Using skeleton and Hough transform variant to correct skew in historical documents

Omar Boudraa, Walid Khaled Hidouci, Dominique Michelucci

Using skeleton and Hough transform variant to correct skew in historical documents

Omar Boudraa[a,*], Walid Khaled Hidouci[a], Dominique Michelucci[b]

[a] *Laboratoire de la Communication dans les Systèmes Informatiques, Ecole nationale Supérieure d'Informatique, BP 68M, 16309, Oued-Smar, Alger, Algérie. http://www.esi.dz.*
[b] *Laboratoire LIB, Université de Bourgogne, BP 47870, 21078, DIJON CEDEX, France.*

Abstract

As a main part of several document analysis systems, Skew estimation represents one of the major research defies, particularly in case of historical documents exploration. In this paper, we propose an original skew angle detection and correction technique. Morphological Skeleton is introduced to considerably diminish the amount of data by eliminating the redundant pixels and preserving only the central curves of the image components. Next, the proposed method uses Progressive Probabilistic Hough Transform (PPHT) to find image lines. At the end, a specific procedure is applied in order to measure the global skew angle of the document image from these identified lines. Experimental results demonstrate the accuracy and the effectiveness of our approach on skew angle detection upon three popular datasets covering many types of documents of diverse linguistic writings (Chinese, Greek and English) and different styles (horizontal or vertical orientations, including figures and tables, multi-columns page layouts).

Keywords: Skew estimation, Document image analysis, Skew correction, Progressive Probabilistic Hough Transform, Morphological Skeleton

*Corresponding author at : Ecole Doctorale (STIC), Ecole nationale Supérieure d'Informatique, BP 68M, 16309, Oued-Smar, Alger, Algérie. http://www.esi.dz.
Email address: o_boudraa@esi.dz (Omar Boudraa)

1. Introduction

Historical documents include invaluable data and informations; accordingly, they remain a fundamental reference for cultural, literary and scientific information retrieval process. However, when a document is digitized, small skew degrees are inevitable. This imprecision is often due to the scanning device tolerance or the false alignment.

Skew estimation and correction operation is a crucial step in most of document analysis systems [1]. By way of illustration, this process can correct this scanning default and enhance the outcome quality of OCR system. Nonetheless, common complications for skew detection systems are to handle with multipart documents comprising non-textual elements, namely: graphics, forms, tables, figures, or even noise.

In our work, we propose to take benefit of Morphological Skeleton and Progressive Probabilistic Hough Transform (PPHT) in skew estimation. Skeleton decreases significantly the quantity of information to treat and conserves the topology of original patterns, which may reduce the non-textual data presence impact; therefore, they can provide more precision to our method [2]. Moreover, PPHT is later used to calculate the skew angle because of its accuracy and its acceptable runtime [3].

This article is structured as follows: In Section 2 we present a survey of previous works dedicated to skew detection subject. Section 3 furnishes a full description of our proposed system. While, Section 4 is devoted to experimentations, tests, certain statistical analyses and an objective comparison between our contribution and other current methods. At the end, a critical discussions and final conclusions are given in Section 5.

2. Related work

Over the last few years, a number of skew detection techniques were proposed. Predominantly, these methods may be categorized into the undermentioned:

(1) Projection profile analysis based methods (2) Nearest-neighbor (NN) clustering based methods (3) Hough transform based methods (4) Cross-correlations based methods (5) Morphological transform based methods (6) Analysis of the background of documents images based methods (7) Statistical mixture model based methods (8) Principal component analysis based methods (9) Radon transform based methods (10) Fourier transform based methods.

2.1. Projection profile analysis based methods

This formalism is very used to approximate the document inclination angle. Mathematically, Projection profile (also called: Histogram) is the result of the projecting action of a two-color image into a one-spatial vector [4].

So as to decrease the measured time, Bloomberg *et al.* proposed some revisions on the elementary form done by Postl [5]. They consist initially of smoothing the image. Then, an exploration procedure is applied to compute projection profiles over some angles. The detected inclination is done by a finer search algorithm that tries to maximize a defined function [6].

To find the skew angle, using Jain and Borah method [7], the projection profile is calculated in a set of pivoted Skeleton images over multiple degrees. The angle corresponding to the maximum value of a criterion function defines the inclination angle of the document text.

Based on hybridization of advanced vertical and horizontal projection profiles, Papandreou *et al.* developed a new technique for skew evaluation in printed documents images, which involves the minimum bounding box area that was used as fundamental criterion [8].

This approach is simple and appropriate for understandable structure documents, but in some cases, it is not suitable for multipart documents containing graphics or other entities.

2.2. Nearest-Neighbor (NN) clustering based methods

Hashizume *et al.* originally proposed a descendant technique based on K Nearest Neighbors. It commences by detecting all the connected-components

(CCs) in the entire document. After that, for each one, authors computed the orientation of the straight line which links it to its spatially Nearest Neighbors CC. Lastly, a histogram of all these orientations is calculated, in where the peak value refers to the inclination angle [9, 10].

Liolios *et al.* attempted to cluster all CCs placed in the same text line. However, this algorithm is able to handle only with a relatively identical font size writing, due to the using of the average width and height of the CCs in their system [11].

Lu and Tan proposed a NN algorithm, which can deal with documents of diverse languages and writings systems like English and Chinese. First of all, they suggested to restrict sizes of detected nearest-neighbors (NN) of components. Once the adjacent NN pairs are located, the inclination angle is finally given by the NN chains slope having the biggest components number [12].

This class of methods is applicable in multi-columns documents. In addition, it can distinguish several skew angles in the same document. Nevertheless, it has some weakness as its dependence on the thresholding quality issue, the prohibitive runtime and the noise influence [13].

2.3. Hough transform based methods

This well-known image processing technique allows detecting straight lines and other forms (e.g. curves and multilines) in an image. It consists of passing from Cartesian coordinates (x, y) to Polar coordinates (ρ, θ) ; accordingly, a line is defined, via the geometrical conversion:

$$\rho = x \cos\theta + y \sin\theta \tag{1}$$

After quantization of (ρ, θ), 2D Table Accumulator $[\rho][\theta]$ is used to count the number of foreground pixels (x, y) lying on a (ρ, θ) line. Its Peaks values define the estimative skew angle [14, 15, 16, 17].

