IJDAR (2014) 17:433–454
DOI 10.1007/s10032-014-0228-5

ORIGINAL PAPER

Efficient skew detection of printed document images based on novel combination of enhanced profiles

A. Papandreou · B. Gatos · S. J. Perantonis · I. Gerardis

Received: 29 October 2013 / Revised: 28 April 2014 / Accepted: 23 July 2014 / Published online: 10 September 2014

Abstract Document skew is often introduced during the capturing process of the document image processing pipeline and may seriously affect the performance of subsequent stages of segmentation and recognition. Skew detection is often accomplished with the use of horizontal projections, while recently, a new approach that is based on vertical projections has been introduced. In this paper, we use the technique of minimum bounding box area in order to combine a horizontal with a new reinforced vertical projection profile method. We are motivated by the fact that the horizontal and the novel vertical projection profiles are found to be complementary to each other. We claim that the proposed approach has more accurate performance compared with other state-of-the-art skew detection algorithms; it deals with all the drawbacks of the projection profile methods; it is more noise and warp resistant and gives accurate results for any kind of printed document image. For these reasons, it can be efficiently applied to historical machine printed or multicolumn documents, documents with figures and tables, while it is robust for any kind of script. Extended experimental results on two databases in different skew angle range, with representative printed documents of all kinds, as well as printed documents of two historical books, prove the efficiency of the proposed approach. There is also a comparison with commercial products in several cases where the contribution of the proposed algorithm is demonstrated at optical character recognition level. Moreover, an analysis of the accuracy performance of the main elements of the proposed technique is also performed.

Keywords Document skew correction · Projection profiles · Document image preprocessing

A. Papandreou
Department of Informatics and Telecommunications,
National and Kapodistrian University of Athens,
Panepistimioupoli, Ilissia, 15784 Athens, Greece

A. Papandreou (✉) · B. Gatos · S. J. Perantonis · I. Gerardis
Institute of Informatics and Telecommunications, National Center for Scientific Research "Demokritos", Agia Paraskevi,
15310 Athens, Greece
e-mail: alexpap@iit.demokritos.gr

B. Gatos
e-mail: bgat@iit.demokritos.gr

S. J. Perantonis
e-mail: sper@iit.demokritos.gr

I. Gerardis
e-mail: gerardis@sch.gr

1 Introduction

In order to proceed with optical character recognition (OCR), document image preprocessing is always essential as a starting step. The task of preprocessing mainly includes the removal of noise as well as image normalization in order to remove unwanted variations, while it can be divided into several steps such as binarization, skew correction, noise removal, and enhancement. In this paper, we focus on the task of skew detection in machine-printed documents.

The detection and correction of the document skew is one of the most important document image analysis steps. Some degree of skew is unavoidable to be introduced when a document is scanned manually or mechanically [1]. This may seriously affect the performance of subsequent stages of segmentation and recognition, while skew of as little as 0.1° may be apparent to a human observer. To this end, a reliable skew estimation and correction technique have to be used in scanned documents as a preprocessing stage in almost all document analysis and recognition systems [2]. Existing skew estimation algorithms mainly consider documents with

plain structure. This limitation appears too restrictive with respect to the diversity of current documents. The requirement of more and more sophisticated tools is promoted by the large diffusion of complex documents characterized by creative geometric layouts [3]. To this end, skew detection is still a challenging task especially for documents with graphics, charts, figures or various font sizes [4]. In the literature, a variety of skew detection techniques are available and fall broadly into the following four categories according to the basic approach they adopt [3,5]: Hough transform [6–10], nearest neighbor clustering [11–14], interline cross- correlation [15–19] and projection profile [20–24]- based methods.

Hough transform (HT) is a popular approach for machine vision. Despite its computational demerit, it has been widely employed to detect lines and curves in digital images. According to Srihari and Govindaraju [6], each black pixel is mapped to the Hough space (ρ, θ) and the skew is estimated as the angle in the parameter space that gives the maximum accumulation along the ρ component. In order to improve the computational efficiency of the HT-based approaches, several variants that reduce the number of points which are mapped in the Hough space have been proposed. Hinds et al. [7] fuse HT and run-length encoding in order to reduce data. They create a burst image which is built by replacing each vertical black run with its length placed in the bottom-most pixel of the run. The bin with maximum value in the Hough space determines the skew angle. Furthermore, Wang et al. [8] use bottom pixels of the candidate objects within a selected region for the HT. Document skew is then estimated by finding a local peak in Hough space. Yu and Jain in [9] fuse hierarchical HT and centroids of connected components. They first compute the connected components and their centroids, and then the HT is applied to the centroids using two angular resolutions. Finally, Singh et al. [10] introduce a block adjacency graph (BAG) in a preprocessing stage and document skew is calculated based on voting using the HT. As mentioned by the authors, the technique is language dependent. Researchers proposed different strategies to reduce the amount of input data, but the computational cost of HT is still very high. Furthermore, prior to applying, it is requisite to extract text regions in the document images. However, this is very hard in document images since layouts may be unknown and complex.

According to nearest neighbor approaches, spatial relationships and mutual distances of connected components are used to estimate the page skew. In the work of Hashizume et al. [11], the direction vector of all nearest neighbor pairs of connected components is accumulated in a histogram and the peak in the histogram gives the dominant skew. This method is generalized by O'Gorman [12] where nearest neighbor clustering uses the K-neighbors for each connected component. However it is accurate for a resolution of 0.5°. Lu and Tan [13] propose an improved nearest neighbor chain-based approach for document skew estimation with high-accuracy and language-independent capability. The methods of this class aim at exploiting the general assumptions that characters in a line are aligned and close to each other. They are characterized by a bottom-up process, which starts from a set of objects and utilizes their mutual distances and spatial relationships to estimate the document skew. Okun et al. [14] propose a method that consists of spatial image-resolution reduction, connected component detection, classification into different categories and text line accumulation. For the last step, a simple technique is applied, which utilizes the first eigenvector of the data covariance matrix. The line slopes are accumulated into an angle histogram, whose peak corresponds to the desired document skew. Few researchers prefer the nearest neighbor clustering techniques for document skew removal since the text lines can be grouped in any direction, and connections with noisy subparts of characters reduce the accuracy of the method. Furthermore, skew detection based on nearest neighbor clustering methods is rather slow due to the component labeling phase that is performed bottom-up and has quadratic time complexity $O(n^2)$.

The interline cross-correlation approaches are based on the assumption that de-skewed images present a homogeneous horizontal structure. Therefore, such approaches aim at estimating the document skew by measuring vertical and horizontal deviations among foreground pixels along the document image. Cross-correlation between lines at a fixed distance is used by Yan [15]. This method is based on the observation that the correlation between two vertical lines of a skewed document image is maximized if one line is shifted relatively to the other so that the character base line levels for the two lines are coincident. The accumulated correlation for many pairs of lines can provide a precise estimation of the page skew. In [16], Gatos et al. proposed that the image should first be smoothed by a horizontal run-length algorithm, and then, only the information existing in two or more vertical lines should be used for skew detection. Chou et al. in [17] proposed piece-wise coverings of objects by parallelograms (PCPs) for estimating the skew angle of a scanned document. A document image is initially split into a number of non-overlapping slabs and subsequently parallelograms at various angles cover the components in each slab. The angle at which maximum white space is obtained indicates the estimated skew angle of the document. In [18], Deya and Noushath have enhanced PCP (e-PCP) with a criterion to classify a document as a landscape or portrait mode, while they have introduced a confidence measure for filtering the estimated skew angles that may not be reliable. Finally, Alireza et al. [19] proposed an algorithm which applies piecewise painting on both directions of a document and creates two painted images. Those images provide statistical information about regions with specific height and width. The points of those regions are grouped, a few lines are obtained

with the use of linear regression and line drawing concepts, and their individual slopes are calculated.

The traditional projection profile approach is a simple solution to detect the skew angle of a document image. It is initially proposed by Postl [20] and is based on horizontal projection profiles. According to this approach, a series of horizontal projection profiles are calculated at a range of angles. The profile with maximum variation corresponds to the best alignment of the text lines. In order to reduce high computational cost, several variations of this basic method have been proposed. Papandreou and Gatos have proposed an improvement of the traditional horizontal projection profile approach by combining it with the minimization of the bounding box area technique [21]. According to this approach, the maximum variations are divided by the bounding box area that includes the text. This amount is minimized when the document is horizontally aligned. Baird [22] proposed a technique for selecting the points to be projected: For each connected component, only the midpoint of the bottom side of the bounding box is projected. An objective function is used to compute the sum of the squares of the profiles. Ciardiello et al. [23] calculate projections on selected subregions, with high density of black pixels per row, of the document image. The algorithm detects the skew angle by maximizing the mean square deviation of the profile. Ishitani [24] 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 to white transitions along the lines. Finally, Papandreou and Gatos [21] have proposed a vertical projection profile approach, taking advantage of the alignment of vertical strokes of the Latin alphabet when the document is deskewed. This approach proved to be more noise and warp resistant.

Projection profile techniques can efficiently deal with clean text, and they are accurate in small angles, while on the other hand, they are computationally expensive, and they are limited to estimating skew angle within $\pm 10^{\circ}$ [2]. Additionally, horizontal projection profile-based approaches cannot deal with noisy or warped documents, with broken characters [2], and they also fail as much as multicolumn documents and documents with figures are concerned. The horizontal projection profile method is a quite accurate skew estimation method as much as plain text document images are concerned since it utilizes the alignment of letters and words in printed lines. It detects the gaps between the lines and the succeeding space when the document is properly deskewed, while the accuracy of the method depends on the width of the document image. Consequently, in multicolumn documents where the lines of different columns are not aligned, in document images with significant noise and in warped documents that do not have straight text lines, there are no gaps to be detected, and the horizontal projections (as well as the Hough transform-based methods) may fail.

On the contrary, the vertical projection profiles [21] can deal with noise, multicolumn documents and tables in a great range of angles accurately and are more robust in warped documents (see Fig. 1). They are based on the right, left or full alignment of a text, on the alignment of vertical strokes or the existence of vertical lines such as borders, and their accuracy depends on the height of the document image. Due to this fact, they are favored concerning multicolumn documents since there is vertical alignment in every column. Furthermore, the noise does not affect the detection of the vertical black runs, which are additionally stressed with the reinforced version of the algorithm. Finally, the warped documents keep their vertical strokes in vertical direction in the correct skew angle, and they mostly maintain their right, left or full alignment of the text.

In Fig. 1a, a typical example of a warped document is demonstrated. As shown in Fig. 1b, the baselines of the document are not straight lines, and consequently, neither horizontal projection profile methods nor Hough transform-based methods can detect the skew of the document accurately. On the other hand, as demonstrated in Fig. 1c, d, when the dominant skew is corrected, the vertical strokes of the letters are almost at 90 degrees.

