Image and Vision Computing 20 (2002) 813–824
www.elsevier.com/locate/imavis

Skew angle estimation for printed and handwritten documents using the Wigner–Ville distribution

E. Kavallieratou*, N. Fakotakis, G. Kokkinakis

Wire Communications Laboratory, Dept. of Electrical Computer Engineering, University of Patras, 26500 Patras, Greece

Abstract

A skew estimation algorithm for printed and handwritten documents, based on the document's horizontal projection profile and its Wigner–Ville distribution, is presented. The proposed algorithm is able to correct skew angles that range between −89 and +89° detecting the right oriented position of the page by the alternations of the horizontal projection profile. It is able of processing successfully handwritten documents, even if they consist of non-parallel text lines. It deals with the presence of graphics, while a few text lines suffice for the application of the algorithm. Furthermore, the latter permits the use of only a part of the page for the skew estimation minimizing the computational complexity. The proposed algorithm was evaluated on a wide variety of pages (i.e. printed, handwritten, multi-column, application forms etc.) achieving a success rate of 100% within a confidence range of ±0.3°. © 2002 Published by Elsevier Science B.V.

Keywords: Skew angle correction; Skew angle detection; Wigner–Ville distribution

1. Introduction

Optical character recognition (OCR) systems facilitate the everyday use of computers by transforming large amounts of documents, either printed or handwritten, into electronic form for further processing. However, typical OCR tasks, e.g. thinning, segmentation, feature extraction or classification, are usually sensitive to the existence of skewing in the page. On the other hand, when a document page is photocopied or scanned, a skew angle may be introduced. Thus, a robust OCR system should include a skew angle detection and correction module.

Several methods have been used to solve the problem of skew detection. O'Gorman [1] proposes a categorization of these methods according to their techniques into three categories: projection profile, Hough transform and nearest neighbor clustering. Moreover, hybrid systems using more than one technique have also been proposed.

The *projection profile* technique is applied by Postl [2] in combination with the Fourier transform. He determines the skew angle from the density of the Fourier space. The same technique is employed by Peake [3] as well, but with a higher accuracy when applied to pages with graphics and photos. The horizontal projection profile used by Akiyama [4] is a faster approach in comparison with the vertical projection profile approaches (Pavlidis [5] and Ciardilello [6]). Finally, Baird [7] proposes a combination of projection profile and connected components techniques.

The Hough transform, despite its computational complexity, is a popular technique of skew estimation. Yin [8], first smoothes the black runs and locates the black–white transitions to emphasize the text lines, then the skew angle is determined by an improved Hough Transform. Srihari [9] as well as Le [10] calculate the skew angle by using Hough transform, while Hinds [11] uses it in combination with run-length encoding, and Wang [12] looks in his approach for suitable peaks in the Hough transform space. Yu [13] presents a very accurate and rather fast approach in a combination of the Hough transform and connected components. In order to reduce the computational cost, Jiang [14] pre-estimates the skew angle and then applies the Hough transform.

O'Gorman [1] and Hashizume [15] preferred the *nearest neighbor clustering* technique in their approaches. Finally, several other techniques have been proposed such as those of Pratt [16], Liu [17], Gatos [18] and Okun [19].

In this paper we present a generic approach to skew angle estimation, based on the Wigner–Ville distribution. The structure of this paper is as follows: in Section 2 through the presentation of common problems in skew angle detection, we try to explain the basic reasons that led us to propose a new technique for skew angle detection. The Wigner–Ville distribution is presented in Section 3 while the proposed algorithm is described in Section 4. Finally, in Sections 5

* Corresponding author. Tel.: +30-61-991-722; fax: +30-61-991-855.
E-mail address: ergina@wcl.ee.upatras.gr (E. Kavallieratou).

0262-8856/02/$ - see front matter
PII: S0262-8856(02)00091-4

and 6, experimental results and conclusions are given, respectively.

2. Common problems in skew angle estimation

Okun [19] refers three types of skew: a global skew, when there is a common orientation for all the page blocks; a multiple skew, when several orientations are present on the page (e.g. handwritten pages); and a non-uniform text line skew, when several orientations are present on the same text line (e.g. Hill and Dale writing). The last case was presented in Ref. [20]. Here we will focus on the first two cases.

Although many of the above-mentioned approaches are satisfactorily accurate [3,13,18] and fast [4,13], either they fail to face some common problems or they are applicable to specific applications. Some common problems in skew angle detection are:

(a) Restriction of the detectable angle range [4,7,15,18].
(b) Restrictions on type or size of fonts [2,19].
(c) Dependence on page layout. Usually the presence of graphics, borders, or tables constitutes a problem [2,14, 21].
(d) A specific document resolution is required [3,14,19].
(e) High computational cost [8,10,11].
(f) Limitation to specific applications [23–25].
(g) Large text areas are required [18,21,22].
(h) Moreover, most of the proposed algorithms are appropriate for machine-printed pages and fail when they deal with handwritten documents. The proposed methods for handwritten or mixed (i.e. both handwritten and printed) documents concern specific applications [23]. Chin [26] illustrates this problem by applying known algorithms to handwritten pages and underlines that the methods which handle printed pages successfully, either do not manage to deal with handwritten pages with comparable accuracy or require a considerably increased computational cost. Furthermore, the hitherto proposed algorithms can estimate the dominant skew angle and cannot deal with the cases of handwritten pages where the text lines may not be parallel to each other (Fig. 1).

Although many of the techniques mentioned in Section 1 can deal with several of the above-described problems, there is not yet a technique able to face all of them at the same time. Thus, the majority of methods usually fail in the case of application forms (Fig. 2) where the fonts, character size, and density of page vary. The proposed method is language, fonts, and size of fonts independent, while it successfully deals with borders, graphics and pages with any resolution. It covers angles that range between +89 and −89°. Moreover, since only a few text lines are sufficient for the application of the algorithm, the computational cost can be minimized, by using a portion of the page instead of the whole document and a dynamic algorithm.

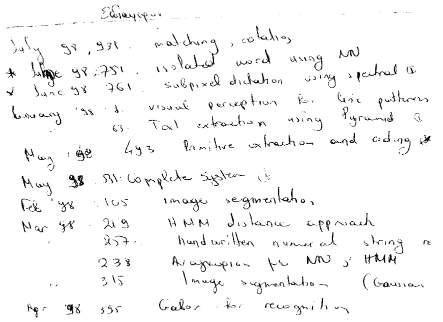


Fig. 1. A handwritten page with non-parallel text lines.

The proposed method is able to handle handwritten documents, even multiple-skewed (Fig. 1), since the algorithm recognizes areas with different skew angles.