Le and Thoma proposed a novel algorithm which can identify the page disposition (Landscape / Portrait) as well as the inclination angle degree of

binarized documents images by means of Hough transform. Moreover, the page disposition feature which is based on internal document analysis is employed in the skew detection procedure [14].

To diminish computational costs (space and time) required in Hough transform technique, Amin and Fisher proposed to assemble the neighboring pixels within connected-components and restrict its application only to the last segment rather than the complete partitioned text blocks [15].

Not long ago, Boukharouba has elaborated an innovative technique to identify the inferior Arabic documents baselines. At the beginning, the skew angle is fixed via Randomized Hough Transform running to vertical black–white mutations inside converted binary image. So, after rectifying the tilted image, the baselines are drawn using the Horizontal Projection Profile [16].

These methods are more accurate and less affected by noise. In return, they are to a certain degree costly in terms of assigned memory space and consuming time.

2.4. Cross-Correlations based methods

The inclination approximation in Cross-correlations based methods is established by computing vertical deviations inside the document image [18]. It has interested a part of specialists to adopt it in the handling of this problematic [18, 19, 20].

In Yan scheme [18], for each pair of lines, the correlations in the vertical orientation belonging to distinctive angles ranges are firstly calculated and accumulated. At that moment, the inclination angle is given by the shift having the maximum number of counted cross-correlations.

To ameliorate the cross-correlations precision (as a consequence of the graphics existence in certain documents images), and to decrease the running time, Chaudhuri and Avanindra proposed to choose arbitrary small regions in the image to evaluate the interline cross-correlations rather than the full image lines [19].

2.5. Morphological transform based methods

With the purpose of eliminating some typographical entities (such as: the descenders and the ascenders) from text objects, Chen and Haralick implicated a preprocessing stage that comprises an iterative morphological transforms [21]. The main goal is to obtain CC for all text lines. For this reason, a least square algorithm is applied to associate a line to all detected CCs. Then, by using a special examination procedure, the sloping angle of the entire document is extracted from the histogram of the diverse detected lines orientations.

Morphological transforms have the advantage to manage facilely grayscale images. But, these operations can affect negatively the shape of some images structures.

2.6. Analysis of the background of documents images based methods

This set of methods tries to inspect the geometrical information dispersion (including: texts, tables, graphics, etc.) zones within the document.

Primarily, Chou *et al.* [22] proposed a quadrilateral enclosing algorithm. After partitionning the document into blocks, parallelograms are then created in each element of the document at different orientations to opt for the best.

Mascaro *et al.* [13] suggested a variant of this method, which makes it more robust against noise and non-word-based objects presence. Likewise, it remarkably decreases the consuming time.

Latterly, Shafii has presented multi-steps algorithm, it is composed of: Preprocessing, Text boundary extraction, Edge pixels spotting, Parallel rectangle area computation and Skew approximation. Shafii was focused on script information as a substitute of background analysis[23].

2.7. Statistical mixture model based methods

Egozi and Dinstein proposed an analytical approach to detect the inclination, where Gaussian noise is used to contaminate each object which is represented by a straight line segment [24]. The Expectation Maximization (EM) algorithm [25] is executed to approximate the statistical model factors and the skew angle

is obtained from the histogram of the inclination angles concerning the detected lines.

The previous method can accept the inconsistency of the handwriting character characteristics. Nevertheless, the ideal lines number to be set in the mixture model as well as the local maximum convergence, in some situations, are the main issues of this approach [24].

2.8. Principal component analysis based methods

The Principal Component Analysis (PCA) is an efficient technique in Data Analysis, especially used in identifying patterns in high dimensions data, and in comparing them. PCA method was demonstrated its importance in particular fields such as : face recognition, image compression and some other image preprocessing methods [26, 27].

The two works done by (Burrow and Bodade *et al.*) are the most common PCA based methods for skew angle removal [28, 29].

Recently, Bodade *et al.* have proposed five submodules in PCA based method, these submodules consist of: Preprocessing, PCA, Skew Correction, 3D Correction and Character Segmentation: all the irrelevant areas of car plate image are removed by using a preprocessing algorithm which aimed to mask the central pixel in a group of same pixel row way and column way respectively [29].

In front of aforementioned pros, the major con of the PCA based method is the inconstency in case of poor visibility conditions.

2.9. Radon transform based methods

Similarly to Hough transform, Radon transform function computes projections of an image matrix along specified directions [30].

Raducanu *et al.* proposed an algorithm for document skew detection and correction using Straight-forward technique (to compute the connected objects) and Radon transform integrating heuristic for comparing between different angles [31]. In the same context, Patel *et al.* proposed an application

of Radon transform on binary images after detection of theirs edges using Canny edge detector, having different skew angles ranging from 0° to 90°. The detected skew angle corresponds to the maximum value in Radon transform coefficients [30].

2.10. Fourier transform based methods

In image processing, Fourier transform is used to convert image spatial domain to frequency domain by decomposing it into its sine and cosine elements [32, 33].

Fabrizio proposed a novel method that is based on Fourier transform and $K-NN$ clustering [34], by taking benefits of the fact that the main frequencies of the image are linked with the orientation of image elements. Thus, in order to reduce noise in the frequency domain, the author [34] improved results by using convex hulls of clustered image regions instead of the original image.

Although Fourier transform is robust to noise and some image degradations, its key limitation is that it does not consider adequately the location of signals in images.

3. Document inclination detection and correction approach

In this part, we introduce our approach. It is basically consisted of Morphological Skeleton running through a repetitive thinning technique comprising the two elementary morphological transformations: namely, Erosion and Dilatation [2], cooperated with a modified Standard Hough Transform (SHT) version that is called Progressive Probabilistic Hough Transform (PPHT) as proposed by Matas *et al.* [35]. Our choices arguments and a full explanation of our algorithm steps are correspondingly done in Section 3.1 and Section 3.2.