As mentioned above, horizontal and vertical projection profiles are complementary to each other and can handle accurately all kinds of documents even if they consist of few lines or columns since one approach is accurate depending on the width and the other on the height of the document image. In this work, a new technique based on combined enhanced projection profiles (CEPP) is introduced, and furthermore, the proposed method is accelerated with a coarse-to-fine technique (CEPP c-f) for faster convergence and more profound combination of the projection profile methods. It is based on both, horizontal and vertical projection profiles while we are motivated by their complementarity. The minimization of the bounding box area is introduced as an appropriate criterion for combining these different kinds of profile. The efficiency of the minimum bounding box area criterion, in order to select between horizontal and vertical profiles, as well as the fact that the reinforced vertical projection profiles and the enhanced horizontal profiles are complementary, are proved through exhaustive testing. We also claim that the proposed approach is more accurate compared with state-of-the-art skew detection algorithms. Classical projection profile methods are obsolete due to their restrictions and the disadvantages listed above. Nevertheless, despite the fact that the proposed method falls broadly in the same category, it deals with all the drawbacks of the projection profile methods due to the complementarity of vertical and horizontal profiles. Furthermore, the proposed technique is noise and warp resistant, and it gives more accurate results in any kind of printed document image; it is not constrained by any range of angles, and it is fast due to the proposed coarse-to-fine con-

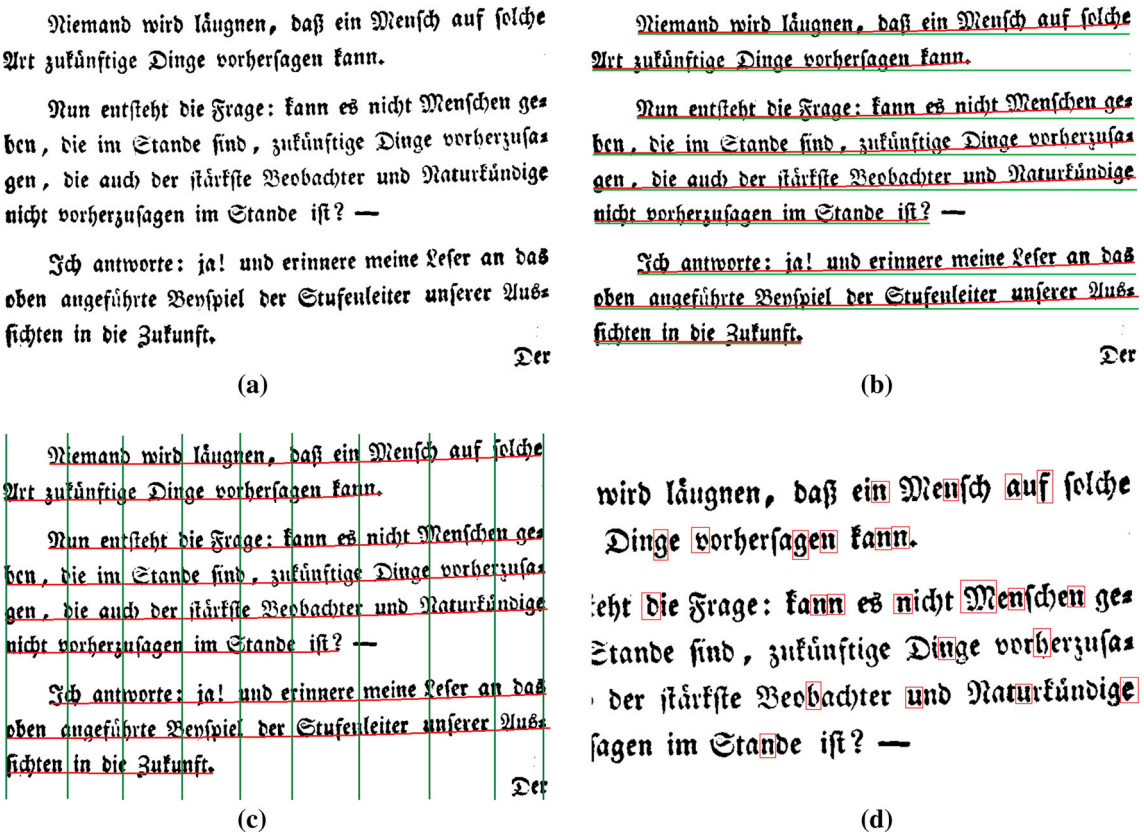


Fig. 1 **a** Warped document having its dominant skew corrected, **b** comparison of the baselines of the document with *horizontal lines*, **c** comparison of the baselines of the document with *vertical lines*, **d** the *vertical strokes of the letters* remain near vertical regardless the curvature of the document

vergence. For these reasons, it can be efficiently applied to historical machine printed or multicolumn documents, documents with figures and tables, while it is robust in any kind of script and in a wide range of angles.

The proposed technique is an extension of our previous work on skew detection in printed documents [21]. The advancement compared with [21] is summarized in the following points: (1) introduction of an innovative criterion which can operate as a skew estimation technique selector. The minimization of the bounding box area proved to be an effective algorithm selector, which can combine two or more different skew estimation techniques. (2) In the proposed method, the vertical projection profiles are reinforced, stressing the existence of black runs, instead of those described in Papandreou and Gatos [21]. (3) In Papandreou and Gatos [21], the maximization of an energy function was used in order to determine the minimum entropy corresponding to the correct skew angle. On the contrary, in the proposed method, the difference between three successive values of each histogram is computed to determine the correct skew angle. Finally, (4) in the proposed method, a coarse-to-fine algorithm is introduced in order to accelerate the convergence of the algorithm. The proposed combination of horizontal and vertical projection profiles is more robust concerning the variations of the document layout and the range of skew angles detected than the methods described in Papandreou and Gatos [21]; a fact that is demonstrated through the extended experimental results and the diversity of the datasets used.

The remainder of the paper is organized as follows. Section 2 focuses on the proposed methodology and it provides a detailed analysis of the steps involved. Extensive experimental results, in multiple datasets and various angles, indicating the accuracy performance of the proposed methodology compared with other state-of-the-art methods are presented and discussed in Sect. 3, while conclusions are drawn and remarks are made in Sect. 4.

2 Proposed methodology

In the proposed CEPP methodology, we combine novel reinforced vertical and enhanced horizontal projection profiles with the use of the minimum bounding box area criterion in order to achieve an accurate and robust document skew detection. The estimation of the foreground area bounding box is commonly used in the literature as a skew estimation technique, since the area of the rectangle that contains all the foreground pixels of a document image is expected to be

Fig. 2 I_x^θ and I_y^θ are the width and the height of the skewed document image

minimum when the document image is properly deskewed [1,25]. Papandreou and Gatos have recently combined it with the classical projection profile technique in order to achieve an improved estimation [21]. In the CEPP approach, we also use it as a criterion, in order to determine which of the two projection profile algorithms (horizontal or vertical), operates better in each case. As demonstrated in the experimental results (Sect. 3.2.5), the bounding box method can perform well in plain text document images but it has low accuracy in documents with complex layout due to the outliers (components that are beyond the main rectangular body of the page). This is why in the proposed work the minimum bounding box area is not used to estimate directly the skew angle but is used mainly as a selection criterion. The main contribution of the bounding box algorithm is the selection between the results of two much more accurate algorithms. Moreover, in our approach, horizontal and vertical histograms are divided by the bounding box area (as in Papandreou and Gatos [21]) in order to normalize the result better. At the correct skew angle there is a great accumulation of foreground pixels in certain lines and columns. Since this may also occur for near diagonal angles where information may be accumulated, we use the division by the bounding box area which, in this case, is larger and thus capable of making a distinction between the correct skew and the near diagonal angle.

In order to estimate the skew, the document image is rotated by a range of angles $\theta \in [\theta_{\min}, \theta_{\max}]$. Let $I^\theta(x, y)$ be the binary document image in angle θ having 1s for black (foreground) pixels and 0s for white (background) pixels and I_x^θ and I_y^θ be the width and the height of the rotated document image (see Fig. 2).

The proposed skew estimation methodology consists of three steps demonstrated in the flowchart in Fig. 3. At a first step, the document image is rotated by a range of angles and the rectangle box that bounds the foreground pixels is computed. Then, two of histograms, a reinforced vertical and an enhanced horizontal, are calculated for every angle. Finally, these histograms are analyzed providing two candidate skew angles, one of which is selected with respect to the minimum bounding box area criterion.

2.1 Computation of the bounding box

In order to compute the minimum rectangle area E^θ that bounds the foreground pixels for each angle θ, we calculate the four extreme points $(x_l^\theta, x_r^\theta, y_u^\theta, y_d^\theta)$ (see Fig. 4) using the following equations:

$$x_l^\theta = \mathrm{x}' : \sum_{\substack{y=1 \\ x=1}}^{\substack{y=I_y^\theta \\ x=x'}} I^\theta(x, y) = 0 \cap \mathrm{x}' = \max \tag{1}$$