3. Wigner–Ville distribution

A very important chapter of signal processing is the representation of *time-varying* or *non-stationary* signals, that is signals whose characteristics vary with time. For such

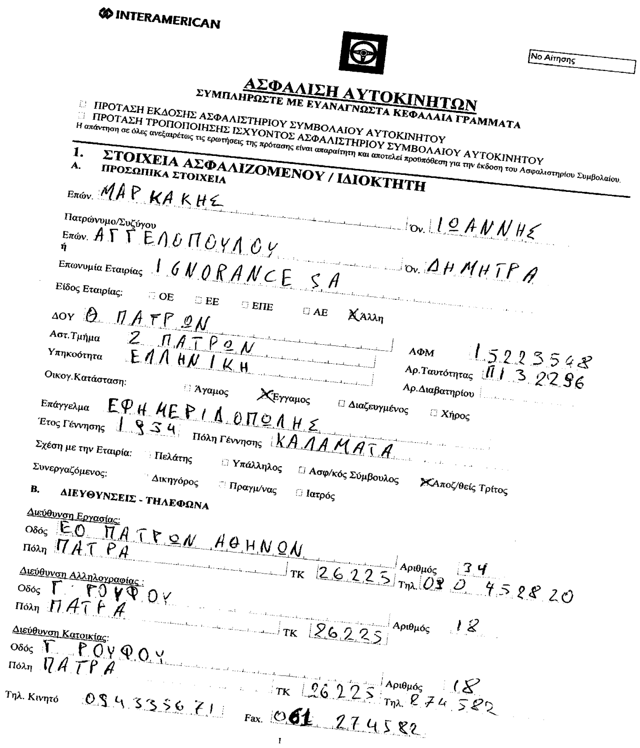
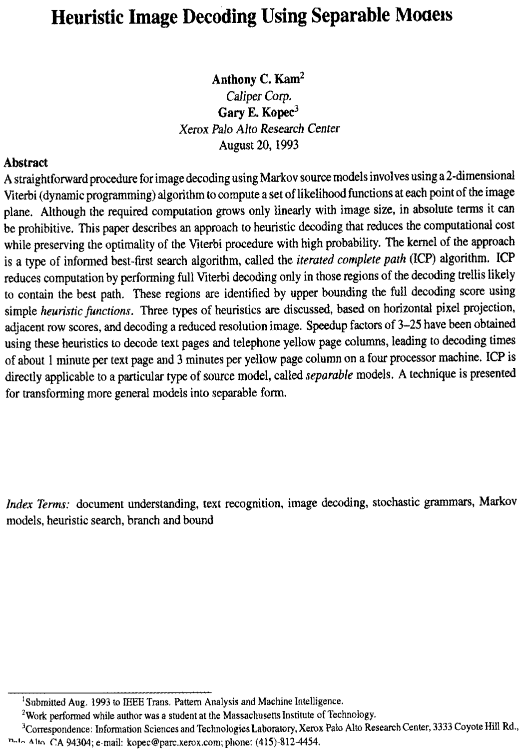


Fig. 2. A skewed application form. Such pages with various fonts, both printed and handwritten, and borders usually complicate the skew angle estimation.



Heuristic Image Decoding Using Separable Models

Anthony C. Kam[2]
Caliper Corp.
Gary E. Kopec[3]
Xerox Palo Alto Research Center
August 20, 1993

Abstract
A straightforward procedure for image decoding using Markov source models involves using a 2-dimensional Viterbi (dynamic programming) algorithm to compute a set of likelihood functions at each point of the image plane. Although the required computation grows only linearly with image size, in absolute terms it can be prohibitive. This paper describes an approach to heuristic decoding that reduces the computational cost while preserving the optimality of the Viterbi procedure with high probability. The kernel of the approach is a type of informed best-first search algorithm, called the *iterated complete path* (ICP) algorithm. ICP reduces computation by performing full Viterbi decoding only in those regions of the decoding trellis likely to contain the best path. These regions are identified by upper bounding the full decoding score using simple *heuristic functions*. Three types of heuristics are discussed, based on horizontal pixel projection, adjacent row scores, and decoding a reduced resolution image. Speedup factors of 3–25 have been obtained using these heuristics to decode text pages and telephone yellow page columns, leading to decoding times of about 1 minute per text page and 3 minutes per yellow page column on a four processor machine. ICP is directly applicable to a particular type of source model, called *separable* models. A technique is presented for transforming more general models into separable form.

Index Terms: document understanding, text recognition, image decoding, stochastic grammars, Markov models, heuristic search, branch and bound

[1]Submitted Aug. 1993 to IEEE Trans. Pattern Analysis and Machine Intelligence.
[2]Work performed while author was a student at the Massachusetts Institute of Technology.
[3]Correspondence: Information Sciences and Technologies Laboratory, Xerox Palo Alto Research Center, 3333 Coyote Hill Rd., Palo Alto, CA 94304; e-mail: kopec@parc.xerox.com; phone: (415)-812-4454.

Fig. 3. A right-oriented document image.

signals the concept of time–frequency distributions has been introduced.

A first class of time–frequency representations is the *atomic decomposition*, or *linear time–frequency representations*. These distributions decompose the signal on the basis of elementary signals (i.e. the atoms), which have to be well localized in both time and frequency. However, there is a trade-off between time and frequency resolutions, as the decomposition is followed by windowing the signal. A good time resolution requires a short window; on the other hand, a good frequency resolution requires a long window. This is a consequence of the time–frequency resolution relation via Heisenberg–Gabor inequality [27].

In contrast, the *energy distributions*, another class of time–frequency representations, distribute the energy of the signal over two description variables: time and frequency.

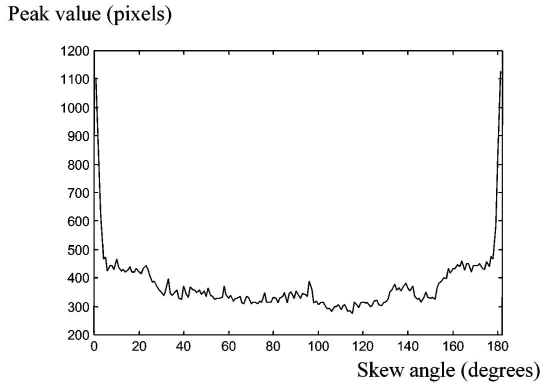


Fig. 5. Maximum peaks of the histograms referred to the document page of Fig. 3, with respect to the skew angle.

In order to devise a joint function of time and frequency, a lot of distributions have been proposed since 1932 [28–32] but only in 1966 Cohen [33] proved that an infinite number of distributions can be generated by the unified formulation:

$$\rho_z(t,f) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \times \mathrm{e}^{\mathrm{j}2\pi v(u-t)} g(v,\tau) z\left(u+\frac{\tau}{2}\right) z\left(u-\frac{\tau}{2}\right) \mathrm{e}^{-\mathrm{j}2\pi f\tau} \mathrm{d}v\,\mathrm{d}u\,\mathrm{d}\tau \tag{1}$$

where $g(v,\tau)$ is an arbitrary function called the *Kernel* function. The Kernel function characterizes the observation mode chosen by the analyst. It determines how the signal energy is distributed in the time and the frequency domain and it corresponds to the windows that are used in the atomic decomposition.

The simplest member of Cohen's class is the Wigner–Ville distribution (WVD):

$$W(t,f) = \int_{-\infty}^{+\infty} z(t+\tau/2) z(t-\tau/2) \mathrm{e}^{-2\pi f\tau} \mathrm{d}\tau, \tag{2}$$

where $z(t)$ represents the analytical signal associated with the signal $s(t)$. The WVD can also be expressed as a

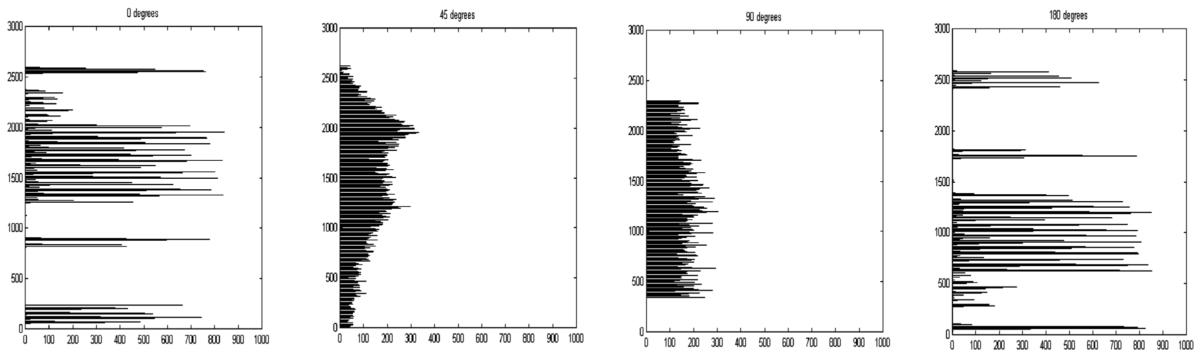


Fig. 4. Horizontal histograms of the document page of Fig. 3 for various skew angles.