3.1. Motivations and justifications

Our proposed approach has its particular benefits when compared to others. In fact, Morphological Skeleton tries to maintain the form topology

while contour pixels are excluded. Consequently, processing may be quicker and needs less space [2]. Equally, this preprocessing can reduce the impact of non-textual information existence. Skeleton methods can mainly be classed into three sets: *Distance-based skeleton methods, Median-based skeleton methods* and *Thinning-based skeleton methods* [36].

Based on Euclidean Geometry, for a given subset of component points, in the first class, the skeleton of a CC is the set of centers of maximal disks in the region. The second class attempts to link preselected key points via median lines in non iterative mode. However, the selected technique belongs to the third class. This selection is justified by Morphological Skeleton speed, and that is commonly used in literature as well as its results are satisfactory enough.

For inclination degree detection stage, Hough transform seems good choice forasmuch its exactness and forcefulness, nevertheless, many efforts has been done with the aim of overcoming its essential weakness (namely: computing time). A performant solution is to limit the pixels list contributing for the voting phase into a list of randomly selected pixels. This is referred to as Probabilistic Hough Transform (PHT) [3].

The performed method incorporates a Hough Transform revised version named Progressive Probabilistic Hough Transform (PPHT): the number of extracted lines is bounded by an approval threshold that is generally represented by the total number of vote-pixels [35].

3.2. Proposed method

The complete phases of our proposed document inclination angle estimation and correction algorithm are exposed and clarified thereafter (see Figure 1).

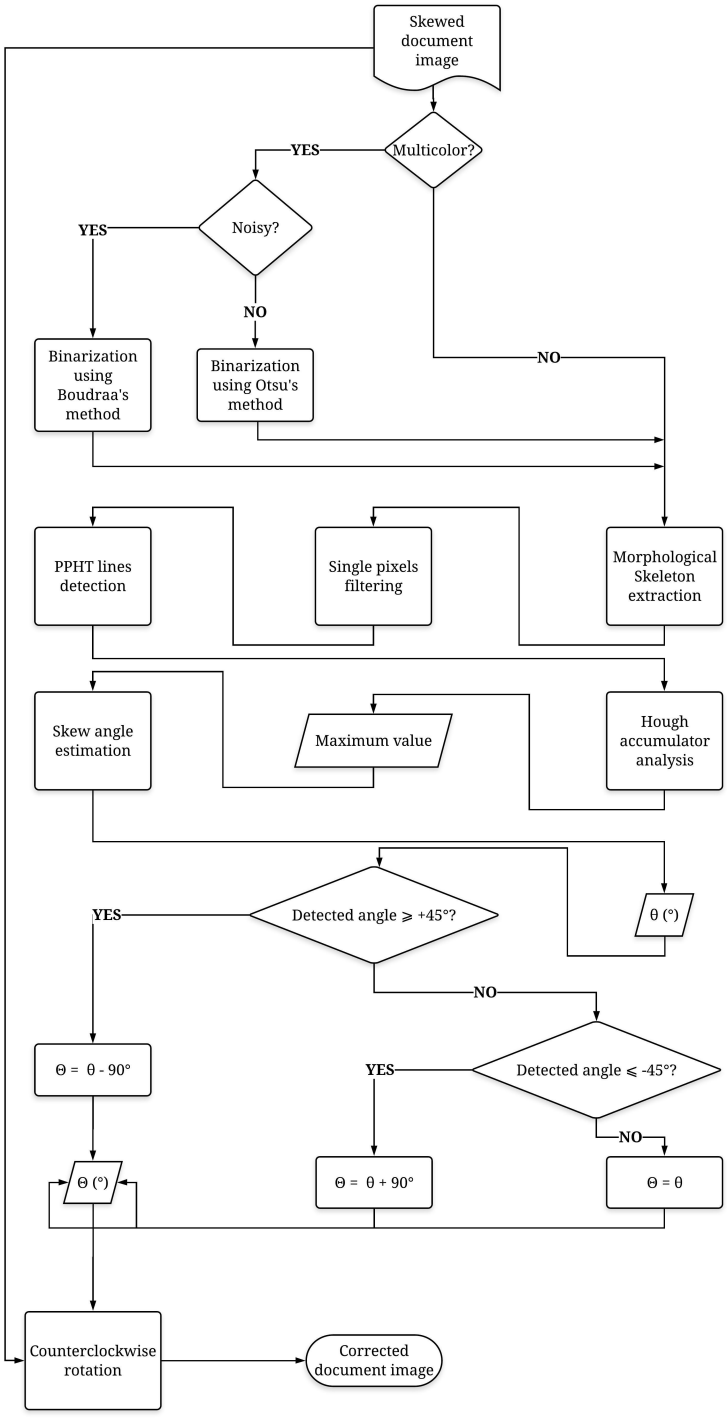


Figure 1: Flowchart of proposed Framework.

3.2.1. Image binarization

Since the input document is generally multicolor, it must be first compressed. This preprocessing task attempts to distinct the foreground from the background. Hence, this process should ideally provide the background in white color and the foreground text in black color [37].

Recently, Boudraa *et al.* suggested a robust binarization method [38]. It is essentially based on a hybrid thresholding approach using three efficient techniques (namely: Otsu, Multi-Level Otsu and Nick). This method also incorporates contrast enhancement and post-processing step that are used to correct and improve the results. The big pixels agglomerations are processed by analyzing the connected components; and a second binarization is then performed using a local Nick's method [39]. Its formula is given by:

$$T = m + k \times \sqrt{\frac{\left(\sum_{i=1}^{NP}(p_i^2 - m^2)\right)}{NP}} \tag{2}$$

Here, k is the Nick variable ($-0.2 \leqslant k \leqslant -0.1$), p_i the intensity value of pixel i, NP the local window pixels number and m the mean intensity value. Unlike the Niblack's method, Nick's method ameliorates appreciably the binarization result by minimizing the false classification of non textual objects as foreground components in case of light images.

It is proved experimentally in [38] that this technique conquers many other binarization algorithms.