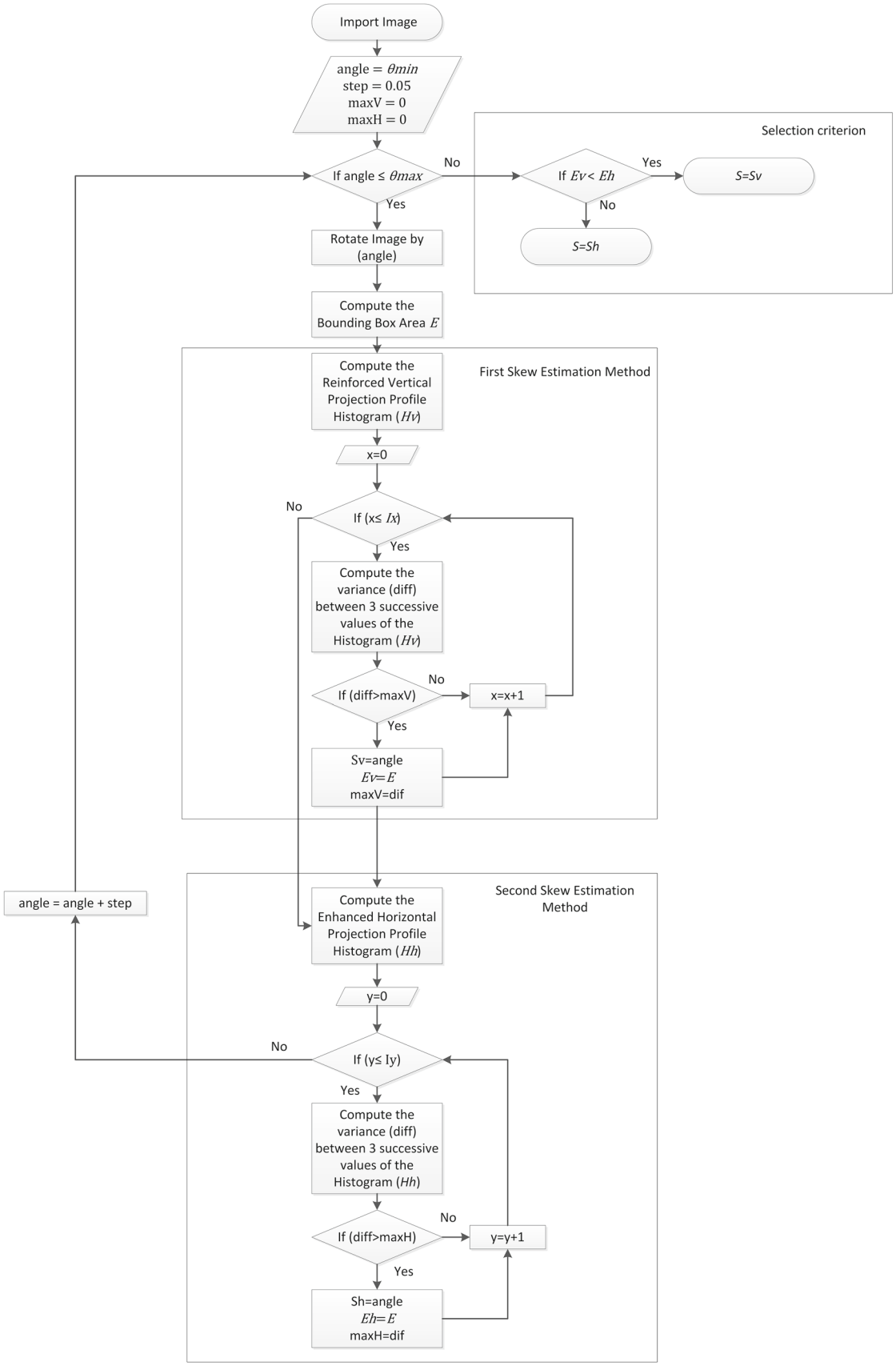


Fig. 3 The flowchart of the proposed CEPP method

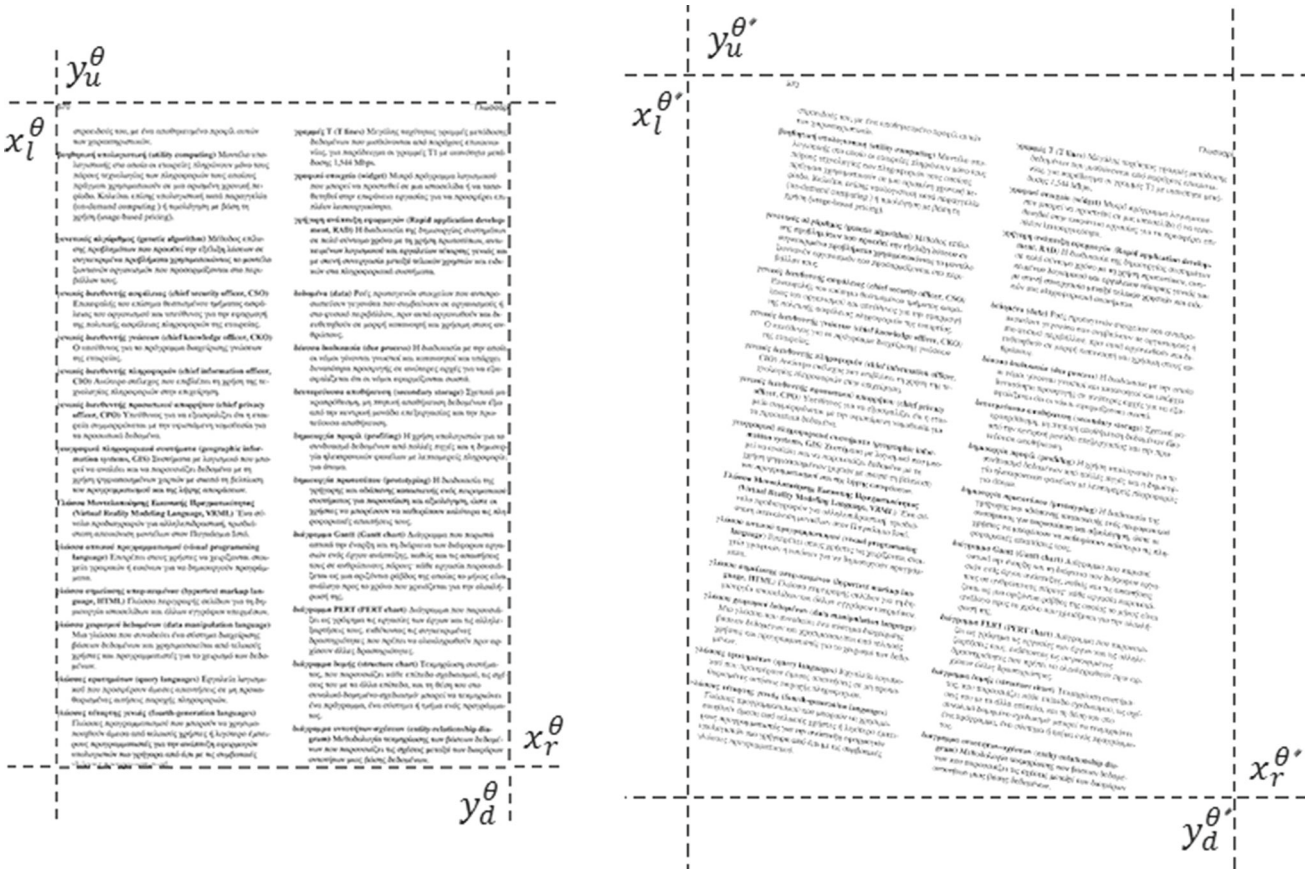


Fig. 4 The *rectangle* that contains all the foreground pixels is defined by the four extreme points $(x_l^\theta, x_r^\theta, y_u^\theta, y_d^\theta)$ and its area is minimum when the document image tends to be properly deskewed

$$y_u^\theta = \mathrm{y}' : \sum_{\substack{y=1 \\ x=1}}^{\substack{y=y' \\ x=I_x^\theta}} I^\theta(x, y) = 0 \cap \mathrm{y}' = \max \tag{2}$$

$$x_r^\theta = \mathrm{x}' : \sum_{\substack{y=1 \\ x=x'}}^{\substack{y=I_y^\theta \\ x=I_x^\theta}} I^\theta(x, y) = 0 \cap \mathrm{x}' = \min \tag{3}$$

$$y_d^\theta = \mathrm{y}' : \sum_{\substack{y=y' \\ x=1}}^{\substack{y=I_y^\theta \\ x=I_x^\theta}} I^\theta(x, y) = 0 \cap \mathrm{y}' = \min \tag{4}$$

$$E^\theta = \left(x_r^\theta - x_l^\theta\right)\left(y_d^\theta - y_u^\theta\right) \tag{5}$$

For each angle θ, after we have calculated the extreme points of the image, we calculate a pair of histograms within these boundaries, the reinforced vertical $H_v^\theta()$ and the horizontal $H_h^\theta()$ histograms, that are detailed in the following section.

2.2 Reinforced vertical projection profile

In order to calculate the reinforced vertical histogram $H_v^\theta()$, we first compute an initial histogram ${}_vH^\theta()$ for each angle θ, in which we focus on the vertical black runs of the document image and stress the existence of long vertical strokes or lines. Those vertical black runs are expected to have significant contribution in the vertical profile, to be in larger number and to have greater values when the document is properly deskewed. The vertical black run profile ${}_vH^\theta()$ is defined as follows:

$${}_vH^\theta(x) = \sum_{i=1}^{B(x)} \sum_{j=1}^{L(i,x)} br(j) \tag{6}$$

where $B(x)$ is the number of black runs in the vertical scan of line x, $L(i, x)$ is the length of the ith black run of line x and $br(j)$ is defined as follows:

$$br(j) = \begin{cases} j, & j \leq 4 \\ 4, & j \geq 4 \end{cases} \tag{7}$$

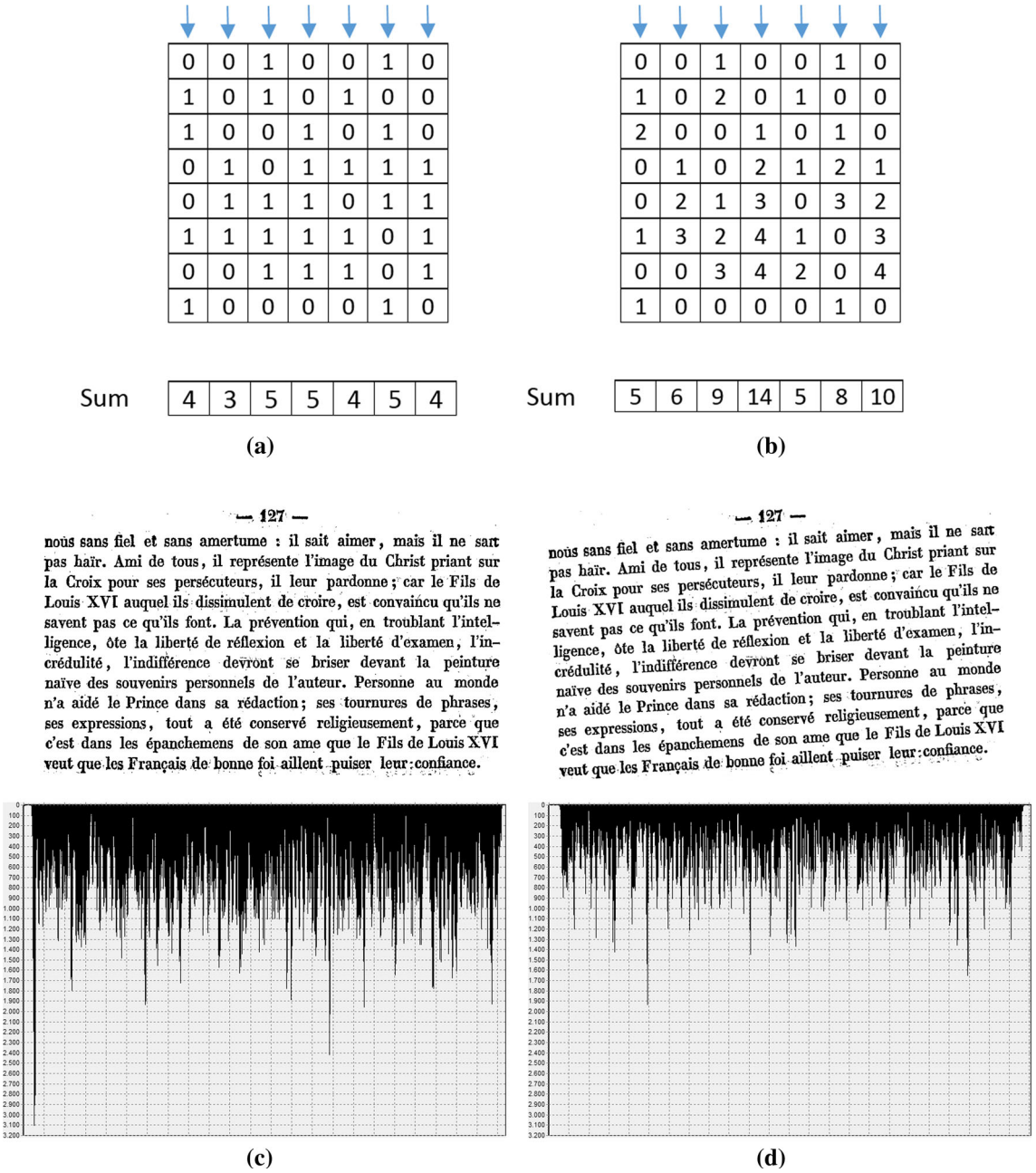


Fig. 5 A comparison between **a** the classical and **b** the reinforced, ${}_vH^{\theta}()$, vertical projection profiles calculation, and an example of the reinforced vertical histogram when **c** the document image is properly aligned and when **d** it has skew

The parameter $br(j)$ is restricted and cannot take values that exceed 4 so that straight lines will not determine the result in an absolute but in a dominant way. Although the value 4 is determined experimentally, it is observed that setting values >4 does not have significant impact on the skew detection performance. A comparison between the classical and the reinforced (${}_vH^{\theta}()$) vertical projection profiles calculation is exhibited, respectively, in Fig. 5a, b, while in Fig. 5 there is also an example of the reinforced vertical histogram when the document image is properly aligned (Fig. 5c) and when it has skew (Fig. 5d).