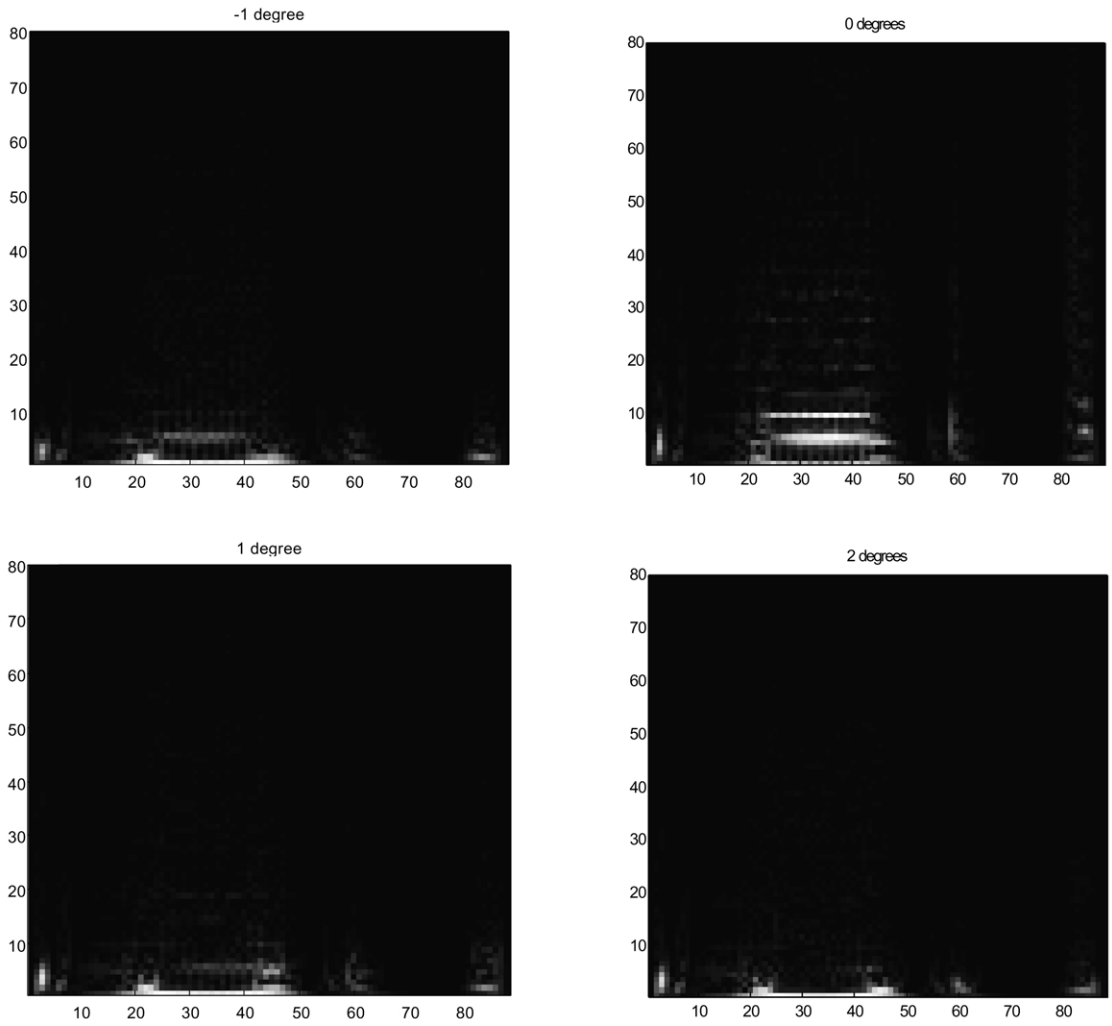


Fig. 6. The WVDs for several histograms of the document page of Fig. 3.

function of the spectrum of the signal $Z(f)$ under analysis as follows:

$$W(t,f) = \int_{-\infty}^{+\infty} Z(f + u/2)Z(f - u/2)e^{j2\pi ut}du. \quad (3)$$

The WVD function is a particularly popular distribution due to the large number of desirable mathematical properties it satisfies. Mecklenbrauker [27] proves the uniqueness of the WVD "in the sense that it is that single energy distribution that possesses all the stated desirable properties". This justifies the numerous applications of WVD in pattern recognition [34,35], signal synthesis [36], seismic signal [37], optics [38], etc.

In the case of the present application, the superiority of WVD over the other Cohen's class distributions has been demonstrated in Ref. [39]. In more detail, the tests with several distributions, members of the Cohen's class, proved WVD to be the best compromise between computational cost and accuracy.

4. Skew angle estimation

Our approach employs the projection profile technique and the WVD. Specifically, the maximum intensity of the WVD of the horizontal histogram of a document image is used as the criterion for its skew angle estimation.

The horizontal histogram of a properly oriented document page presents the maximum peaks and the most intent alternations between peaks and dips than any other histogram-by-angle of the same page [3,40]. In Fig. 4, the horizontal histograms of the document image of Fig. 3 for various skew angles are shown, while in Fig. 5 the highest peak values of the histograms of the same page are illustrated with respect to the corresponding skew angle. As expected, the peaks become maximum at 0 and 180°, that is, when the page is right or reverse oriented.

The WVD of the histograms represents their time–frequency distribution, where in this case the time increases according to the height of the page. Consequently, the WVD presents maximum intensity for the histograms of 0 and 180°, which indicate the most intent alternations between

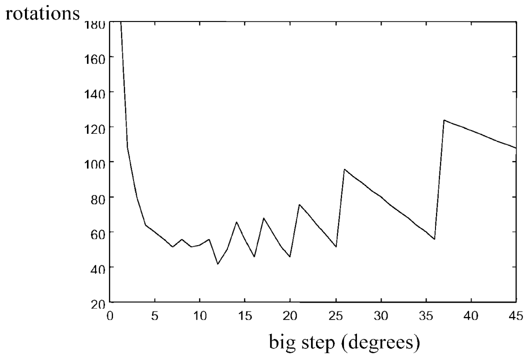


Fig. 7. The relation between big step and rotations. The required rotations are minimized for a big step equal to 12°.

peaks and dips i.e. for the histogram corresponding to the right or reverse oriented page. The closer the skew angle to 0 and 180°, the higher the value of the maximum intensity. The latter guarantees the success of our algorithm for skew angle ranging between −89 and +89° with respect to the right page orientation. Otherwise the page may end up oriented at reverse position. In Fig. 6 the WVDs of histograms for several skew angles are shown.

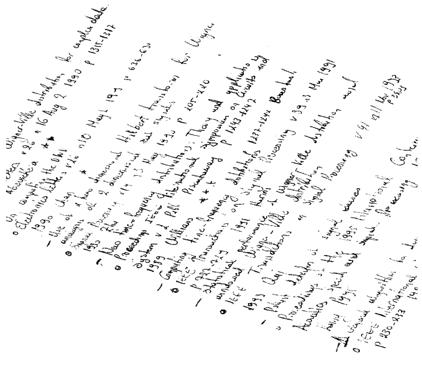
However, the calculation of the histograms for every skew angle and the application of the Wigner–Ville function would imply unnecessary computational complexity and cost. Thus, at the beginning, a big step is used for a rough estimation, and then a more accurate estimation follows within a smaller area.