In our implementation, if the original image is smooth and neat, we choose global Otsu's Method. Otherwise (i.e. in case of degraded images), we use hybrid Boudraa's Method. Furthermore, since the relevant information (foreground text) in the binary image is set to 0 (black intensity), we use a mask image containing inverted binary image rather than the obtained binary image, because both Morphological Skeleton and PPHT deal with white intensity pixels. While, this mask image remains hidden in the developed GUI and it is placed on the algorithm background.

3.2.2. Skeleton extraction

A repetitive thinning-based technique using an aggregation of morphological (*Opening* and *Erosion*) and Boolean operators (Or (|), And (&) and Not (!)), which represents an application of Lantuejoul's formula [40], is then used:

Algorithm 1 Repetitive Morphological Skeleton [41]

Require: Binary image

Ensure: Single pixel width image skeleton generation

1: $img \leftarrow input\_binary\_image$;
2: **while** (Not_empty(img)) **do**
3: $skel \leftarrow skel \mid (img \ \& \ (!(Opening(img))))$;
4: $img \leftarrow Erosion(img)$;
5: **end while**
6: **return** $skel$;

In this procedure, an *Opening* transformation is a composition of *Erosion* succeeded by *Dilation* using the identical *Structuring Element*, that is assigned to a 3x3 cross-shaped *Structuring Element* (i.e. we adopt 4-connexity). For each reiteration, the skeleton is compressed by calculating the union of the present *Erosion* less the *Opening* using this *Erosion*, the current binary image is eroded once more until this image has empty-white pixels.

3.2.3. Isolated pixel filtering

The retained components gotten at the issue of the preceding step are reviewed to identify and eliminate any segregated single pixels. However, accent signs might be accidently removed perceiving them like isolated pixel.

3.2.4. Lines detection

To extract lines, we run PPHT algorithm, as designated in [35]. This alternative is preferred in case of images having a few extended lines. Here, rather than taking all the points into consideration in voting step, only an random subset is considered. Additionally, each line is characterized by its extremal points (see Figure 2).

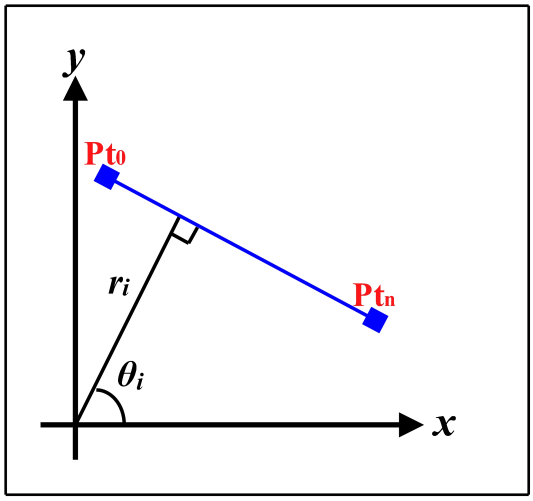


Figure 2: Line spotting in PPHT algorithm.

While, this algorithm includes three factors to fix manually [42] :

- ***Threshold:*** Lines which acquire enough votes number (> threshold) in the accumulator are the only retained.
- ***MinLineLength:*** Lines with votes smaller than this value are excluded.
- ***MaxLineGap:*** Largest acceptable space between interior fragments within the same line.

3.2.5. Hough accumulator array analysis

Here, we inspect the Hough accumulator to bring out its maximum value, relating to the line which defines the global inclination angle of the document image, knowing that Standard Hough Transform (SHT) and its relevance; depend on the fact that the greatest number of collinear points are on or near to the text baseline [3].

3.2.6. Skew angle estimation

To approximate the overall skew of the document, we need to compute the angle between the chosen line that is represented by its extremal points and the

X-axis, by using a geometrical formula:

$$\theta = \arctan\left(\frac{y_n - y_0}{x_n - x_0}\right) \times \frac{180^\circ}{\pi} \tag{3}$$

where, (x_0, y_0) and (x_n, y_n) are the Cartesian coordinates of the two ends of this straight line. Here, right part of above equation is appended to have degree representation ($^\circ$), instead of the radian representation (π).

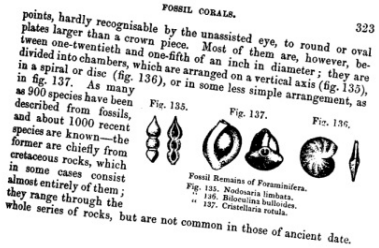
Additionally, with the purpose of making it able to work with different text orientations (i.e. vertical and horizontal), we associate a supplementary equation, as:

$$\Theta = \begin{cases} \theta + 90^\circ & \text{if } \theta \leqslant -45^\circ \\ \theta - 90^\circ & \text{if } \theta \geqslant +45^\circ \\ \theta & \text{Otherwise} \end{cases} \tag{4}$$

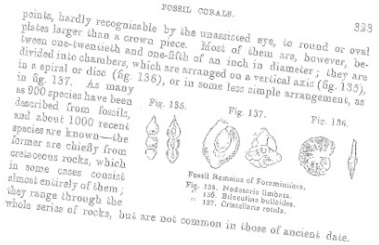
This extension is in some pratical cases beneficial, because some traditional common East Asian letters (such as: Chinese, Japanese, Mongolian and Korean) are written vertically running from top to bottom and starting at the right side of document.

3.2.7. Skew correction

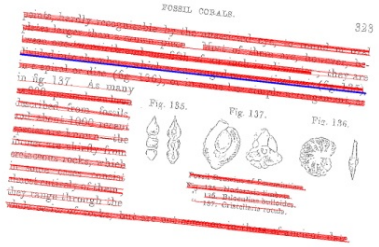
As a final step, to rectify this inclination, the initial image is rotated by the detected angle, considering the counterclockwise as direct (positive) way. Figure 3 illustrates an exhaustive example of skew detection by mean of our suggested method using as input an historical printed document extracted from PRIMA 2011[43].