At a next step, we divide the values of ${}_vH^{\theta}()$ by the area E^{θ} of the box that bounds the foreground pixels in order to calculate $H_v^{\theta}()$:

$$H_v^{\theta}(x) = {}_vH^{\theta}(x)/E^{\theta} \tag{8}$$

As explained in Papandreou and Gatos [21], when the document image is properly deskewed, the area of the bounding box E^{θ} is expected to be minimum (see Fig. 4) and the concentration of foreground pixels will be higher in certain columns of the image. So, by dividing these high values of the histogram by the low values of the area of the bounding box,

the algorithm is facilitated to distinguish the correct skew angle from the rest of the angles where the concentration of the foreground pixels is lower and the area of the bounding box is bigger. Despite the fact that the bounding box minimization criterion may have low accuracy in documents with complex layout due to outliers, the division by the bounding box area is capable of distinguishing between big accumulations of foreground pixels in the histograms due to the alignment in the correct skew angle and accumulations in the near diagonal angles where information may be accumulated.

The reinforced vertical projection profiles take advantage of the black borders of the document image that originate from the image capturing processes (see Fig. 6a, b). In some rare cases where the vertical border (or frame) does not identify the skew angle of the text, the $br(j)$ upper limit assures that the border will have a dominant but not decisive contribution to the calculation of the skew angle. Even if the reinforced vertical projection profiles fail to detect the correct skew angle, the proposed method has the advantage of combining methods. The horizontal projection is likely to find the correct skew angle and in this way the outcome will be correct.

The use of reinforced vertical projection profiles is also suitable for multicolumn documents. The proposed projection profiles, unlike the classical horizontal projection profiles, have extremely accurate performance in multicolumn documents, since this kind of documents has several areas fully aligned. This results in several high peaks which correspond to high concentration of black pixels (see Fig. 6c).

2.3 Horizontal projection profiles

At this step, we calculate the initial horizontal histogram ${}_hH^\theta()$ for each angle θ, which is defined as follows:

$$ {}_hH^\theta(y) = \sum_{x=x_l^\theta}^{x_r^\theta} I^\theta(x, y) \tag{9} $$

and then divide it by the area of the bounding box E^θ in order to form $H_h^\theta()$ as follows:

$$ H_h^\theta(y) = {}_hH^\theta(y)/E^\theta. \tag{10} $$

2.4 Skew estimation and minimum bounding box area criterion

In order to estimate the correct skew of the document image, we seek for the angles θ_v and θ_h in which the difference between three successive values of, respectively, $H_v^\theta()$ and $H_h^\theta()$ histograms is maximized. θ_v, θ_h are defined as follows:

$$ \theta_v = \arg\max_\theta(D_v(\theta)) \tag{11} $$

$$ \theta_h = \arg\max_\theta(D_h(\theta)) \tag{12} $$

where θ varies in the range of the image document skew expected while $D_v(\theta)$ and $D_h(\theta)$ are defined as:

$$ D_v(\theta) = \arg\max_x \Big(\Big| H_v^\theta(x) - H_v^\theta(x-1) + H_v^\theta(x) - H_v^\theta(x-2) \Big| \Big) \tag{13} $$

$$ D_h(\theta) = \arg\max_y \Big(\Big| H_h^\theta(y) - H_h^\theta(y-1) + H_h^\theta(y) - H_h^\theta(y-2) \Big| \Big) \tag{14} $$

where $x \in [x_l^\theta, x_r^\theta]$ and $y \in [y_u^\theta, y_d^\theta]$ and values of $H_v^\theta(x)$ and $H_h^\theta(y)$ with x and y exceeding their range are considered to be zero. We use the difference between three successive values in Eqs. (13, 14) in order to have a smoothed function that is more robust in degradations, significant noise, broken characters, etc.

In order to decide which of the two candidate skew angles, θ_v or θ_h, corresponds to the estimated skew of the document image, we compare the areas of the bounding boxes in each of the two angles E^{θ_v} and E^{θ_h} while the final skew angle θ_s is the one that corresponds to the minimum bounding area.

$$ \theta_s = \begin{cases} \theta_h, & E^{\theta_h} \le E^{\theta_v} \\ \theta_v, & \text{otherwise} \end{cases} \tag{15} $$

2.5 Faster convergence: coarse-to-fine technique

In order to make our method faster and reduce the computational cost, a coarse-to-fine technique, similar to the one proposed by Li et al. [26], is applied. In that way, the number of angles that the document image is rotated is reduced without restricting the range of the angles that the document is rotated. This coarse-to-fine technique, furthermore, enables the extension of the range of the expected skew angle since it accelerates the proposed algorithm multiple times, depending on the range of angles of the expected skew, without affecting significantly the accuracy of the proposed method. The flowchart of CEPP using this coarse-to-fine technique (CEPP c-f) is presented in Fig. 7.

Given that we have a wide range of angles $[\theta_{\min}, \theta_{\max}]$ in which the document skew is expected, a classical projection profile method [20–24] would rotate the document by every angle from $\theta_{\min}$ to $\theta_{\max}$ with a certain step which corresponds to the required accuracy. In that way, the algorithm would have to perform $[(\theta_{\max} - \theta_{\min} + 1) \times 20]$ rotations in order to achieve an accuracy of 0.05°. Instead, in the proposed method the document image is rotated from $\theta_{\min}$ to $\theta_{\max}$ with step 1° and the dominant skew angle θ_{d1} is defined. Afterward, the document image is rotated from $(\theta_{d1} - 0.5)$ to $(\theta_{d1} + 0.5)$ with step 0.5° in order to find the new dominant

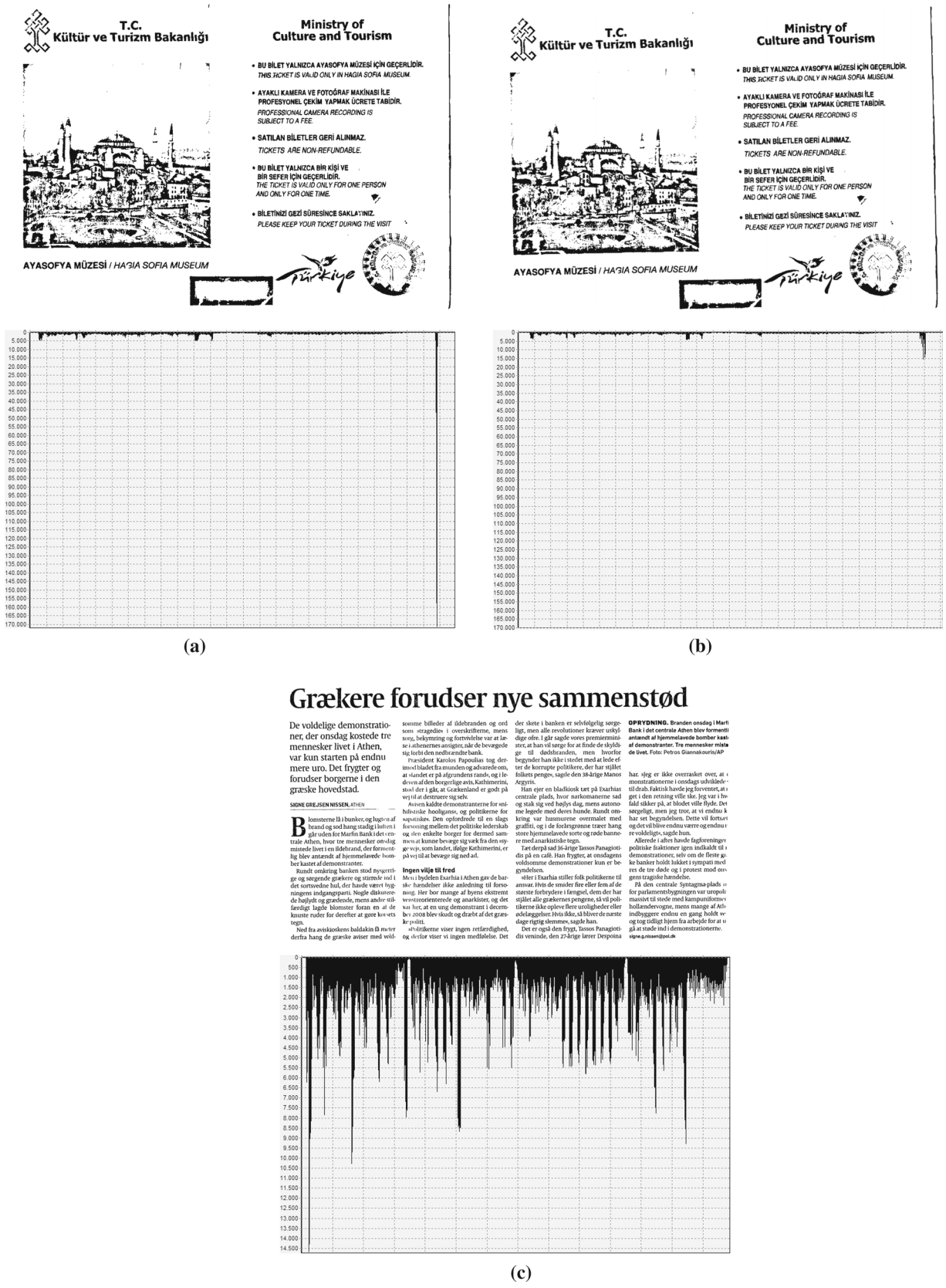


Fig. 6 The contribution of the black borders of the document image to the reinforced vertical histogram ${}_{v}H^{\theta}()$ when its in **a** zero skew angle and **b** with 1° skew. In **c**, there is an example of the reinforced vertical histogram in multicolumn documents

