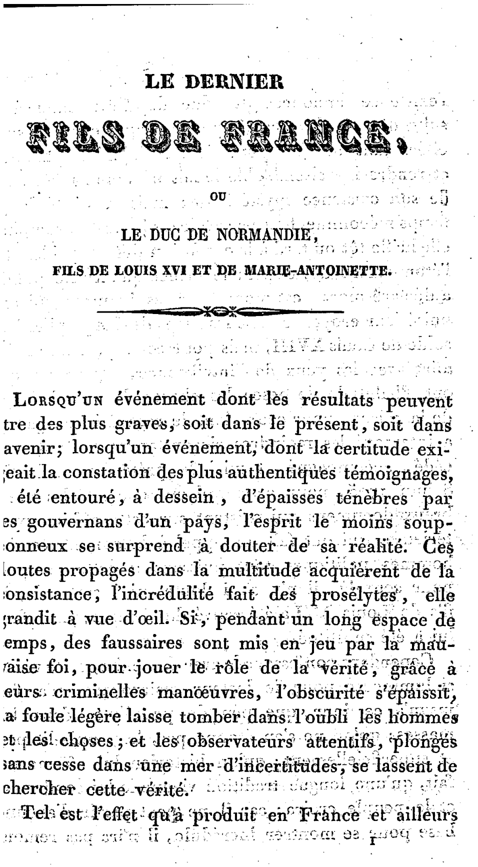
B. *Testing only in Text Region of the Images*

Afterwards the images of the test–set were segmented excluding the page borders. We also rotated each image for 11 different angles, from -5° to 5° with 1° step, while the accuracy that the algorithms presumed adequate was 0.2°, as their step in the successive rotations in order to estimate skew was 0.2°. The results presented are from a 1584 different estimations for each of the algorithms (see Table III).

TABLE III. Results From Degraded Segmented Images

Different Skew Estimation Techniques	***Error Deviation in Degrees***	***Percent of Correct Estimation***
Simple Vertical Projection Profile Technique	**0.27**	**40.35%**
Simple Horizontal Projection Profile Technique	0.40	27.65%
Combined Vertical Projretion Profile Technique	**0.22**	**46.65%**
Combined Horizontal Projretion Profile Technique	0.40	28.85%

The results demonstrate that in degraded images with noise and warp the proposed method is more accurate than the horizontal projection profiles method. Also the contribution of the minimizing of the bounding box technique that we combined is clear in all cases. Last but not least it is easily observed that the borders of the scanned documents help in the accuracy of all the methods tested, since all algorithms had better results in full pages than in segmented documents.



LE DERNIER

FILS DE FRANCE,

OU

LE DUC DE NORMANDIE,

FILS DE LOUIS XVI ET DE MARIE-ANTOINETTE.

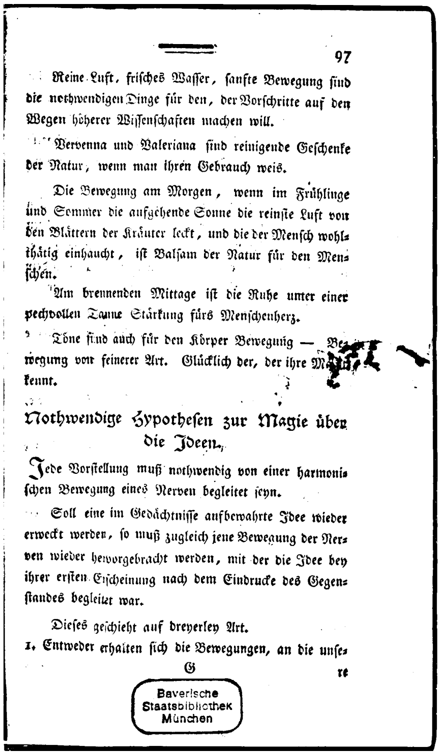
Lorsqu'un événement dont les résultats peuvent tre des plus graves, soit dans le présent, soit dans avenir; lorsqu'un événement, dont la certitude exi- ;eait la constation des plus authentiques témoignages, été entouré, à dessein, d'épaisses ténèbres par es gouvernans d'un pays, l'esprit le moins soup- onneux se surprend à douter de sa réalité. Ces loutes propagés dans la multitude acquièrent de la onsistance, l'incrédulité fait des prosélytes, elle ;randit à vue d'œil. Si, pendant un long espace de emps, des faussaires sont mis en jeu par la mau- raise foi, pour jouer le rôle de la vérité, grâce à eurs criminelles manœuvres, l'obscurité s'épaissit, a foule légère laisse tomber dans l'oubli les hommes et les choses; et les observateurs attentifs, plongés sans cesse dans une mer d'incertitudes, se lassent de chercher cette vérité.

Tel est l'effet qu'a produit en France et ailleurs

Figure 5. Sample of french historical document with significant noise and warp from our test-set.