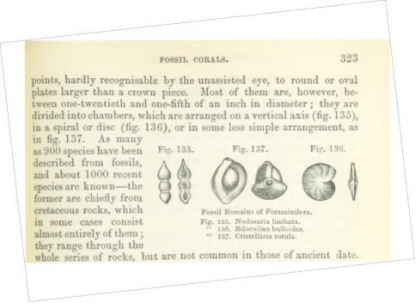
(a)



(b)



(c)



(d)

Figure 3: Detection and correction of inclination (case of historical printed document) [43]: (a) Binary image (Slope angle = 6.20°); (b) Morphological Skeletons & small pixels cleaning; (c) Lines detection, Blue line denotes the peak value; (d) Image redressing.

4. Experimental results

Here, a recapitulation of performed tests and results is done to prove the efficacy and the powerfulness of our skew estimation approach. Therefore, we test it on a diversity of document images (historical and present, clear and blurred, good and degraded, simple and composite) taken from three well-known benchmarks (see. Section 4.1). The efficiency of all skew estimation techniques depends on four common evaluation criteria (as showed in Section 4.2).

Addedly, PPHT factors were adapted conforming to input image layouts and were fixed experimentally, in a directed and careful manner, seeking to obtain high correctness outcomes.

At that moment, we compare our technique to a Histogram-based method, like those suggested by Postl [5], a method based on the analysis of the background area (i.e. Mascaro *et al.* [13]), Edge pixels spotting and text areas calculation-based method as proposed by Shafii [23], as well as a Standard Hough Transform-based method applied on extracted contours of objects using Canny filter [3]. We selected these techniques since they are popular and widely-used for inclination estimation and a lot of algorithms are based on them. Moreover, with the intention of raising the precision rate, we used a scale of 0.1° instead of 1° in the search process. We similarly compare our approach with the top three best-performing methods in (DISEC 2013) Competition; a Fourier transform based method (**LRDE-EPITA-a**), a Maximum-likelihood Estimation based method (**Ajou-SNU**) and a Standard Hough Transform based method (**LRDE-EPITA-b**) [44](see. Section 4.3).

As a final point, a brief runtime study of our approach comparing to the four implicated methods is given in Section 4.4.

4.1. Datasets

For tests, we reassembled a wide range of documents covering various types of data and some particular defies, for instance: vertical and horizontal text-lines, pictures, charts, diagrams, multi languages and casts, including or not noise and shade, taken from three famous datasets.

The first dataset comprises 200 binary images of the DISEC 2013 sample that were nominated from the benchmarking set and were distributed to the contestants in order to test their methods in the event. These images were arbitrarily pivoted in ten distinct angles, sorted out from -15° to $+15^{\circ}$. Likewise, they comprehend several outlines and languages like Latin, Chinese and Greek [44].

The second dataset consists of a part of historical images of famous dataset (PRIMA 2011) [43]. The database comprises samples with undulated text-lines and different-orientation paragraphs, from which we selected 8 images. Hence, we rotated these images at ten various angles to provide skewed images; the peak angle was limited to $\pm 20^{\circ}$. Consequently, a set of 80 images was created.

The third dataset contains a synthetic document images that has been produced from Epshtein work [45]. The dataset is generated from 8 images, with added Gaussian noise and blur. All images are pivoted 10 times inside a range of -15° to $+15^{\circ}$. Shown in Figure 4 are some samples of the tested images.

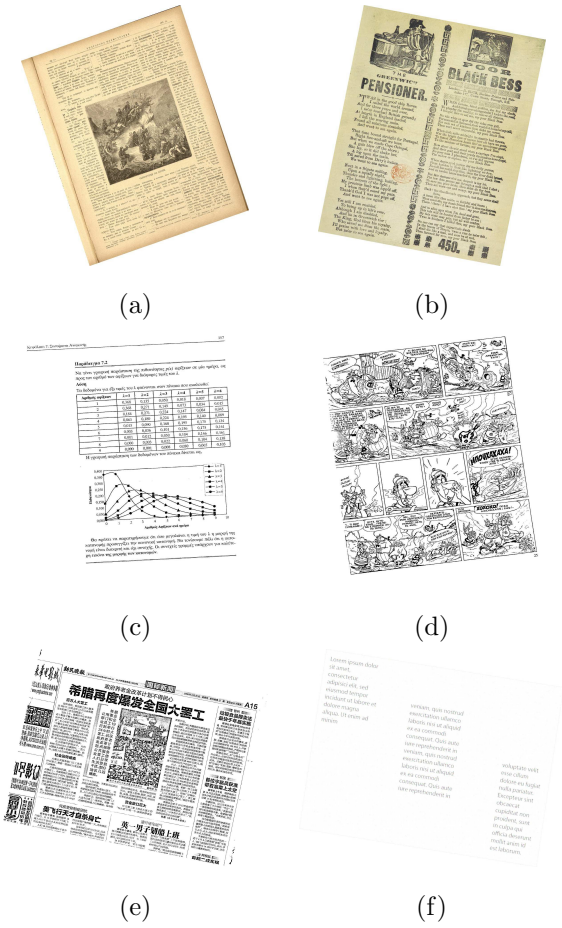


Figure 4: Sample of tested images: (a) Historical Russian printed book page with graphic (Skew angle = 18.90°); (b) Historical English printed document (Skew angle = 11.10°); (c) Document with tables (Skew angle = 4.41°); (d) Document with dominant graphics (Skew angle = 7.63°); (e) Newspaper with Chinese script, horizontal and vertical text orientations (Skew angle = -11.66°); (f) Noisy and low contrast document image (Skew angle = -8.21°).

4.2. Evaluation criteria

To validate the inclination angle estimation performance, achieved results are valuated and then compared by mean of the succeeding indicators, as recommended in [22, 44]:

- **Average Error Deviation (AED):** this norm calculates the mean error deviation, as next:

$$AED = \frac{\sum_{j=1}^{N} E(j)}{N} \tag{5}$$

where, $E(j)$ symbolizes the unsigned difference rounded to the second unit, between the real angle and the detected angle for the image j in a set of N images to evaluate.