Consequently, the proposed algorithm for skew angle estimation and correction for a given page, described in Fig. 8, consists of six steps:

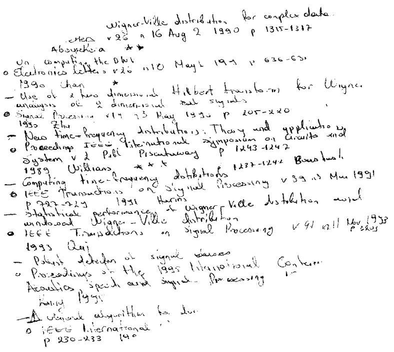
Step 1
The document is rotated both to the right and left by an angle (*big step* Section 4.1) within the area ±89°. For each angle the horizontal projection profile is extracted.
Step 2
For each projection, the WVD as well as the maximum intensity of the distribution are calculated.
Step 3
The angle whose histogram presents the maximum intensity, i.e. it is closer to the right page position, is selected (Section 4.2) and the document is rotated by this angle (*angle1*).
Step 4
The procedure is repeated from Step 1 to Step 3 while the document is rotated within a smaller area (around the angle1) by one degree at the time and a more exact angle is calculated (*angle2*).
Step 5
The procedure is repeated from Step 1 to Step 3 while the document is rotated within an even smaller area (around the angle1 + angle2) by 0.1° at the time and a more exact angle is calculated (*angle3*).
Step 6
The estimated skew angle is angle1 + angle2 + angle3.

```
Angle=-84
while Angle<=84{
       rotate_page[Angle]
       calculate horizontal histogram[Angle]
       calculate WVD[Angle]
       extract maximum intensity curve[Angle]
       Angle=Angle+12
}
select the maximum intensity curve[angle1] with the highest peak
rotate_page[angle1*12]

Angle=-6
while Angle<=6{
       rotate_page[Angle]
       calculate horizontal histogram[Angle]
       calculate WVD[Angle]
       extract maximum intensity curve[Angle]
       Angle=Angle+1
}
select the maximum intensity curve[angle2] with the highest peak
rotate_page[angle2*1]

Angle=-0.5
while Angle<=0.5{
       rotate_page[Angle]
       calculate horizontal histogram[Angle]
       calculate WVD[Angle]
       extract maximum intensity curve[Angle]
       Angle=Angle+0.1
}
select the maximum intensity curve[angle3] with the highest peak
rotate_page[angle3*0.1]
```

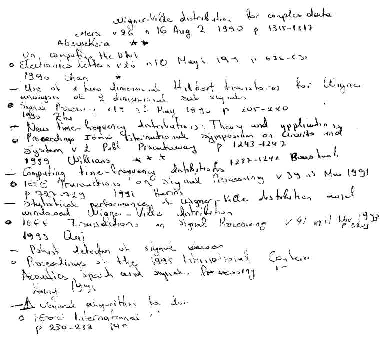
Fig. 8. The proposed algorithm.



(a)



(b)



(c)

Fig. 9. (a) A skewed handwritten document, (b) after the first estimation, rotated by 60° and (c) after the final estimation rotated by 62.0°.

4.1. Step selection

As already mentioned, the calculation of the histogram and the WVD for every angle is unnecessary. Moreover, since the rotation part is the most computationally expensive part of our algorithm, it is desirable to minimize the number of required rotations. We calculated the overall necessary rotations for various steps. The results are shown in Fig. 7. We note that for

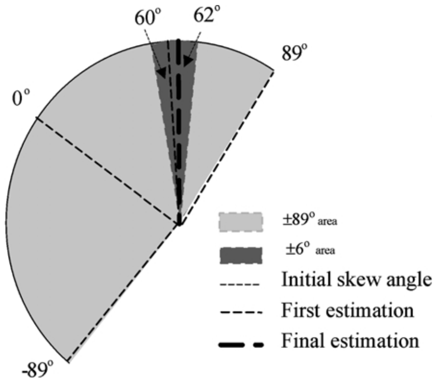


Fig. 10. The procedure for the handwritten document of Fig. 9(a).

a big step of 12° the number of rotations is minimized. Consequently, 12° is the indicated step.

However, the whole area between $+89$ and $-89°$ and only that should be covered. In the opposite case, if the covered area is greater than that of $\pm 89°$, it is possible to end up with the page reversed. On the other hand, if the covered area is smaller than that and the skew angle is greater than the covered area, the system will fail to correct the page.

Thus, the document is rotated from -84 to $+84°$ by a step of 12° and after a first skew estimation and correction of the document, the same procedure is followed, this time within the area -6 and $+6°$, for a step of 1°. Finally, we repeat the procedure for a step of 0.1° in the area of $\pm 0.5°$ around the pre-estimated angle.

In Fig. 10 the whole procedure is shown for the handwritten document of Fig. 9. Initially the area of $\pm 89°$ is covered and a skew angle of 60° is first estimated (Fig. 9(b)). The procedure is repeated within a smaller area of $\pm 6°$ with respect to the first estimation and then within the area of $\pm 1°$. The final estimation is 62.0°. The successfully corrected page is shown in Fig. 9(c).

4.2. The maximum intensity as criterion

As already mentioned the maximum intensity of the WVD is used as the criterion for the skew angle estimation. The WVD is calculated and the corresponding curve of maximum intensity is extracted for each histogram. Thus, for every angle a curve of maximum intensity is available.

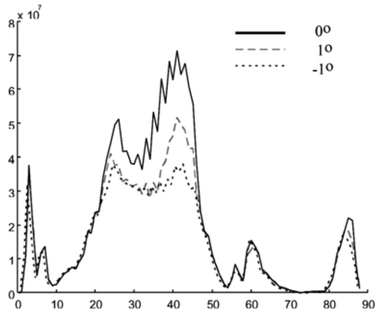


Fig. 11. Curves of maximum intensity extracted from Fig. 3.

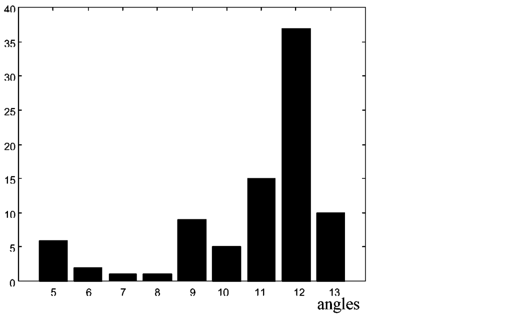


Fig. 12. Number of times that several curves of maximum intensity for different skew angles present maximum.

In Fig. 11 several curves of maximum intensity referred to the page of Fig. 3 are shown.

In order to define the curve that corresponds to the right-oriented page we select the curve that presents the most maxima in the time domain. Specifically, in order to focus on the peaks of the curves, only the curves with values above a threshold (a tenth of the maximum peak proved a good threshold in our experiments) are examined. In Fig. 12, the number of the maximum values, measured on several curves of the page shown in Fig. 2 is presented, with respect to the corresponding skew angle of this page. Obviously, the estimated skew angle of the page is $+12°$.

4.3. Reduction of the computational cost

It is worth mentioning that a few text lines are sufficient for skew angle detection. Thus, the use of the whole page is not necessary. On the other hand, this method is imposed in case of multi-column documents, where the use of a part of the page is safer, and at the same time it reduces significantly the computational time and complexity.