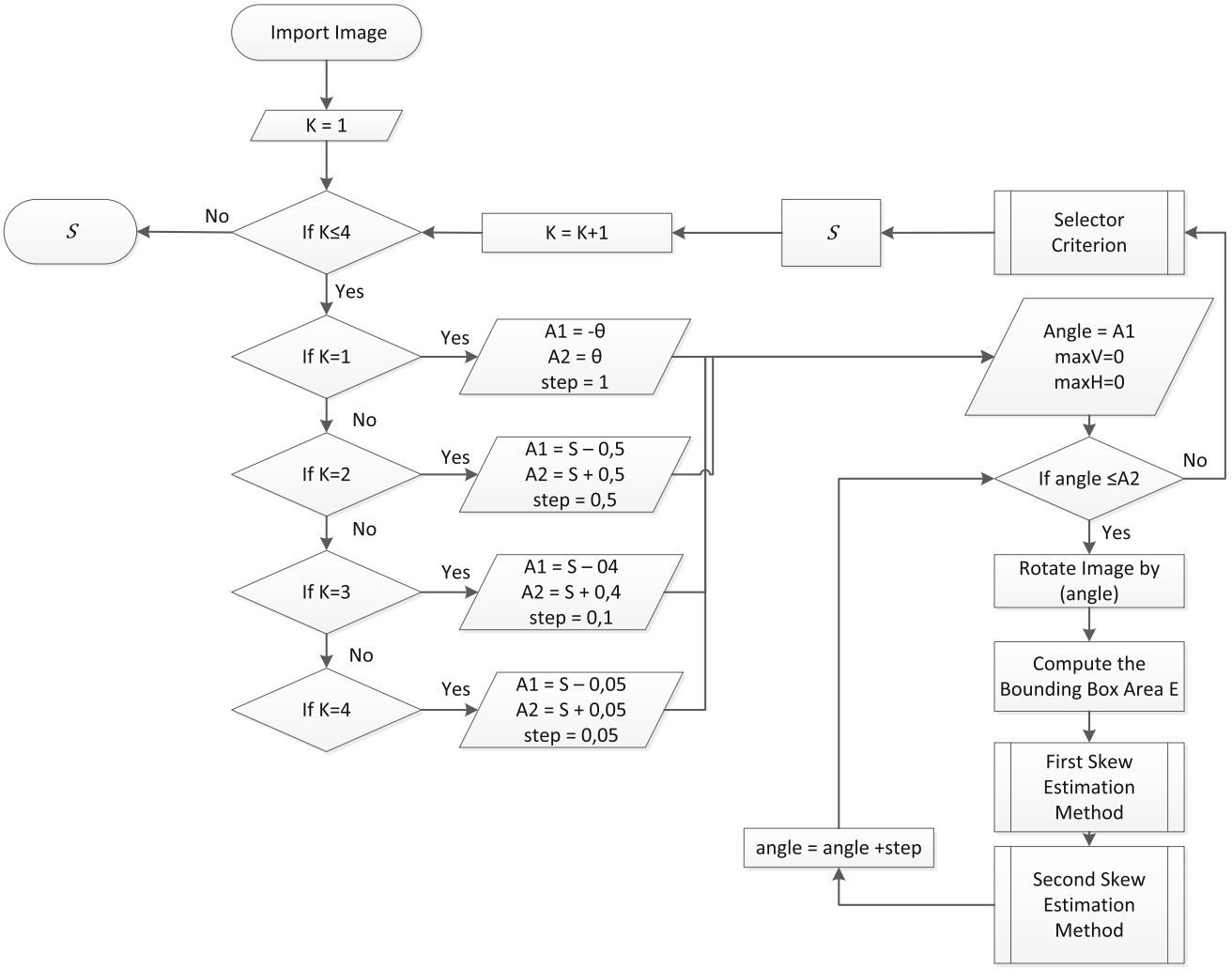


Fig. 7 Flowchart of the coarse-to-fine technique used for faster convergence

skew angle θ_{d2}. Then, the document image is rotated from $(\theta_{d2}-0.4)$ to $(\theta_{d2}+0.4)$ with step $0.1°$ and the dominant skew angle is re-adjusted in θ_{d3}, while finally, the image is rotated from $(\theta_{d3}-0.05)$ to $(\theta_{d3}+0.05)$ and so the skew angle θ_s is determined. With the use of this coarse-to-fine technique the rotations required are reduced to $(\theta_{\max}-\theta_{\min}+13)$. As a conclusion the proposed algorithm accomplishes a faster convergence and performs 9 times faster in a range between $-5°$ and $5°$ and 17, 5 times faster in a range between $-45°$ to $45°$.

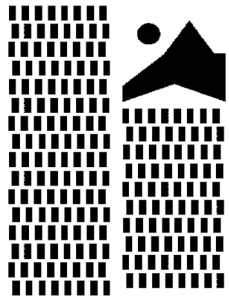
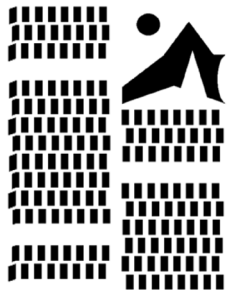
It is to be noted that the accuracy of $0.05°$ is quite satisfactory for document images scanned in a resolution of 300 dpi or greater, while for lower resolution document images the last iterative step of the CEPP c-f can be omitted without affecting the accuracy of the method.

2.6 Implementation and computational complexity

In order to optimize the computational cost of the proposed algorithm, the following implementation is suggested.

First step: Rotate the document image by the desired angle.
Second step: Detect the rectangle that contains the foreground pixels.
Third step: If the rectangle area computed is smaller than the minimum area found save the value of the area as minimum and also save the rotated angle.
Fourth step: Scan inside the detected rectangle vertically, according to the reinforced projection profiles and save the value calculated along with the previous two values. At the same time, if a foreground pixel is detected, accumulate it in the horizontal histogram.
Fifth step: If the variation among the three successive values of the vertical projections is greater than the maximum difference found, save this value as maximum as well as the rotated angle.
Sixth step: After the scanning of the rectangle is done, process the horizontal histogram, divide the values by the area of the rectangle and find the maximum difference

Table 1 Results for each iterative step of the proposed algorithm

Artificial Document Images		Vertical Profile (°)	Horizontal Profile (°)	Selection Criteria (°)	Vertical Profile (°)	Horizontal Profile (°)	Selection Criteria (°)
	**Rotated Angle (°)**	**5.5°**			**-41.65°**		
	Step 1 (1° step)	7.00	6.00	5.00	-40.00	-41.00	-41.00
	Step 2 (0.5° step)	5.50	5.50	5.50	-40.50	-41.50	-41.50
	Step 3 (0.1° step)	5.50	5.70	5.50	-41.70	-41.80	-41.70
	Step 4 (0.05° step)	5.50	5.45	5.50	-41.70	-41.65	-41.65
	**Rotated Angle (°)**	**-5.65°**			**42.35°**		
	Step 1 (1° step)	-4.00	-6.00	-6.00	25.00	42.00	42.00
	Step 2 (0.5° step)	-5.50	-5.50	-6.00	42.50	42.50	42.00
	Step 3 (0.1° step)	-5.70	-5.80	-5.60	42.40	42.50	42.40
	Step 4 (0.05° step)	-5.65	-5.65	-5.65	42.45	42.50	42.45

between three successive values for the rotated angle. If this maximum is greater than the general maximum value detected for the horizontal profiles save it as maximum and also save the rotated angle.

Seventh step: Continue with the first step until all the desired angles are examined.

Eighth step: If all angles are examined, compare the distances of the skew angle detected from the horizontal and vertical projections with the angle detected from the minimum bounding box area method and qualify the angle with the minimum distance.

In this way the proposed algorithm scans the rectangle that contains the information only once in each angle in order to compute the outcome of both vertical and horizontal profiles. The time complexity of the proposed implementation is detailed below.

During skew detection, the calculation of the maximum difference between three successive values is done for a constant number of times in a constant number of iterations. In the case of CEPP c-f method for a range of angles between $-5°$ and $5°$ and accuracy $0.05°$, the number of iterations is 23. We find the maximum values among the variations computed and since finding the maximum value involves a constant number of comparisons the time complexity of the proposed algorithm mainly depends on the time complexity of the calculation of the variation values. The variation calculation for each angle involves iteratively checking each pixel inside the rectangle area that contains the foreground pixels. Hence, the overall time complexity is $O((I_x + 1) \times I_y)$, which is linear with respect to the total number of pixels in the image. (I_x+1) stands for the number of columns of the image, I_x, plus one that stands for the histogram of the horizontal profiles which is also of the same size as the height of the document image.

3 Qualitative and quantitative evaluation of the method

3.1 Qualitative testing on the combination of vertical and horizontal profiles

As mentioned in the introduction and Sect. 2, horizontal and vertical projection profile approaches are complementary one to the other. To summarize, (1) horizontal projection profiles can accurately deal with single column printed documents with minimum noise and are accurate for a restricted range of angles and (2) vertical projection profiles are very efficient in dealing with noise, warp, multicolumn documents and tables while they perform well for a great range of angles.

The two types of projection profiles, besides their suitability for different cases, can also contribute in a cooperative way in the process of a single complex document due to the coarse-to-fine technique, since in each step of the iterative process a different method may detect the dominant skew. In this way, the enhanced projection methods are combined in a more profound way. In order to demonstrate how the combination of the enhanced projection profiles can contribute

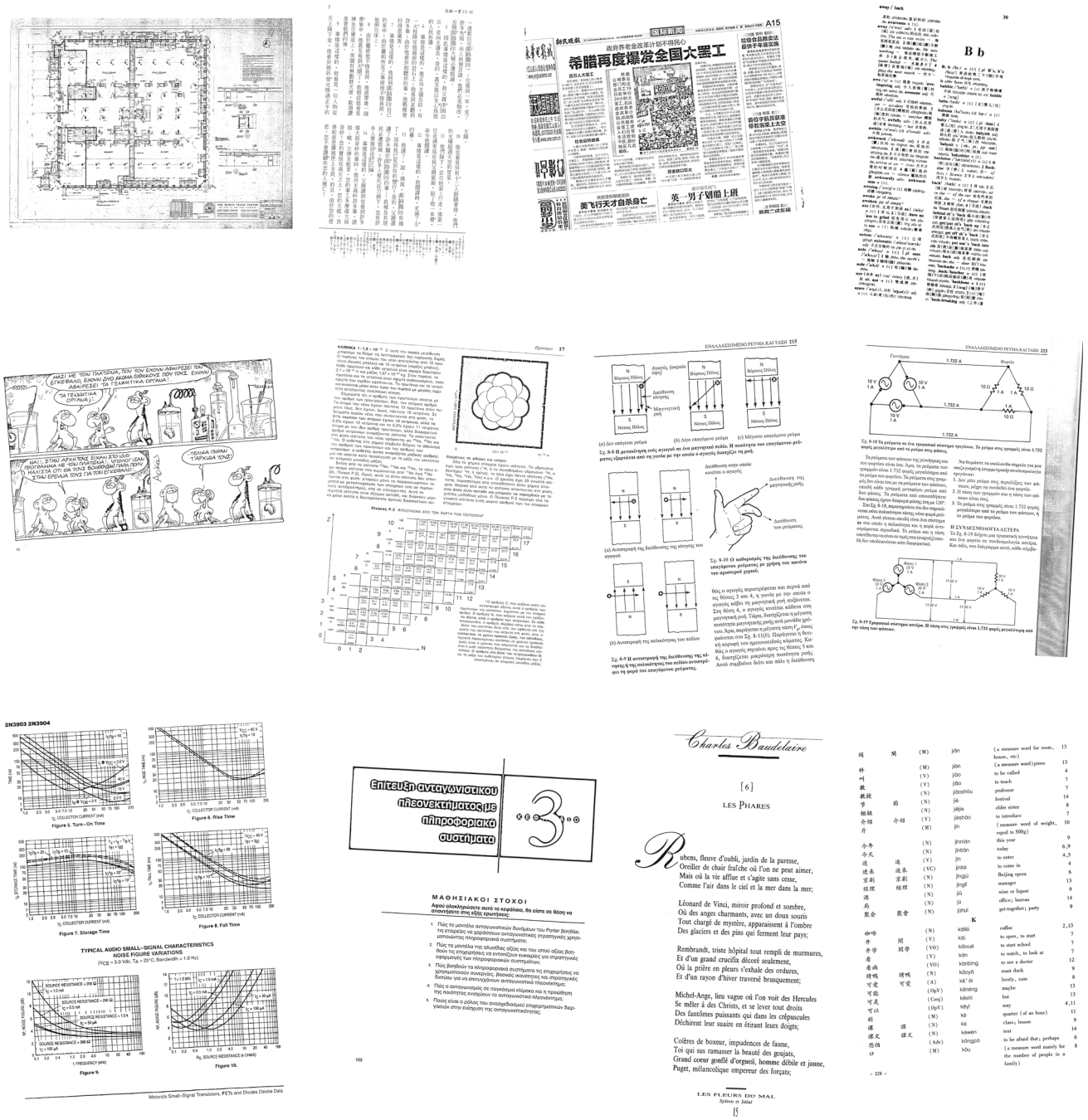


Fig. 8 Sample images of dataset A

to an accurate skew estimation, two artificial images have been created and rotated by common and wide range skew angles (see images of Table 1). These images involve the major drawback cases of both vertical and horizontal projection profiles since they represent double column documents with no vertical or horizontal alignment, while in the second case there are also vacant spaces and significant warping simulated in the left side of the image.