- **Average Error Deviation of the Top 80% (ATOP80):** it quantifies the mean error deviation for the top 80% detections, as equation:

$$ATOP80 = \frac{\sum_{j=1}^{0.8\times N} E_{80}(j)}{0.8 \times N} \tag{6}$$

Here, $E_{80}(j)$ is found by arranging the differences in $E(j)$ from minor to major value. This new class is used to avoid account of outliers.

- **Variance of Error Estimation (VEE):** it specifies the variation quantity or the error distribution, as:

$$VEE = \frac{\sum_{j=1}^{N} (E(j) - AED)^2}{N} \tag{7}$$

- **Variance of Error Estimation of the Top 80% (VTOP80):** in this case, the images with the worst 20% of errors estimations are dropped from the calculation, this criterion is given by:

$$VTOP80 = \frac{\sum_{j=1}^{0.8\times N} (E_{80}(j) - ATOP80)^2}{0.8 \times N} \tag{8}$$

4.3. Numerical results

Table 1 shows the approximating results of the inclination angle concerning the first dataset (DISEC 2013) which involves a wide collection of document styles and writing systems. Withal, this dataset comprises 20 images with vertical text orientation. Our suggested method reached very good results, that was really close to LRDE-EPITA-a method results, and the best method among the six other applicant methods in term of obtained AED, that confirm

obviously its high precision in treating diverse documents types under a big number of inclination angles.

The results belonging to the second dataset are illustrated in Table 2. This dataset includes a selection of images emanated from ancient low-quality Latin documents issued of PRIMA 2011 dataset with both a unique and two-column dispositions and some of them have few pictures. The results display a minor deterioration in all four measures compared to the previous values. Nevertheless, our method achieved somewhat well, with very sufficing value.

The last dataset to be analyzed is that supplied by Epshtein [45]. Its results are registered in Table 3. This dataset contains artificial images with inserted Gaussian noise and shade. Relatively, all three methods were slightly influenced by the presence of degradations, as revealed by the augmented error rates. Even so, our method was the least influenced.

To recap, the AED of the first three competed methods over all tested images and over the best 80% images stay excellent and reasonably constant (numerical values are at interval of 0.051 and 0.983) with an amelioration in case of our suggested method on the first as well as the third datasets (see Figure 5). Equally, the maximum retained AED was (0.279) over the second dataset, which approve evidently its high accurateness and practicality.

Table 1: Absolute Error Degrees (°) of the eight tested methods applied to the first dataset.

First dataset	Our Method	Projection Profile-based method	Background Analysis-based method	Edge-based Method	Standard Hough Transform-based method	LRDE-EPITA-a (winning method)	Ajou-SNU (2^{nd}-rank method)	LRDE-EPITA-b (3^{rd}-rank method)
Average (All images)	0.078	0.290	0.539	1.514	6.120	0.072	0.085	0.097
Average (Top 80%)	0.051	0.195	0.081	0.353	4.374	0.046	0.051	0.053
Variance (All images)	0.004	0.065	8.342	9.540	22.781	0.003	0.01	0.001
Variance (Top 80%)	0.001	0.014	0.003	0.273	12.752	0.001	0.001	0.002

Table 2: Absolute Error Degrees (°) of the five tested methods applied to the second dataset.

Second dataset	Our Method	Projection Profile-based method	Background Analysis-based method	Edge-based method	Standard Hough Transform-based method
Average (All images)	0.279	0.313	0.983	0.123	5.168
Average (Top 80%)	0.188	0.211	0.261	0.070	2.725
Variance (All images)	0.076	0.067	6.821	0.028	38.598
Variance (Top 80%)	0.013	0.018	0.098	0.001	11.312

Table 3: Absolute Error Degrees (°) of the five tested methods applied to the third dataset.

Third dataset	Our Method	Projection Profile-based method	Background Analysis-based method	Edge-based method	Standard Hough Transform-based method
Average (All images)	0.253	0.306	0.359	3.791	5.248
Average (Top 80%)	0.119	0.227	0.141	1.760	3.440
Variance (All images)	0.260	0.043	0.325	23.617	23.016
Variance (Top 80%)	0.008	0.019	0.021	7.565	11.725

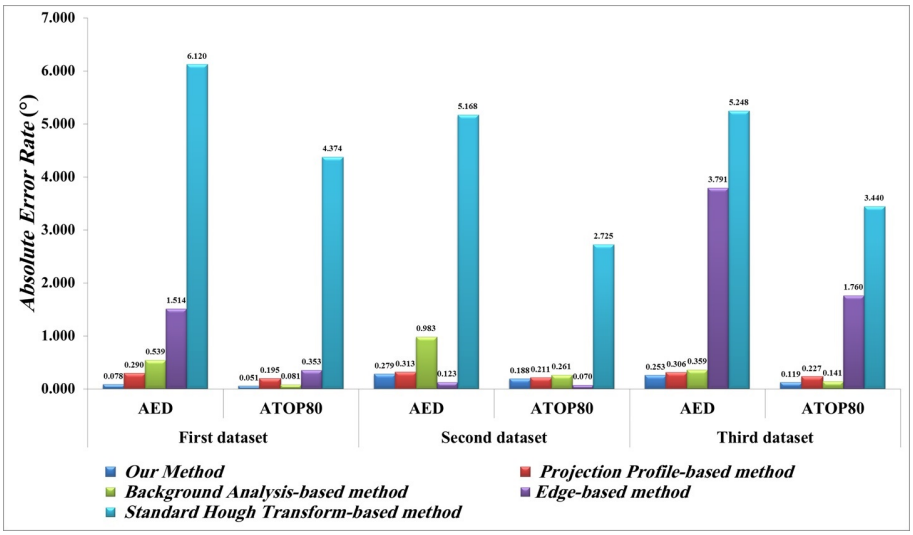


Figure 5: Average errors deviations over all tested images and over the best-80% images.