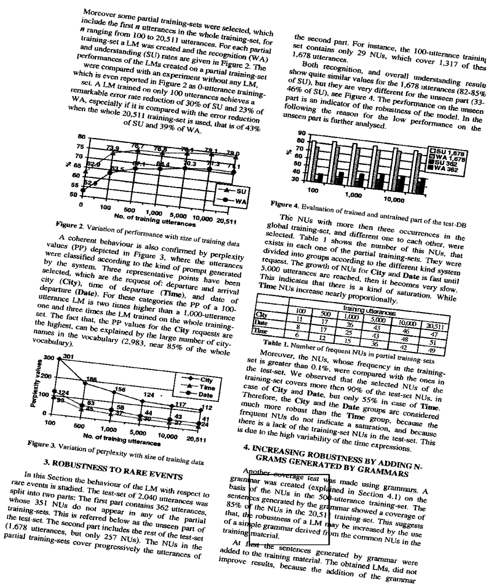
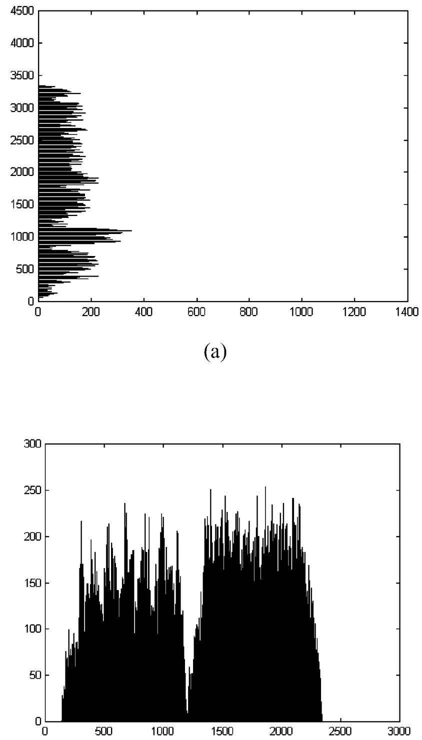


Fig. 13. A two-column skewed document.



(a)

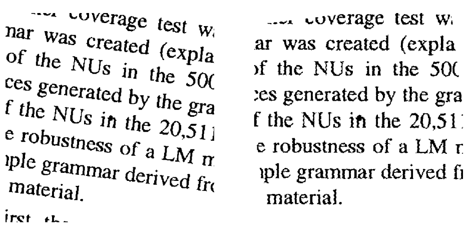
(b)

Fig. 14. The horizontal (a) and vertical (b) histogram of the page of Fig. 13.

However, the selection of the appropriate part to be used may constitute a problem. For selecting such a part we extract a segment of the page centered on the point whose co-ordinates are the peaks of the horizontal and vertical projection profiles. In this way, the part of the page with the maximum density in black pixels is located.

This technique guarantees the selection of a segment with text in the most cases of document images, application forms and pages with graphics, when the text parts of the page present a higher density in black pixels than graphics, tables, etc. However, this assumption does not hold in all pages that include pictures and the proposed technique may fail to select a part with text. In such cases the use of the whole page is rather recommended.

In Fig. 13 a two-column document with tables and graphics is shown. The selected part shown in a rectangle in Fig. 13, with dimensions 400×400 pixels, is centered on the pixels (1276,1564), the maxima of the projections of its horizontal and vertical histograms (Fig. 14(a) and (b),



(a) (b)

Fig. 15. The selected part (a) before and (b) after the correction.

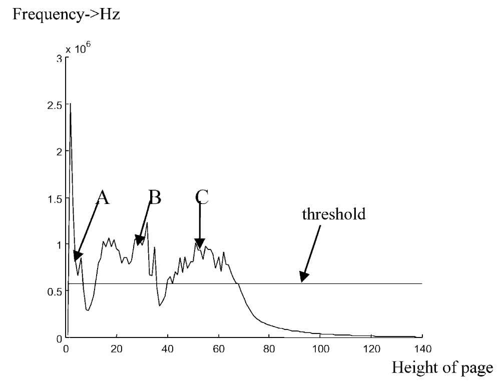


Fig. 16. The average maximum intensity curve, the selected threshold and the extracted areas (A, B, and C) of different skew angle for the page of Fig. 1.

respectively). In Fig. 15 the selected part is shown before and after the application of our method.

4.4. Application of the method to handwritten documents

In handwritten documents, as the one in Fig. 1, more than one skew angles are very often met in the same page. Most of the proposed approaches handle this case by estimating and correcting the skewing according to the dominant angle.

Our approach deals with this problem processing the page by areas. The segmentation into areas is based on the fact that the variation of the angle of successive non-parallel text lines causes wider valleys in the horizontal projection profile of the page, in comparison to parallel lines. These valleys are, then, interpreted into deeper dips at the corresponding curve of maximum intensity of the WVD.

In order to localize these dips and use them as boundaries of areas of different skew angles, we follow the procedure below:

- The average value of the curves of maximum intensity is calculated.
- A threshold depending on the average value is established (the 3/4 of the average value of the average curve is a good threshold in our experience) (Fig. 16).

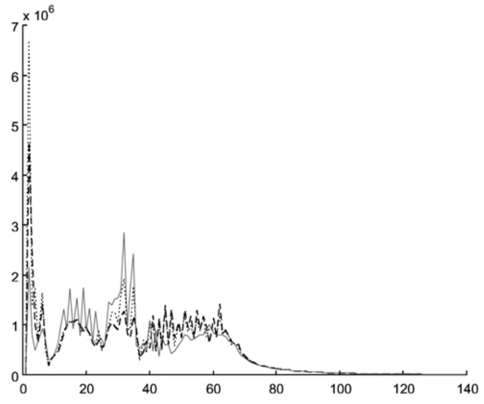


Fig. 17. Several curves of maximum intensity for the page of Fig. 1.

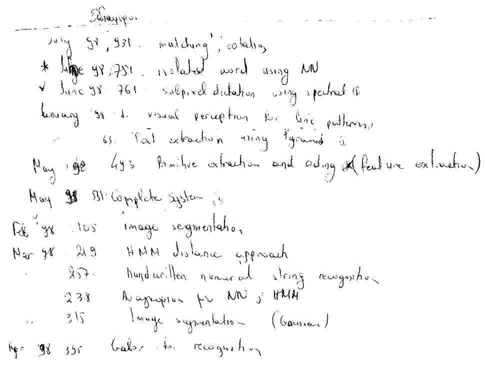


Fig. 18. The corrected page of Fig. 1.

- The dips below this threshold define the boundaries of the areas.
- The dominant curve of maximum intensity in each area provides the corresponding skew angle.

Fig. 17 presents the dominant curves of maximum intensity for the corresponding areas of Fig. 16. In Fig. 18 the corrected page of Fig. 1 is shown. However, in case the text features multiple changes in the line width, the use of the above threshold may segment the page into many areas. Nevertheless, this fact does not affect the accuracy since each area will be corrected separately.

5. Experimental results

The reliable and accurate evaluation of a skew angle estimation system is a very hard task since in many cases it is inevitably based on subjective criteria. Specifically, the perfect orientation of a handwritten document may not be unique and an estimation of skew angle with a confidence range of $\pm 0.5°$ is generally acceptable. Moreover, even the most carefully scanned printed documents may be slightly skewed by 0.1°. Thus, when dealing with cases where the skew angle has to be estimated with a confidence range of $\pm 0.1°$ the employment of scanned documents is not sufficient.

Documents in postscript form are a very good source of unambiguously perfectly oriented printed document images. We took advantage of this fact by extracting 30 printed document images from postscript documents and artificially

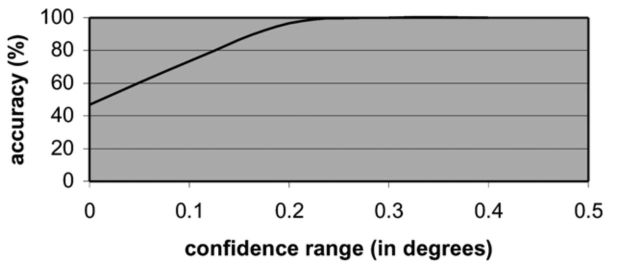


Fig. 19. The confidence range of the estimated skew angle vs. the achieved accuracy.