The results of each projection profile method used in each iterative step are demonstrated in Table 1 . The result of the projection profile method that is closer to the outcome of the minimum bounding box area algorithm is qualifying in each step and is considered as an input angle for the next step. The selected result is highlighted in order to demonstrate the swapping among the contributing projection profile methods during the convergence process. Steps 1–4 correspond to the successive iterations of CEPP c-f with respective step of rotation for the projection profile methods from 1° to 0.05°.

As can be observed in the cases demonstrated above, the optimal algorithm is altered during the converenge process

Table 2 Description of the different scripts in dataset A

Language–script of the documents	No. of documents		Percentage of documents (%)
Latin script		58	29
English	42		
Italian	5		
French	5		
Danish	4		
Turkish	2		
Greek		83	41.5
Cont. Greek	81		
Ancient Greek	2		
Bulgarian		3	1.5
Russian		4	2
Chinese		49	24.5
Japanese		3	1.5
Total		200	100

of CEPP c-f and a more profound combination of the two projection profile methods is evident. It is worth mentioning, neither the reinforced vertical nor the enhanced horizontal projection profiles alone could achieve the accuracy of the proposed method.

3.2 Quantitative testing on the accuracy of CEPP on real document images

3.2.1 Datasets

In order to test the estimation accuracy of the proposed method, we created two datasets. Dataset A is formed by contemporary document images of various types and dataset B is formed by two historical books.

Dataset A For the construction of dataset A, we scanned in 300 dpi resolution, manually deskewed, binarized [27] and rotated by a range of angles 200 document images from various types of documents, representative of the most realistic cases that an algorithm might come up against (see Fig. 8). The document images used contain figures, tables, diagrams, block diagrams, architectural plans, electrical circuits, and are obtained from newspapers, scientific journals, scientific books, literature books, poetry anthologies, course books, dictionaries, travel guides, museum guides, museum tickets, menus, comic books, official state documents and various other sources. The image documents of dataset A are written mainly in English, Chinese and Greek, while there are several documents written in Japanese, French, Bulgarian, Russian, Danish, Italian, Turkish and ancient Greek. There are representative cases of various sizes of image documents, any kind of mixed content, vertical and horizontal writing, multisized fonts and multiple different number of columns in the same document, while some of the documents suffer from degradations and distortion. A description of the languages and the layout types that are included in dataset A are detailed in Tables 2 and 3, respectively. Dataset A is publicly available [28].

As can be observed, only 30 % of the dataset is written in Latin script, 40 % is written in Greek (contemporary and ancient), 25 % is written in Chinese and the rest is written in different non-Latin script languages.

From the above table, it is derived that only 20 % of the documents contain only plain text, while most of the remaining documents have a complex layout combining figures, tables, diagrams and multi-column formats. Due to this fact

Table 3 Description of the different layouts in dataset A

	Plain text	Figures	Tables	Diagrams	Multi-column	Vertical writing	Total
Latin script							77
English	5	18	15	4	14	-	56
Italian	4	–	–	–	1	–	5
French	5	–	–	–	–	–	5
Danish	–	4	–	1	3	–	8
Turkish	–	2	–	–	1	–	3
Bulgarian	–	–	1	–	2	–	3
Russian	3	1	–	–	–	–	4
Greek							114
Cont. Greek	3	45	31	6	26	1	112
Ancient Greek	2	–	–	–	–	–	2
Chinese	16	9	5		21	8	59
Japanese	3	–	–	–	–	–	3
Total	41	79	52	11	68	9	260

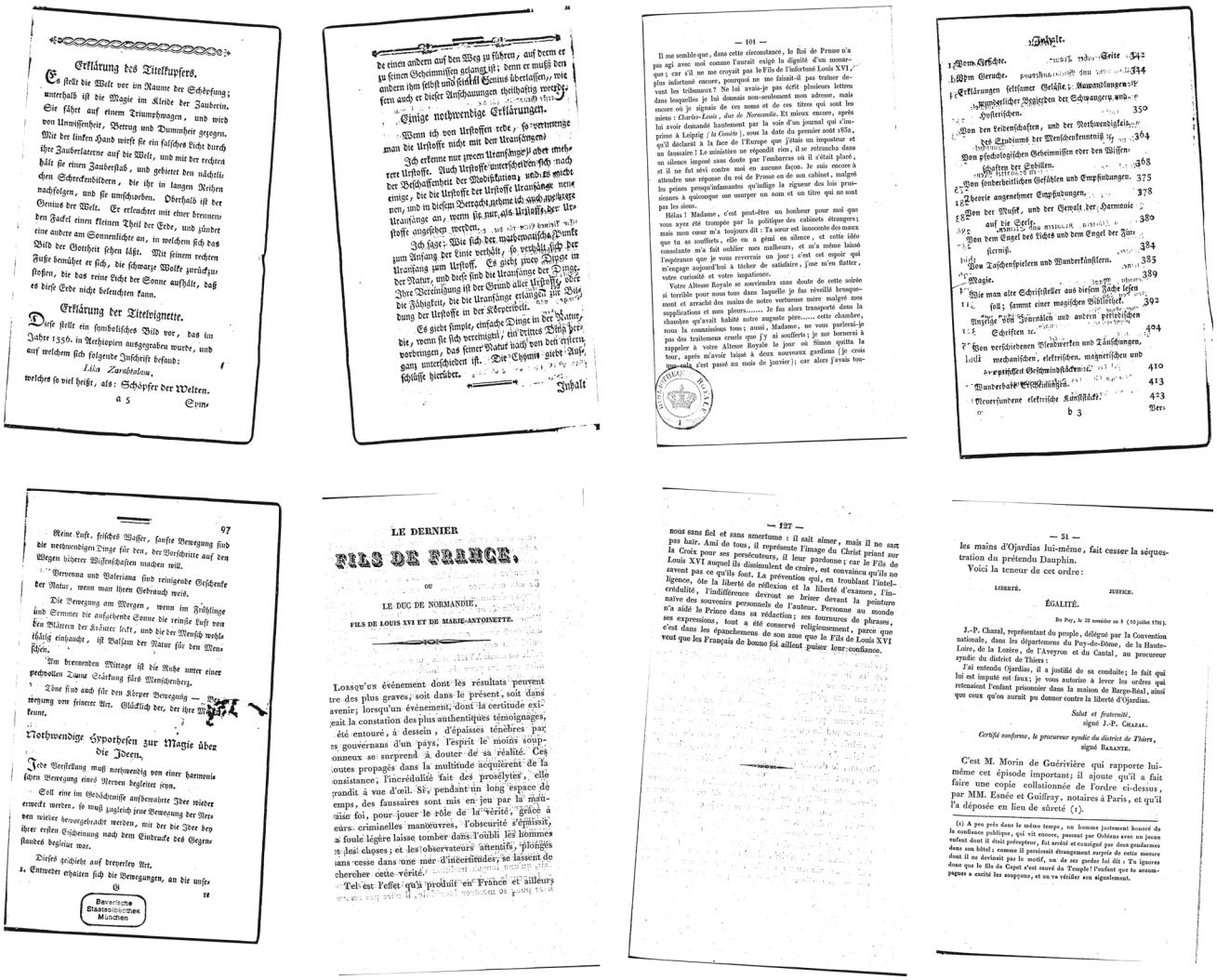


Fig. 9 Sample images of dataset B

the total number accumulated (260) exceeds the number of documents (200). 35 % of the documents have multiple columns, 45 % of the documents have either figures or diagrams while 25 % of the documents contain various styles of tables. It should also be noticed that 5 % of the dataset was written in vertical orientation.

Dataset B In order to test the accuracy performance of the proposed method in severe cases of degraded documents, we formed dataset B. Dataset B contains two entire historical books, one from Eckartshausen, which was published in 1788 and is owned by the Bavarian State Library [29] (126 document pages), and a French historical book, which was published in 1838, and is owned by the Bibliothèque nationale de France [30] (152 document pages). Samples of the dataset are demonstrated in Fig. 9. The two books consist of 278 pages and none of them was excluded, while some of the images of the books suffer from several problems such as warp, noise, broken characters and degradations (see Fig. 10). Although all pages of dataset B suffer from a level of noise and warp, in Table 4 the number of pages with extended noise (see Fig. 10a) and significant warp (see Fig. 10b) is presented.

Over 73 % of the German book and almost 62 % of the French suffer from warping while over 16 % of dataset B suffers from extended noise. The use of dataset B aims at proving the robustness of the proposed algorithm in extended degradations, noise, warp and broken characters.

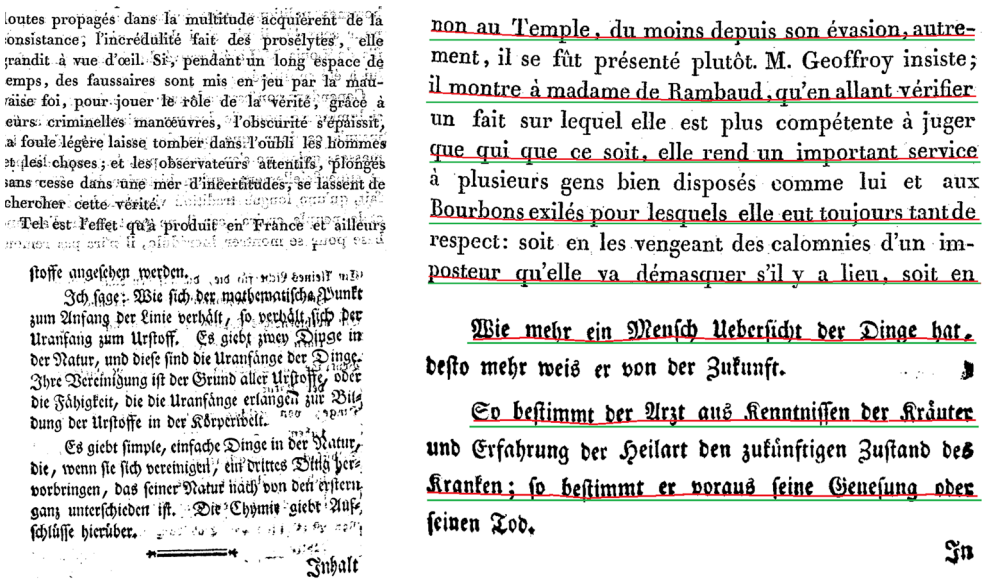
All the printed document images were scanned in 300 dpi resolution, deskewed manually, binarized with the method described in Gatos et al. [27] and were rotated by different angles, while dataset B is a superset of the one used in Papandreou and Gatos [21].

3.2.2 *Evaluation metrics*

For every document image j of the datasets used, the distance $E(j)$ between the ground-truth and the estimation of each tested algorithm was calculated. It should be noted that the estimations of each algorithm were rounded to the sec-

Table 4 Description of the degradations in dataset B

	German (126 documents)		French (152 documents)		Total (278 documents)	
Noise	4	3.18 %	42	27.63 %	46	16.54 %
Warp	92	73.01 %	94	61.84 %	186	66.90 %
Total	96	76.19 %	136	89.47 %	232	83.45 %



(a) **(b)**

Fig. 10 Image examples from dataset B with **a** extended noise and **b** significant warp

ond decimal place. In order to measure the performance of the submitted methods, the following three criteria were used: (1) the average error deviation (AED), (2) the percentage of correct estimations (CEs) and (3) the percentage of estimations that the error was less than $0.2°(E < 0.2)$. The definition of the above criteria is given below.