4.4. Computational time

All the experiments were performed on 64-bits Windows OS machine with 2.0 GHz AMD Quad-Core CPU (4 cores, 4 threads) and 4 GB of memory. We implemented our and Projection-based algorithms in Open Computer Vision (Intel) integrated in QT/C++ platform with no too much optimization. The other algorithms are executed in MATLAB. The registered calculating time (in Sec) using our and the four other methods is arranged in Table 4.

Table 4: Mean and Standard Deviation of Computational time (in Sec) using our proposed method when compared to the other methods.

Runtime (Sec)	Dataset	Our method	Projection Profile-based method	Background Analysis-based method	Edge-based method	Standard Hough Transform-based method
Mean (μ)	First	66.44	67.43	66.42	14.53	12.28
	Second	87.66	59.68	43.29	20.40	18.40
	Third	19.93	18.31	25.15	8.35	7.91
	Global	60.82	54.79	52.11	14.46	12.67
Standard Deviation (σ)	First	39.16	38.03	34.45	3.67	6.64
	Second	51.99	34.84	23.25	12.76	9.20
	Third	15.46	19.18	25.28	4.27	5.88
	Global	45.25	39.25	34.80	7.97	7.95

It can be certainly realized from Table 4 that the registered computational times (in terms of Mean and Standard Deviation) were clearly very close for the first three implemented methods over all datasets implicated in experimentations (the last two methods are faster but less accurate). Nonetheless, the running time was remarkably increased in case of our approach. This might be because of the consuming time during the binarization step which needs in some particular case more material resources in order to produce better results. Another justification is that our implementation uses an interpreted languages such as Matlab need smaller executable code size and less external function calls. Based on these results, we can say that the time consumed by our presented method is reasonable enough.

5. Conclusion and discussions

In this article, we suggested a new solid approach for both skew estimation and correction in either present or historical documents. Our algorithm is consisting of Morphological Skeleton along with Progressive Probabilistic Hough Transform (PPHT).

Based on different tests and experimentations, statistical results demonstrated the efficiency and strength of our method, and revealed that it has attained high accurateness in skew angle estimation over three popular datasets involving several documents types (articles, maps, books, newspapers, etc.) of diverse linguistic writings (such as: English, Greek or Chinese), dissimilar styles (horizontal or vertical alignments, multi-columns layouts, containing figures and forms) and different challenges (noise presence, low luminosity and contrast, etc.). On the other hand, our proposed method has enclosed some insufficiencies as: the algorithm variables number that is considerably high. Furthermore, this implementation can only reveal the global document image slope.

In conclusion, with the aim of enhancing this effort and developing our work,

some interesting prospects may be stated. For instance:

- Improve and develop a new extension of our algorithm allowing to calculate several angles of lines slopes in old handwritten documents.
- Make use of Machine Learning techniques (e.g. Deep Learning) in order to automatically set parameters based on the input image features.

6. Acknowledgments

This research was supported by LCSI and LIB Laboratories. We thank our colleagues from ESI (Algiers, Algeria) and ESIREM (Dijon, France) who provided insight and expertise that greatly assisted the research, although they may not agree with all of the interpretations of this paper.

[1] A. Giotis, G. Sfikas, B. Gatos, C. Nikou., A survey of document image word spotting techniques, Pattern Recognition 68 (2017) 310–332.

[2] A. Rosenfeld., Multiresolution image processing and analysis, Vol. 12, Springer Science & Business Media, 2013.

[3] P. Mukhopadhyay, B. Chaudhuri., A survey of Hough Transform, Pattern Recognition 48 (2015) 993–1010.

[4] R. Gonzalez, R. Woods., Digital Image Processing, 4th Edition, Pearson, London, United Kingdom, 2017.

[5] W. Postl., Detection of linear oblique structures and skew scan in digitized documents, Pattern Recognition Letters (1986) 687–689.

[6] D. Bloomberg, G. Kopec, L. Dasari., Measuring document image skew and orientation, in: L. M. Vincent and H. S. Baird, (Eds.), Proc. SPIE, Document Recognition II, Vol. 2422, 1995, pp. 302–316.

[7] B. Jain, M. Borah., A comparison paper on skew detection of scanned document images based on horizontal and vertical projection profile

analysis, International Journal of Scientific and Research Publications 4 (6) (2014) 1–6.

[8] A. Papandreou, B. Gatos, S. Perantonis, I. Gerardis., Efficient skew detection of printed document images based on novel combination of enhanced profiles, IJDAR 17 (2014) 433–454.

[9] A. Hashizume, P. Yeh, A. Rosenfeld., A method of detecting the orientation of aligned components, Pattern Recognition Letters 4 (7) (1986) 125–132.

[10] F. Farahani, A. Ahmadi, M. Zarandi., Hybrid intelligent approach for diagnosis of the lung nodule from CT images using spatial kernelized fuzzy c-means and ensemble learning, Mathematics and Computers in Simulation 149 (2018) 48–68.

[11] N. Liolios, N. Fakotakis, G. Kokkinakis., Improved document skew detection based on text line connected-component clustering, in: Internat. Conf. Image Process, Vol. 1, 2001, pp. 1098–1101.

[12] Y. Lu, C. Tan., A nearest-neighbor chain based approach to skew estimation in document images, Pattern Recognition Letters 24 (2003) 2315–2323.

[13] A. Mascaro, G. Cavalcanti, C. Mello., Fast and robust skew estimation of scanned documents through background area information, Pattern Recognition 31 (2) (2010) 1403–1411.

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

[15] A. Amin, S. Fisher., A document skew detection method using the Hough transform, Pattern Anal. Appl 3 (3) (2000) 243–253.

[16] A. Boukharouba., A new algorithm for skew correction and baseline detection based on the randomized Hough Transform, Journal of King Saud University, Computer and Information Sciences 29 (1) (2017) 29–38.

[17] P. Yu, V. Anastassopoulos, A. Venetsanopoulos., Pattern recognition based on morphological shape analysis and neural networks, Mathematics and Computers in Simulation 40 (5–6) (1996) 577–595.

[18] H. Yan., Skew correction of document images using interline crosscorrelation, CVGIP: Graphical Models and Image Process 55 (6) (1993) 538–543.