Table 1
Experimental data: document images extracted from PS form

Image no	Resolution in dpi	Rate of text in the page (%)	Border	Columns	Graphics	Different type of fonts	Skew angle in degrees	Estimated skew angle
1	80	100	No	1	No	3	85	85
2	280	20	No	1	No	1	78.8	78.5
3	300	70	Yes	1	No	3	72.3	72.0
4	240	80	No	1	Yes	1	66	66
5	280	80	No	1	Yes	3	59.7	59.7
6	260	100	No	2	No	3	53.4	53.5
7	220	90	Yes	1	No	2	47.1	47.0
8	200	30	Yes	1	No	2	40.7	40.5
9	100	50	Yes	1	Yes	4	34.3	34.3
10	100	60	No	1	Yes	2	27.9	28
11	80	80	No	2	Yes	2	21.6	21.5
12	160	100	Yes	1	No	3	15.2	15.0
13	180	100	Yes	1	No	4	8.9	8.9
14	60	60	No	1	Yes	2	2.5	2.5
15	50	100	No	1	No	2	−3.9	−3.9
16	200	30	No	1	No	2	−10.3	−10.3
17	280	30	Yes	1	No	5	−16.5	−16.5
18	110	90	No	2	Yes	2	−23.1	−23
19	60	30	No	1	No	2	−29.4	−29.5
20	80	50	No	1	Yes	1	−35.8	−36
21	230	40	No	1	No	3	−42.2	−42
22	190	20	No	1	No	2	−48.6	−48.5
23	130	70	Yes	2	No	3	−55	−55
24	90	50	Yes	1	Yes	1	−61.4	−61.5
25	110	100	No	1	No	2	−67.8	−68
26	130	50	Yes	1	Yes	3	−74.2	−74
27	160	70	Yes	2	No	3	−76.5	−76.5
28	280	100	Yes	1	No	4	−80.5	−80.5
29	100	50	No	1	Yes	1	−82.5	−82.5
30	160	100	No	1	No	2	−85	−85

rotated them resulting in skewed documents. The proposed system was, then, applied to the skewed document images for estimating the skew angle. The artificial and the estimated skew angles are presented in Table 1. As can been seen the accuracy of the estimation procedure does not depend on the resolution level, the presence of borders, the number of columns, the presence of graphics nor the number of different fonts included in the page.

The third column of Table 1 is a rough indication of the presence of text in the page. The performance of the proposed method is satisfactory even in cases where only a limited percentage of the image (<40%) is covered by text. In other words, a few text lines are generally adequate for achieving a good estimation of the skew angle.

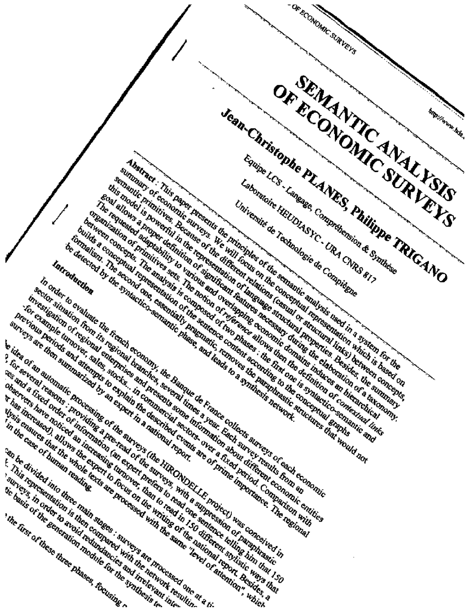
The relation of the confidence range of the estimated skew angle with the achieved accuracy is shown in Fig. 19. The performance of our system with a confidence range greater than 0.3° is perfect. Note that in most cases a deviation of 0.3° or less is not likely to be detected with the naked eye.

In order to test the performance of the proposed system in handwritten documents we applied the skew estimation procedure to documents taken from the following sources:

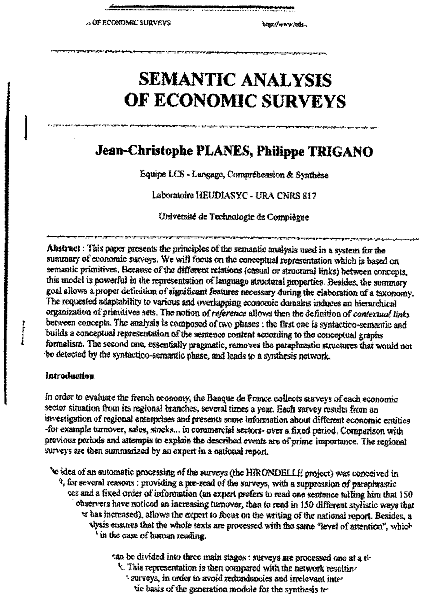
- A collection of over 250 application forms containing both printed and handwritten text. These forms were produced in the framework of the European project ACCeSS (LE-1 1802). An example is shown in Fig. 2.
- 300 documents of the AIM-DB [41]. A typical document of this database consists of a printed and a handwritten part. Moreover, it has been paid special attention at keeping the handwritten text lines in parallel.
- A collection of scanned handwritten student notes of any level of complexity. An example is shown in Fig. 1.

As mentioned above, the definition of the perfectly oriented document in such cases is an ambiguous task. Thus, the performance of our system is not easy to be measured with a confidence range lower than 0.5°. However, our system succeeded to correct all the mentioned document images within this confidence range.

In case of student notes, where non-parallels text lines were often met, our system first corrects the page by the dominant skewed angle and then finds the areas with different skew angles and corrects the local skew angle by



(a)



.. OF ECONOMIC SURVEYS http://www.hds..

SEMANTIC ANALYSIS
OF ECONOMIC SURVEYS

Jean-Christophe PLANES, Philippe TRIGANO

Equipe LCS - Langage, Compréhension & Synthèse

Laboratoire HEUDIASYC - URA CNRS 817

Université de Technologie de Compiègne

Abstract : This paper presents the principles of the semantic analysis used in a system for the summary of economic surveys. We will focus on the conceptual representation which is based on semantic primitives. Because of the different relations (casual or structural links) between concepts, this model is powerful in the representation of language structural properties. Besides, the summary goal allows a proper definition of significant features necessary during the elaboration of a taxonomy. The requested adaptability to various and overlapping economic domains induces an hierarchical organization of primitives sets. The notion of *reference* allows then the definition of *contextual links* between concepts. The analysis is composed of two phases : the first one is syntactico-semantic and builds a conceptual representation of the sentence content according to the conceptual graphs formalism. The second one, essentially pragmatic, removes the paraphrastic structures that would not be detected by the syntactico-semantic phase, and leads to a synthesis network.

Introduction

In order to evaluate the french economy, the Banque de France collects surveys of each economic sector situation from its regional branches, several times a year. Each survey results from an investigation of regional enterprises and presents some information about different economic entities -for example turnover, sales, stocks... in commercial sectors- over a fixed period. Comparison with previous periods and attempts to explain the described events are of prime importance. The regional surveys are then summarized by an expert in a national report.

..e idea of an automatic processing of the surveys (the HIRONDELLE project) was conceived in ..9, for several reasons : providing a pre-read of the surveys, with a suppression of paraphrastic ..ces and a fixed order of information (an expert prefers to read one sentence telling him that 150 observers have noticed an increasing turnover, than to read in 150 different stylistic ways that ..r has increased), allows the expert to focus on the writing of the national report. Besides, a ..alysis ensures that the whole texts are processed with the same "level of attention", which ..t in the case of human reading.

..can be divided into three main stages : surveys are processed one at a ti.. ..k. This representation is then compared with the network resultin.. ..surveys, in order to avoid redundancies and irrelevant inte.. ..tic basis of the generation module for the synthesis te..