The AED criterion is described by:

$$\text{AED} = \frac{\sum_{j=1}^{N} E(j)}{N} \tag{16}$$

where N equals to the number of images of each dataset.

The CE criterion is determined as:

$$\text{CE} = \frac{\sum_{j=1}^{N} K(j)}{N} \quad \text{where} \quad K(j) = \begin{cases} 1 & \text{if} \quad E(j) = 0 \\ 0 & \text{otherwise} \end{cases}. \tag{17}$$

Finally, the $E < 0.2$ criterion is determined as:

$$E < 0.2 = \frac{\sum_{j=1}^{N} L(j)}{N} \quad \text{where}$$
$$L(j) = \begin{cases} 1 & \text{if} \quad E(j) < 0.2° \\ 0 & \text{otherwise.} \end{cases} \tag{18}$$

The threshold of 0.2° was chosen due to the fact that a skew angle smaller than this threshold is well accepted from the scientific community.

3.2.3 Experimental results

Four experiments were performed, two with each dataset (A and B) comparing the proposed algorithm with six other state-of-the-art methods. The first two experiments were in common skew angles (from −5° to 5°) and the other two in wide range skew angles (from −42° to 42°).

Common skew angles experiments At first all the documents of dataset A and dataset B were rotated by 11 different angles, from −5° to 5° with step 1°, and 2,200 (dataset A) and 3,058 (dataset B) image documents with known ground truth were created. In order to make a comparison of the CEPP method with current state-of-the-art skew estimation techniques, we also implemented (1) the classical projection profiles (PPs) of Postl et al. [20], (2) the Papandreou and Gatos enhanced horizontal profile (eHPP) [21], (3) the Papandreou and Gatos enhanced vertical profile (eVPP) [21], (4) the classical Hough transform (HT) of Srihari and Govindaraju [6], (5) the Yan cross-correlation (CC) [15] and (6) the Alireza et al. piece-wise painting (PPA) [19] algorithms. The range of expected skew for all projection algorithms was set from −6° to 6° while the step of rotation, and consequently its accuracy, was 0.05°. Besides CEPP, CEPP c-f algorithm was implemented and tested in which the coarse-to-fine technique was applied. The results presented in Table 5 demonstrate that our method outperforms the current state-of-the-art

Table 5 Results of state-of-the-art algorithms in common skew angles

Skew estimation technique	Dataset A			Dataset B		
	Av. error deviation (°)	Correct estimations (%)	Error <0.2° (%)	Av. error deviation (°)	Correct estimations (%)	Error <0.2° (%)
Projection profiles (PPs) [20]	0.234	44.30	63.66	0.190	46.03	59.21
Enhanced horizontal profile (eHPP) [21]	0.132	64.95	83.00	0.143	51.00	64.20
Enhanced vertical profile (eVPP) [21]	0.181	65.99	80.92	0.247	39.63	53.04
Hough transform (HT) [6]	0.138	56.49	78.04	0.330	25.82	44.83
Cross-correlation (CC) [15]	1.649	34.41	58.47	1.210	26.51	43.03
Piece-wise painting (PPA) [19]	1.282	11.61	60.76	0.307	40.61	60.20
Proposed method (CEPP)	0.055	**74.50**	**91.60**	**0.058**	**81.32**	**89.01**
Proposed method* (CEPP c-f)	**0.051**	75.49	91.03	0.083	77.30	85.21

skew estimation methods. methods. It should also be noted that the bold in the result tables denote the best performance.

In Table 5 it is demonstrated that CEPP method has an average error deviation around 0.055° regardless of the type of the document image and CEPP c-f method has a similar performance (0.051°) in contemporary documents, while its performance in historical documents is not significantly lower than CEPP's. The results in dataset A show that CEPP can handle any kind of printed image document (see Fig. 8), regardless of layout and script, with an almost zero average error deviation. Also, by observing the results from dataset B, it is derived that the proposed method can handle historical books that suffer from multiple degradations (see Fig. 9) and have common skew, with average error deviation smaller than the one observed by the human eye (0.1°), while almost 90 % of the 3,058 image documents were deskewed under the well accepted threshold of 0.2°. It can be observed that dataset B has significant warping and noise since all the methods that project horizontally or track horizontal lines have lower accuracy performance than the performance that they have in dataset A. On the other hand, there are multiple different layouts in dataset A and as a result the cross-correlation and painting methods accuracy is decreased.

Wide range of skew angles experiments In order to examine the operating range of each state-of-the-art algorithm and prove the robustness of CEPP method, all the document images of dataset A and dataset B were rotated by 15 different angles, from −42° to 42° with step 6°, and 4,170 (dataset A) and 3,000 (dataset B) image documents with known ground truth were created. The proposed method was compared with all the above mentioned state-of-the-art skew estimation techniques, where the coarse-to-fine technique was applied in all of the projection profile algorithms for faster convergence. So, we compared with: (1) the classical projection profiles (PPs c-f), (2) enhanced horizontal profile (eHPP c-f), (3) enhanced vertical profile (eVPP c-f), (4) the classical Hough transform (HT) of Srihari and Govindaraju [6], (5) the Yan cross-correlation (CC) [15] and (6) the Alireza et al. piece-wise painting (PPA) [19] algorithms. The range of expected skew for all projection algorithms was from −43° to 43° while their accuracy was 0.05°. For all these methods the average skew error deviation for every document image skew angle in degrees was calculated and the results are presented in Table 6.

As demonstrated in Table 6, the CEPP c-f method outperforms the state-of-the-art techniques, while it is robust in every range of skew since it has the same accuracy performance as in the first experiment of common skew angles. The proposed algorithm has a smaller average error deviation than the skew observed by the human eye (0.1°) in both datasets, while it can be applied to the whole spectrum of the expected skew. Furthermore, from Table 6, the operating range of each state-of-the-art algorithm implemented can be observed. Most of the state-of-the-art algorithms have an acceptable accuracy performance between 6° and −6°, with the exception of PPA that seems to operate in the range of −12° to 12° and eVPP that seems to be robust in all angles with document images of standard book layout (dataset B) and an operating range between 24° and −30° when the layout varies. It is also interesting to observe that in dataset B, due to the warping and noise, the PP fails along with the eHPP in angles beyond ±6°, while in dataset A, unlike PP, eHPP seems to be robust in angles for 30° to −30°.

3.2.4 *Qualitative comparative experiments with commercial products*

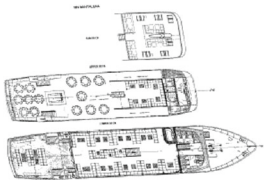
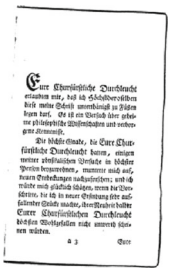
Some qualitative tests comparing the proposed algorithm with highly acknowledged commercial products in randomly chosen cases of our datasets are also performed. Through these results it is demonstrated that CEPP and CEPP c-f can make a significant contribution in the OCR level.

In Table 7 there is a comparison between the results of CEPP c-f method and scan tailor (ST) software in samples of

Table 6 Average error deviation (°) of Skew algorithms in wide range angles

Skew angle (°)	Skew estimation techniques													
	Dataset A							Dataset B						
	PPc-f	eHPPc-f	eVPPc-f	HT	CC	PPA	CEPP c-f	PPc-f	eHPPc-f	eVPPc-f	HT	CC	f PPA	CEPPc-f
−42	4.976	7.035	14.550	40.476	43.791	54.527	0.038	9.683	9.501	0.250	41.633	46.901	58.519	0.058
−36	3.353	0.547	3.983	34.464	37.750	39.025	0.045	6.906	6.882	0.249	36.205	40.953	47.313	0.088
−30	3.362	0.513	0.555	27.602	30.860	27.211	0.044	9.219	9.164	0.243	29.877	33.182	23.239	0.094
−24	1.787	0.162	0.221	21.141	23.858	17.163	0.053	4.423	4.366	0.259	22.778	26.910	3.089	0.100
−18	1.344	0.169	0.186	14.516	17.253	7.876	0.051	3.993	3.948	0.253	15.892	19.195	1.447	0.104
−12	1.442	0.154	0.183	6.637	11.548	3.270	0.045	3.251	3.160	0.292	5.664	11.456	0.648	0.087
−6	0.936	0.140	0.170	0.249	3.339	1.974	0.036	0.594	0.461	0.268	0.199	2.023	0.354	0.083
0	0.190	0.131	0.177	0.130	1.374	0.746	0.039	0.143	0.132	0.231	0.210	1.338	0.257	0.067
6	0.239	0.154	0.184	0.195	2.430	2.372	0.038	0.426	0.363	0.241	0.193	1.824	0.418	0.091
12	1.372	0.156	0.432	6.743	8.994	3.236	0.038	2.314	2.155	0.239	7.041	10.951	0.565	0.086
18	1.694	0.169	0.201	14.542	15.905	5.472	0.056	3.551	3.337	0.240	17.116	18.392	1.073	0.094
24	2.956	0.169	0.527	21.515	21.467	11.191	0.062	5.018	4.862	0.247	24.983	25.409	3.493	0.094
30	3.473	0.161	1.628	28.116	27.924	15.078	0.050	6.120	5.971	0.263	30.961	32.986	13.362	0.088
36	3.683	0.924	4.033	34.496	33.152	23.753	0.055	7.247	7.119	0.263	36.868	39.907	33.277	0.101
42	4.935	5.890	14.972	40.597	40.501	33.622	0.055	9.691	9.322	0.259	42.652	46.152	52.867	0.094
Total	2.383	1.098	2.800	19.428	21.343	16.434	**0.047**	4.838	4.716	0.253	20.818	23.839	16.004	**0.088**

Table 7 Comparative results with scan tailor skew estimation algorithm

Sample document images with skew	ST software skew estimation results	CEPP method skew estimation results	Sample document images with skew	ST software skew estimation results	CEPP method skew estimation results
					
6° skew	6.88°	6°	−5° skew	−2.81°	−4.8°
					
3° skew	1.56°	2.9°	−5° skew	−4.69°	−5.05°
					
−4° skew	−4.5°	−4°	5° skew	6°	5°

both datasets A and B, while it should be mentioned that ST software does not perform when the skew exceeds ±7.5°.

In Table 8, there are also comparative word recognition accuracy results from ABBYY Fine Reader 2011 (FR'11) software, using its own skew estimation algorithm and the proposed one. Some document images from the ones used in the previous experiments were picked and the FR'11 software was used for OCR. Then, the document images were deskewed with the proposed method and were fed in the OCR FR'11 engine, which provided improved OCR results. In that way, the contribution of CEPP c-f method is also indirectly demonstrated through recognition improvement. It should be mentioned that the FR'11 software cannot detect skew that exceeds ±20° and is not efficient with vertical writing.

As demonstrated in Table 8, there are cases in which the CEPP c-f method improves the OCR results, even after the use of layout analysis and dictionaries, proving its efficiency and contribution in the recognition accuracy.