[19] S. Avanindra, B. Chaudhuri., Robust detection of skew in document images, IEEE Trans. Image Process 6 (2) (1997) 344–349.

[20] P. V. D. Grinten, W. Krijger., Processing of the auto and cross-correlation functions to step response, Mathematics and Computers in Simulation 5 (3) (1963) 160–161.

[21] S. Chen, R. Haralick., An automatic algorithm for text skew estimation in document images using recursive morphological transforms, in: Proc. Internat. Conf. on Image Processing, Vol. 1, Austin, USA, 1994, pp. 139–143.

[22] C. Chou, S. Chu, F. Chang., Estimation of skew angles for scanned documents based on piecewise covering by parallelograms, Pattern Recognition 40 (2) (2007) 443–455.

[23] M. Shafii., Optical Character Recognition of Printed Persian/Arabic Documents, Ph.D. thesis, University of Windsor, Canada (2014).

[24] A. Egozi, I. Dinstein., Statistical mixture model for documents skew angle estimation, Pattern Recognition Letters, Elsevier 32 (2011) 1912–1921.

[25] A. Dempster, N. Laird, D. Rubin., Maximum likelihood from incomplete data via the EM algorithm, J. Roy. Statist, Soc. 39, Series B (1977) 1–38.

[26] R. Verma, G. Latesh., Review of Illumination and Skew Correction Techniques for Scanned Documents, in: International Conference on

Advanced Computing Technologies and Applications, ICACTA, 2015, pp. 322–327.

[27] A. Almhdie, O. Rozenbaum, E. Lespessailles, R. Jennane., Image processing for the non-destructive characterization of porous media. Application to limestones and trabecular bones, Mathematics and Computers in Simulation 99 (2014) 82–94.

[28] P. Burrow., Arabic Handwriting Recognition, Ph.D. thesis, University of Edinburgh, England (2004).

[29] R. Bodade, R. Pachori, A. Gupta, P. Kanani, D. Yadav., A Novel Approach for Automated Skew Correction of Vehicle Number Plate Using Principal Component Analysis, in: IEEE International Conference on Computational Intelligence and Computing Research, 2012, pp. 1–6.

[30] J. Patel, A. Shah, H. Patel., Skew Angle Detection and Correction using Radon Transform, International Journal of Electronics, Electrical and Computational System IJEECS 4 (Special Issue ICRDESM-15) (2015) 1–6.

[31] B. Raducanu, C. Boiangiu, A. Olteanu, A. Ștefanescu, F. Pop, I. Bucur., Skew Detection Using the Radon Transform, in: International Conference on Control Systems and Computer Science (CSCS-18), 2014, pp. 1–5.

[32] S. Milan, V. Hlavac, R. Boyle., Image processing, analysis, and machine vision, 4th Edition, Cengage Learning, Boston, Massachusetts, United States, 2014.

[33] S. Belhaj, H. B. Kahla, M. Dridi, M. Moakher., Blind image deconvolution via Hankel based method for computing the GCD of polynomials, Mathematics and Computers in Simulation 144 (2018) 138–152.

[34] J. Fabrizio., A precise skew estimation algorithm for document images using KNN clustering and Fourier transform, in: International Conference on Image Processing ICIP, 2014, pp. 2585–2588.

[35] J. Matas, C. Galambos, J. Kittler., Robust Detection of Lines Using the Progressive Probabilistic Hough Transform, CVIU 78 (1) (2000) 119–137.

[36] G. Klette., Skeletons in Digital Image Processing, Communication and Information Technology Research Technical Report 112, University of Auckland, New Zealand, 2002.

[37] K. Khurshid., Analysis and Retrieval of Historical Document Images, Ph.D. thesis, Paris Descartes University, France (2009).

[38] O. Boudraa, W. Hidouci, D. Michelucci., A robust multi stage technique for image binarization of degraded historical documents, in: In Electrical Engineering-Boumerdes (ICEE-B), 2017 5th International Conference on IEEE, 2017, pp. 1–6.

[39] K. Khurshid, I. Siddiqi, C. Faure, N. Vincent., Comparison of Niblack inspired binarization methods for ancient documents, in: Document Recognition and Retrieval XVI, San Jose, CA, USA, 2009, p. 72470U.

[40] E. Dougherty., An introduction to morphological image processing, SPIE Optical Engineering Press (1992) 161.

[41] F. Abecassis., OpenCV - Morphological Skeleton, `http://felix.abecassis.me/2011/09/opencv-morphological-skeleton/`, [Online; accessed on July-2016] (2011).

[42] The OpenCV Reference Manual, Release 3.0.0-dev, `http://docs.opencv.org/3.0-beta/opencv2refman.pdf`, [Online; accessed on July-2016].

[43] A. Antonacopoulos, C. Clausner, C. Papadopoulos, S. Pletschache., Historical document layout analysis competition, in: Proc. of the International Conference on Document Analysis and Recognition, 2011, pp. 1516–1520.

[44] A. Papandreou, B. Gatos, G. Louloudis, N. Stamatopoulos., ICDAR 2013 document image skew estimation contest (DISEC 2013), in: ICDAR 2013, 2013, pp. 1444–1448.

[45] B. Epshtein., Determining document skew using interline Spaces, in: Proc. of the Int. Conference on Document Analysis and Recognition (ICDAR 2011), 2011, pp. 27–31.

Author biographies

Omar Boudraa is a researcher at the LCSI Laboratory of Heigh School of Computer Sciences (ESI), Oued Smar, Algiers, Algeria. His teaching and research interests are in image processing, historical document analysis, networks and systems administration.

Walid Khaled Hidouci is full professor. Since 2010, he heads the Advanced Data Bases team at LCSI research laboratory. His research interests include: algorithms, database systems, artificial intelligence, parallel and distributed computing and Unix system administration.

Dominique Michelucci has full professor degree and he is a researcher at Computer Science Laboratory of Burgundy (LIB, Dijon, France). His pedagogic and research interests are in image synthesis, geometric transformations and computations of images, modelisation, artificial intelligence, optimization and computer programming.