(b)

Fig. 20. A skewed document page with a resolution of 100 dpi (a) before and (b) after the correction.

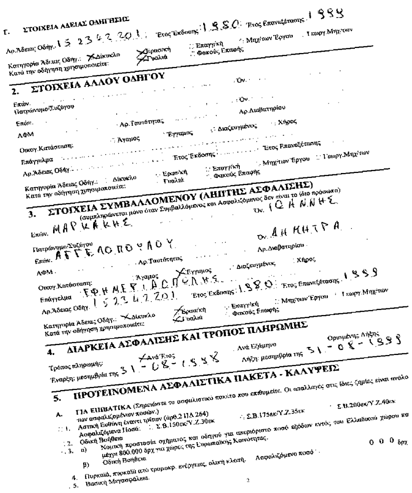
repeating the procedure within the range of $\pm 10°$. The technique described in Section 4.4 detects the areas of different orientation in the same page which reaches a success rate of 93%. When it fails to detect an area of different skew angle, the dominant skew angle of the whole page is used.

It is worth mentioning that the required CPU time depends on the size and the resolution of the page. Some characteristic examples of the application of the presented approach are shown in Figs. 20–23, while in Table 2 performance details concerning some example pages shown in this paper are given. The CPU time refers to a PC Pentium III at 750 MHz.

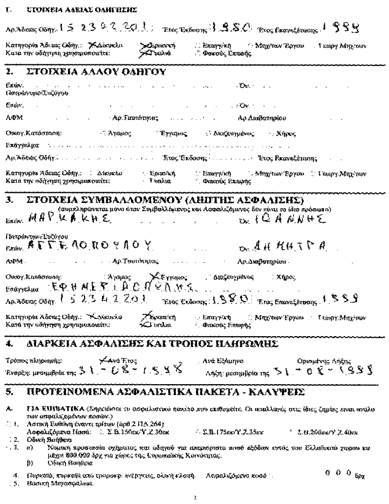
6. Conclusions

We presented a fast and robust algorithm for the skew angle estimation in both printed and handwritten documents based on the horizontal histogram of the page and its WVD.

Our approach handles successfully pages with both printed and handwritten text. It is independent of resolution and type and size of fonts. As far as pages with graphics, tables, photos and borders are concerned, it is quite robust, since a few text lines are sufficient for its application. The latter permits the use of only a small part of the page for the skew estimation, which is necessary for the successful handling of multi-column documents, and at the same time

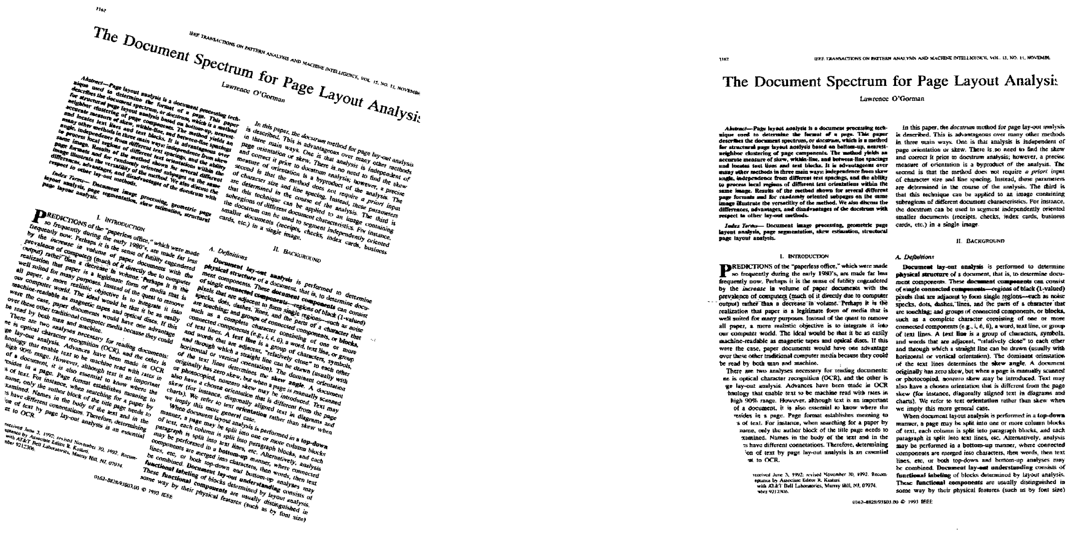


(a)



(b)

Fig. 21. Skewed page of application form with a resolution of 100 dpi (a) before and (b) after the correction.

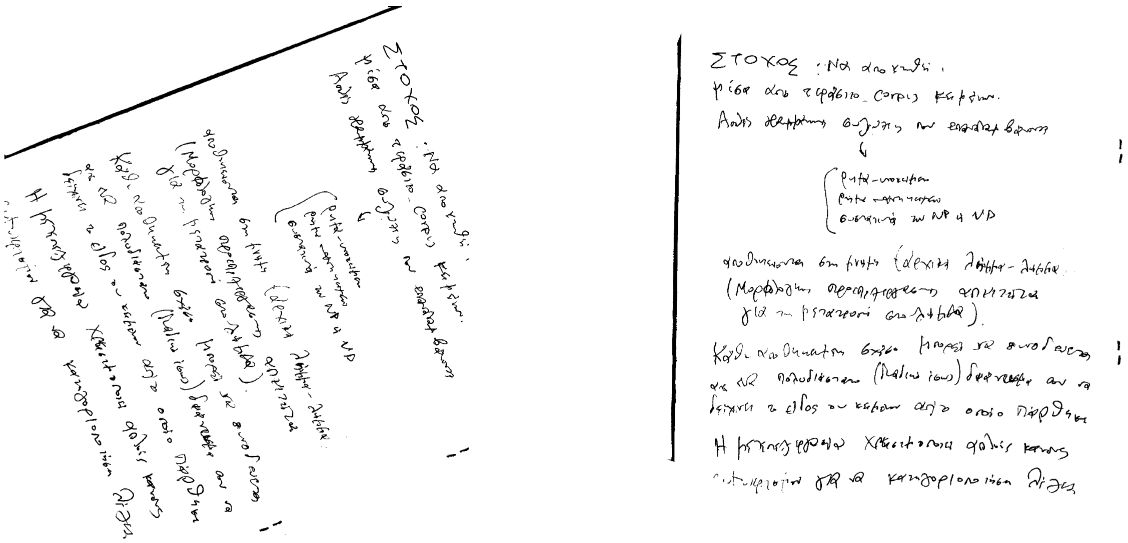


The Document Spectrum for Page Layout Analysis

Lawrence O'Gorman

(a) (b)

Fig. 22. A skewed page of two-column document with a resolution of 100 dpi (a) before and (b) after the correction.



(a) (b)

Fig. 23. A skewed page of handwritten document (a) before and (b) after the correction.

Table 2
Performance example for some of the pages shown in this paper

Figure	Resolution (dpi)	Size (pixels)	CPU time (whole page used) (s)	CPU time (window 400 × 400p) (s)	Reduction (%)	Estimated angle (deg)
2	300	2550 × 3700	6.11	0.82	86.58	− 6.2
8	300	2550 × 4200	7.23	0.83	88.52	62.0
13	300	2550 × 3500	–	0.82	–	− 6.9
20	100	850 × 1083	2.25	0.79	64.89	− 36.3
21	100	850 × 1399	2.90	0.82	71.72	7.1
22	100	850 × 1399	–	0.80	–	− 15.8
23	300	2550 × 3250	5.72	0.79	86.19	− 68.5

it reduces the computational time-cost. However, the localization of the appropriate portion in pages with pictures remains an open problem for us.