3.2.5 Analysis of the individual components of the CEPP method

Finally, experiments to analyze the proposed method in its components were made in order to measure the influence of the components of the proposed technique. In Tables 9 and 10, the skew estimation results of (1) the reinforced vertical projection profile, (2) the enhanced horizontal projection profile and (3) the minimum bounding box area [25] algorithms for the experiments in common and wide range are demonstrated, respectively.

From Tables 9 and 10, it can be observed that the CEPP method outperforms each of its components. Furthermore, the reinforced vertical and enhanced horizontal profiles used, outperform eVPP and eHPP, respectively. The minimization of the bounding box area technique, although not proved robust for both datasets, is a very efficient criterion in order to select among various methods and candidate angle estimations.

Table 8 ABBYY Fine Reader word recognition accuracy results using different skew estimation algorithms

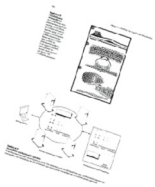
Document image	FR'11 skew estimation algorithm (%)	CEPP (%)	Document image	FR'11 skew estimation algorithm (%)	CEPP (%)
					
18° skew, 400 words	99	100	18° skew, 264 words	96.24	98.75
					
18° skew, 183 words	71.85	81.56	−18° skew, 77 words	15.58	29.87
					
−18° skew, 22 words	77.77	88.88	18° skew, 863 ideograms	59.67	92.35

Table 9 Results of the component algorithms in common skew angles

Skew estimation technique	Dataset A			Dataset B		
	Av. error deviation (°)	Correct estimations (%)	Error <0.2° (%)	Av. error deviation (°)	Correct estimations (%)	Error <0.2° (%)
Reinforced vertical profile	0.128	69.37	82.34	0.190	54.31	63.70
Enhanced horizontal profile	**0.048**	71.11	88.54	0.123	69.06	73.18
Minimization of the bounding box [23]	0.732	40.16	47.24	0.082	**83.12**	85.06
CEPP	0.055	**74.50**	**91.60**	**0.058**	81.32	**89.01**
CEPP c-f	0.051	75.49	91.03	0.083	77.30	85.21

Table 10 Results of the component algorithms in wide range skew angles

Skew estimation technique	Dataset A			Dataset B		
	Av. error deviation (°)	Correct estimations (%)	Error <0.2° (%)	Av. error deviation (°)	Correct estimations (%)	Error <0.2° (%)
Reinforced vertical profile	0.163	59.51	81.50	0.187	53.35	65.49
Enhanced horizontal profile	0.051	66.44	87.55	0.118	69.01	74.60
Minimization of the bounding box [23]	0.692	52.15	76.46	0.121	**73.81**	83.86
CEPP c-f	**0.047**	**67.74**	**89.33**	**0.088**	73.71	**84.72**

4 Conclusion

In this paper, a new technique based on combined reinforced projection profiles (CEPP) was introduced. We are motivated by the fact that the horizontal and vertical projection profile approaches are found to be complementary, so we introduce the minimum bounding box area as the appropriate criterion in order to combine novel reinforced horizontal and enhanced vertical projection profiles. Despite the fact that the proposed method falls broadly in the category of projection profile methods, CEPP deals with all the drawbacks of the projection profile methods. In more detail, it is noise and warp resistant, it gives more accurate results in any kind of printed document image, it is not restrained by any range of angles and can achieve fast convergence if the coarse-to-fine search version (CEPP c-f) is used. For these reasons, it can be efficiently applied to historical machine-printed or multicolumn documents, documents with figures and tables, while it is robust for any kind of script and a wide range of angles.

Besides the efficiency and complementarity of the vertical and horizontal projection profiles that was proved, the proposed method introduces the following innovative elements: (1) An innovative criterion which can operate as a skew estimation technique selector. The minimization of the bounding box area proved to be an effective algorithm selector which can combine two and perhaps more different skew estimation techniques. (2) Novel reinforced vertical projection profiles which take advantage of the vertical lines and strokes in a way which is dominant but not absolute were proposed. Finally, (3) the candidate skew angles that are computed by the reinforced vertical and the horizontal projection profiles are determined by the maximization of the difference between three successive values of the corresponding histograms instead of two. In this way, the calculations have greater confidence and provide improved results especially in degraded and noisy images.

The proposed skew estimation technique was extensively tested in two different datasets, one with document images of entire historical books with serious degradations and another of various document images representative of every kind of realistic case that an algorithm might come up against and it outperformed the state-of-the-art techniques. The CEPP method proved to be accurate, with an average error deviation which is not observed by human eye, it is noise and warp resistant, while it deals with documents with broken characters. The proposed method can be efficiently applied to any range of expected skew angles, it applies in vertical and horizontal oriented text, it deals with figures, tables, various font sizes and multicolumn documents, while it is faster due to a coarse-to-fine convergence technique.

To sum up, skew detection is a preprocessing task that has concerned the computer vision community for many years. The major difficulty though is given by the fact that many of the proposed state-of-the-art algorithms address very different specific problems. Also, they mainly consider documents with plain structure. This limitation appears too restrictive with respect to current documents. Furthermore, it is common for authors to declare performance computed on their own databases, generally not available, making a strict comparison very difficult [3]. To this end, we propose CEPP and CEPP c-f from a generic point of view, since skew detection is still an interesting and challenging issue and needs further improvement in order to meet the modern requirements of OCR on-the-fly, especially regarding documents with graphics, charts, figures or various font sizes. CEPP and CEPPc-f methods are moving in that direction since they are fast and accurate without any restrictions in layout, script, range of angles or context. Furthermore, this paper, with the extensive experimental results and the publicly available generic dataset [28], provides a benchmark for the scientific community and tackles the difficulty of strict comparison between different methods.

Acknowledgments The research leading to these results has received funding from the European Union's Seventh Framework Program under Grant agreement No. 215064—IMPACT as well as from the European Union's Seventh Framework Programme (FP7/2007-2013) under Grant agreement No. 600707—tranScriptorium.

References

1. Sarfraz, M., Rasheed, Z.: Skew estimation and correction of text using bounding box. In: Proceedings of Fifth International Conference on Computer Graphics, Imaging and Visualization, (CGIV '08), pp. 259–264 (2008)
2. Sadri, J., Cheriet, M.: A new approach for skew correction of documents based on particle swarm optimization. In: Proceedings of 10th International Conference on Document Analysis and Recognition, ICDAR'09, pp. 1066–1070 (2009)
3. Cattoni, R., Coianiz, T., Messelodi, S., Modena, C.M. : Geometric layout analysis techniques for document image understanding: a review. Tech. Rep. 9703–09, IRST, Trento, Italy (1998)
4. Sharif, A.E., Movahhedinia, N.: On skew estimation of Persian/Arabic printed documents. J. Appl. Sci. **8**(12), 2265–2271 (2008)
5. Jung, K., Kim, K.I., Jain, A.K.: Text information extraction in images and video: a survey. Pattern Recogn. **37**(5), 977–997S (2004)
6. Srihari, N., Govindaraju, V.: Analysis of textual images using the Hough transform. Mach. Vis. A **2**(3), 141–153 (1989)
7. Hinds, J., Fisher, L., D'Amato, D.P.: A document skew detection method using run-length encoding and the Hough transform. In: Proceedings of the 10th International Conference Pattern Recognition. IEEE CS Press, Los Alamitos, CA, pp. 464–468 (1990)
8. Wang, J., Leung, M.K.H., Hui, S.C.: Cursive word reference line detection. Pattern Recogn. **30**(3), 503–511 (1997)
9. Yu, B., Jain, A.K.: A robust and fast skew detection algorithm for generic documents. Pattern Recogn. **29**(10), 1599–1629 (1996)
10. Singh, C., Bhatia, N., Kaur, A.: Hough transform based fast skew detection and accurate skew correction methods. Pattern Recogn. **41**, 3528–3546 (2008)

11. Hashizume, A., Yeh, P.S., Rosenfeld, A.: A method of detecting the orientation of aligned components. Pattern Recogn. Lett. **4**, 125–132 (1986)
12. Gorman, L.: The document spectrum for page layout analysis. IEEE Trans. Pattern Anal. Mach. Intell. **15**(11), 1162–1173 (1993)
13. Lu, Y., Tan, C.L.: A nearest-neighbor chain based approach to skew estimation in document images. Pattern Recogn. Lett. **24**, 2315–2323 (2003)
14. Okun, O., Pietikainen, M., Sauvola, J. : Robust document skew detection based on line extraction. In: Proceedings of the 11th Scandinavian Conference on Image Analysis (SCIA'99), June 7–11, Kangerlussuaq, Greenland, pp. 457–464 (1999)
15. Yan, H.: Skew correction of document images using interline cross-correlation. CVGIP: Graph. Models Image Process. **55**(6), 538–543 (1993)
16. Gatos, B., Papamarkos, N., Chamzas, C.: Skew detection and text line position determination in digitized documents. Pattern Recogn. **30**(9), 1505–1519 (1997)
17. Chou, C.-H., Chu, S.-Y., Chang, F.: Estimation of skew angles for scanned documents based on piecewise covering by parallelograms. Pattern Recogn. **40**, 443–455 (2007)
18. Deya, P., Noushath, S.: e-PCP: A robust skew detection method for scanned document images. Pattern Recogn. **43**, 937–948 (2010)
19. Alireza, A., Umapadam, P., Nagabhushanm, P., Kimura, F.: A painting based technique for skew estimation of scanned documents. International Conference on Document Analysis and Recognition, pp. 299–303 (2011)
20. Postl, W.: Detection of linear oblique structures and skew scan in digitized documents. In: Proceedings of the 8th International Conference on Pattern Recognition, pp. 687–689 (1986)
21. Papandreou, A., Gatos, B.: A novel skew detection technique based on vertical projections. In: Proceedings of the 11th International Conference on Document Analysis and Recognition, ICDAR '11, pp. 384–388 (2011)
22. Baird, H.S.: The skew angle of printed documents. In: Proceedings of the SPSE 40th Symposium Hybrid Imaging Systems, Rochester, NY, pp. 739–743M (1987)
23. Ciardiello, G., Scafuro, G., Degrandi, M.T., Spada, M.R., Roccotelli, M.P.: An experimental system for office document handling and text recognition. In: Proceedings of the 9th International Conference on Pattern Recognition, pp. 739–743 (1988)
24. Ishitani, Y.: Document skew detection based on local region complexity. In: Proceedings of the 2nd International Conference on Document Analysis and Recognition, Tsukuba, Japan, pp. 49–52 (1993)
25. Safabakhsh, R.: Document skew detection using minimum-area bounding rectangle. In: Proceedings of the International Conference on Information Technology: Coding and Computing ITCC 00, pp. 253–258 (2000)
26. Li, S., Qinghua, S., Jun, S.: Skew detection using wavelet decomposition and projection profile analysis. Pattern Recogn. Lett. **28**(5), 555–562 (2007)
27. Gatos, B., Pratikakis, I., Perantonis, S.J.: Adaptive degraded document image binarization. Pattern Recogn. **39**, 317–327 (2006)
28. https://www.iit.demokritos.gr/~alexpap/dataset_A.rar
29. von Eckartshausen, C.: Aufschlüsse zur Magie aus geprüften Erfahrungen über verborgene philosophische Wissenschaften und verdeckte Geheimnisse der Natur, Bavarian State Library (1778)
30. Le Dernier fils de France, ou le Duc de Normandie, fils de Louis XVI et de Marie-Antoinette, Bibliothèque nationale de France (1838)