IV. Conclusion

A new efficient skew detection technique is presented, that can be applied in degraded documents without any previous segmentation, denoising or dewarping. Our technique proves to be resistant in noise and warp in contradiction to what is claimed for projection profile based approaches [9]. Our technique has better and better results as more and more strokes from different rows are vertically aligned and as it is easily observed the black borders of the images contribute in the correct estimation of the proposed method since in segmented images the results are worse than in full unprocessed images. Furthermore the contribution of the minimization of the bounding box technique, which we combined with the projection profile approach, is clearly demonstrated, since it improves both horizontal and vertical projection profile approaches. In addition we observed that our vertical approach is complimentary to the horizontal, since it has excellent results in documents that the horizontal approach fails, and we expect even better results for smaller step of rotation in our algorithm, since accuracy is very important in the alignment of the vertical strokes, especially as the height of the document increases. Finally, we have to point out that since most of the documents scanned from a book suffer from warp, noise and the image includes the page outline, our technique should be more efficient, especially in historical books with major degradations.



97

Reine Luft, friſches Waſſer, ſanfte Bewegung ſind die nothwendigen Dinge für den, der Vorſchritte auf den Wegen höherer Wiſſenſchaften machen will.

Vervenna und Valeriana ſind reinigende Geſchenke der Natur, wenn man ihren Gebrauch weis.

Die Bewegung am Morgen, wenn im Frühlinge und Sommer die aufgehende Sonne die reinſte Luft von den Blättern der Kräuter leckt, und die der Menſch wohlthätig einhaucht, iſt Balſam der Natur für den Menſchen.

Am brennenden Mittage iſt die Ruhe unter einer pechvollen Tanne Stärkung fürs Menſchenherz.

Töne ſind auch für den Körper Bewegung — Bewegung von feinerer Art. Glücklich der, der ihre Macht kennt.

Nothwendige Hypotheſen zur Magie über die Ideen.

Jede Vorſtellung muß nothwendig von einer harmoniſchen Bewegung eines Nerven begleitet ſeyn.

Soll eine im Gedächtniſſe aufbewahrte Idee wieder erweckt werden, ſo muß zugleich jene Bewegung der Nerven wieder hervorgebracht werden, mit der die Idee bey ihrer erſten Erſcheinung nach dem Eindrucke des Gegenſtandes begleitet war.

Dieſes geſchieht auf dreyerley Art.

1. Entweder erhalten ſich die Bewegungen, an die unſe-

G re

Figure 6. Sample of german historical document with significant noise and warp from our test-set.

Acknowledgment

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

References

[1] K. Jung, K. I. Kim and A. K. Jain, "Text information extraction in images and video: a survey," Pattern Recogn vol. 37 issue 5, pp. 977-997, 2004.

[2] S. N. Srihari and V. Govindaraju, "Analysis of textual images using the Hough transfor,". Mach Vis A, pp. 141–153, 1989.

[3] W. Postl, "Detection of linear oblique structures and skew scan in digitized documents," Proc. 8th international conference on pattern recognition, pp. 687-689, 1986.

[4] H. S. Baird, "The skew angle of printed documents," Proc. SPSE 40th symposium hybrid imaging systems, Rochester, NY, pp 739–743M, 1987.

[5] G. Ciardiello, G. Scafuro, M. T. Degrandi, M. R. Spada and M. P Roccotelli, "An experimental system for office document handling and text recognition," Proc. 9th international conference on pattern recognition, pp. 739–743, 1988.

[6] Y. Ishitani, "Document skew detection based on local region complexity," Proc. 2nd international conference on document analysis and recognition, Tsukuba, Japan, pp. 49–52, 1993.

[7] D. S. Bloomberg, G. E. Kopec and L. Dasari, "Measuring document image skew and orientation," Document Recognition II SPIE vol. 2422, pp 302–316, 1995.

[8] L. Gorman, "The document spectrum for page layout analysis," IEEE Trans Pattern Anal Mach Intell vol. 15 issue 11, pp. 1162–1173, 1993.

[9] J. Sadri, M. Cheriet, "A new approach for skew correction of documents based on particle swarm optimization," Proc. 10th international conference on document analysis and recognition, ICDAR '09, pp. 1066–1070, 2009.

[10] T. Akiyama and N. Hagita; "Automatic entry system for printed documents," Publisher Elsevier Science, Pattern Recognition vol. 23 , issue 11, pp.1141 – 1154, 1990.

[11] Beker, Henry, Piper and Fred "Cipher Systems: The Protection of Communications," Wiley-Interscience, p. 397, 1982. Fletcher and Pratt "Secret and Urgent: the Story of Codes and Chiphers," Blue Ribbon Books, pp. 254-255, 1939.

[12] CorpusDe Thomas Tempe, 2007.

[13] AL Beutelspacher, "Kryptologie," Wiesbaden: Vieweg, p. 10, 2005.

[14] R. Safabakhsh and Shahram Khadivi, "Document Skew Detection Using Minimum-Area Bounding Rectangle," Proc. The International Conference on Information Technology: Coding and Computing ITCC 00, pp. 253 – 258, 2000.

[15] Carl von Eckartshausen, "Aufschlüsse zur Magie aus geprüften Erfahrungen über verborgene philosophische Wissenschaften und verdeckte Geheimnisse der Natur", Bavarian State Library, 1778.

[16] "Le Dernier fils de France, ou le Duc de Normandie, fils de Louis XVI et de Marie-Antoinette", Bibliothèque nationale de France, 1838.