We believe that the most important novelty of our method lies in the ability to handle document pages by areas and deal with pages with multiple skewed text lines, a very common characteristic of handwritten pages.

Our short-term research focuses on the improvement of the technique of segmenting the page in areas with different orientations, since the proposed method may unsuccessfully segment areas of the same orientation. Although this is not a problem as far the resulting accuracy is concerned, it increases the computational cost since the skew angle correction algorithm is called unnecessarily.

References

[1] L. O'Gorman, The document spectrum for page layout analysis, IEEE Trans. Pattern Anal. Mach. Intelligence 15 (11) (1993) 1162–1173.

[2] W. Postl, Detection of linear oblique structures and skew scan in digitized documents, Proceedings of the Eighth International Conference on Pattern Recognition, IEEE CS Press, Los Alamitos, CA, 1986, pp. 687–689.

[3] G.S. Peake, T.N. Tan, A general algorithm for document skew angle estimation, IEEE Int. Conf. Image Process. 2 (1997) 230–233.

[4] T. Akiyama, N. Hagita, Automated entry system for printed documents, Pattern Recogn. 23 (11) (1990) 1141–1154.

[5] T. Pavlidis, J. Zhou, Page segmentation by white streams, Proc. First Int. Conf. Doc. Anal. Recogn. (ICDAR), Int. Assoc. Pattern Recogn. (1991) 945–953.

[6] G. Ciardiello, G. Scafuro, M.T. Degrandi, M.R. Spada, M.P. Roccotelli, An experimental system for office document handling and text recognition, Proc. Ninth Int. Conf. Pattern Recogn. (1988) 739–743.

[7] H.S. Baird, The skew angle of printed documents, Proc. SPSE 40th Conf. Symp. Hybrid Imaging Syst., Rochester, NY (1987) 21–24.

[8] P.-Y. Yin, Skew detection and block classification of printed documents, Image Vis. Comput. 19 (8) (2001) 567–579.

[9] S.N. Shihari, V. Govindaraju, Analysis of textual images using the Hough transform, Mach. Vis. Appl. 2 (1989) 141–153.

[10] D.S. Le, G.R. Thoma, H. Wechsler, Automated page orientation and skew angle detection for binary document image, Pattern Recogn. 27 (10) (1994) 1325–1344.

[11] J. Hinds, L. Fisher, D.P. D'Amato, A document skew detection method using run-length encoding and the Hough transform, Proceedings of the 10th International Conference Pattern Recognition, IEEE CS Press, Los Alamitos, CA, 1990, pp. 464–468.

[12] J. Wang, M.K.H. Leung, S.C. Hui, Cursive word reference line detection, Pattern Recogn. 30 (3) (1997) 503–511.

[13] B. Yu, A.K. Jain, A robust and fast skew detection algorithm for generic documents, Pattern Recogn. 29 (10) (1996) 1599–1629.

[14] H. Jiang, C. Han, K. Fan, A fast approach to the detection and correction of skew documents, Pattern Recogn. Lett. 18 (1997) 675–686.

[15] A. Hashizume, P.S. Yeh, A. Rosenfeld, Method of detecting the orientation of aligned components, Pattern Recogn. 4 (3) (1986) 125–132.

[16] W.K. Pratt, P.J. Capitant, W. Chen, E.R. Hamilton, R.H. Wallis, Combined symbol matching facsimile data compression system, Proc. IEEE 68 (7) (1980) 786–796.

[17] J. Liu, C. Lee, R. Shu, An efficient method for the skew normalization of a document image, Proc. 12th Int. Conf. Pattern Recogn. (1992) 122–125.

[18] B. Gatos, N. Papamarkos, C. Chamzas, Skew detection and text line position determination in digitized documents, Pattern Recogn. 30 (9) (1997) 1505–1519.

[19] O. Okun, M. Pietikainen, J. Sauvola, Document skew estimation without angle range restriction, IJDAR 2 (1999) 132–144.

[20] E. Kavallieratou, N. Fakotakis, G. Kokkinakis, New algorithms for skewing correction and slant removal on word-level, ICECS'99 (1999) 1159–1162.

[21] B.B. Chaudhuri, U. Pal, Skew angle detection of digitized Indian script documents, IEEE Trans. Pattern Anal. Mach. Intelligence 19 (2) (1997) 182–186.

[22] U. Pal, B.B. Chaudhuri, Automatic separation of words in multi-lingual multi-script Indian documents, Proc. Fourth Int. Conf. Doc. Anal. Recogn., Ulm, Germany (1997) 576–579.

[23] B. Yu, A.K. Jain, A generic system for form dropout, IEEE Trans. Pattern Anal. Mach. Intelligence 18 (11) (1996) 1127–1134.

[24] J. Kanai, A.D. Bagdanov, Projection profile based skew estimation algorithm for JBIG compressed images, Int. J. Doc. Anal. Recogn. 1 (1998) 43–51.

[25] A.L. Spitz, Analysis of compressed document images for dominant skew, multiple skew, and logotype detection, Comput. Vis. Image Understanding 70 (3) (1998) 321–334.

[26] W. Chin, A. Harvey, A. Jennings, Skew detection in handwritten scripts, Proc. IEEE Speech Image Technol. Comput. Telecommun. (1997) 319–322.

[27] W. Mecklenbrauker, A tutorial on non-parametric bilinear time–frequency signal representations, Time and Frequency Representations of Signals and Systems, Springer, Wien, pp. 12–68.

[28] E.P. Wigner, On the quantum correction for thermodynamic equilibrium, Phys. Rev. 40 (1932) 749–759.

[29] J. Ville, Theorie et applicacions de la notion de signal analytique, Cable Transmission 2A (1948) 61–74.

[30] M. Born, P. Jordan, Zur Quantenmechanic, Z. Phys. 34 (1925) 858–888.

[31] C.H. Page, Instaneous power spectra, J. Appl. Phys. 23 (1952) 103–106.

[32] W. Rihaczek, Signal energy distribution in time and frequency, IEEE Trans. Inf. Theory IT-14 (1968) 369–374.

[33] L. Cohen, Generalized phase-space distribution functions, J. Math. Phys. 7 (1966) 781–786.

[34] B. Boashash, B. Lovell, L. White, Time frequency analysis and pattern recognition using singular value decomposition of the Wigner–Ville distribution, Adv. Algorithms Architect. Signal Process., Proc. SPIE 828 (1987) 104–114.

[35] G. Cristobal, J. Bescos, J. Santamaria, Application of Wigner distribution for image representation and analysis, Proc. IEEE Eighth Int. Conf. Pattern Recogn. (1986) 998–1000.

[36] K.B. Yu, S. Cheng, Signal synthesis from Wigner distribution, Proc. IEEE ICASSP 85 (1985) 1037–1040.

[37] P. Boles, B. Boashash, The cross Wigner–Ville distribution—a two dimensional analysis method for the processing of vibrosis seismic signals, Proc. IEEE ICASP 87 (1988) 904–907.

[38] O. Kenny, B. Boashash, An optical signal processing for time–frequency signal analysis using the Wigner–Ville distribution, J. Elec. Electron Engng (1988) 152–158.

[39] E. Kavallieratou, N. Fakotakis, G. Kokkinakis, Skew angle estimation using Cohen's class distributions, Pattern Recogn. Lett. 20 (1999) 1305–1311.

[40] M. Shridar, F. Kimura, Handwritten address interpretation using word recognition with and without lexicon, Proc. IEEE Int. Conf. Syst., Man Cybernet., Piscataway, NJ, USA 3 (1995) 2341–2346.

[41] U. Marti, H. Bunke, A full English sentence database for off-line handwriting recognition, Proc. Fifth Int. Conf. Doc. Anal. Recogn., ICDAR, Bangalore (1999) 705–708.