An Efficient Document Skew Detection Method Using Probability Model and Q Test

Kai Huang 1, Zixuan Chen 1, Min Yu 2,* , Xiaolang Yan 1 and Aiguo Yin 3

1 Institute of VLSI Design, Zhejiang University, Hangzhou 310027, China; huangk@vlsi.zju.edu.cn (K.H.); chenzixuan@zju.edu.cn (Z.C.); xlyan@zju.edu.cn (X.Y.)
2 College of Information Science & Electronic Engineering, Zhejiang University, Hangzhou 310027, China
3 Zhuhai Pantum Printing Technology Co., Ltd., Zhuhai 519000, China; yin@pantum.com

Received: 10 December 2019; Accepted: 27 December 2019; Published: 30 December 2019

Abstract: Document skew detection is one of the key technologies in most of the document analysis systems. However, existing skew detection methods either have low accuracy or require a large amount of computation. To achieve a good tradeoff between efficiency and performance, we propose a novel skew detection approach based on bounding boxes, probability model, and Dixon’s Q test. Firstly, bounding boxes are used to pick out the eligible connected components (ECC). Then, we calculate the slopes of the skew document with the probability model. Finally, we find the optimal result with Dixon’s Q test and projection profile method. Moreover, the proposed method can detect the skew angle in a wider range. The experimental results show that our skew detection algorithm can achieve high speed and accuracy simultaneously compared with existing algorithms.

Keywords: document image; skew detection; bounding box; probability model; Dixon’s Q test; projection profile method

1. Introduction

With the rapid development of technology and the strengthening of environmental awareness, the use of electronic documents is more and more extensive. Compared with traditional paper documents, electronic documents have advantages of small footprint, easy saving, easy modification, and easy transmission. Therefore, we often scan some paper documents to convert them into electronic documents [1–4]. During the scanning process of the documents, the documents may be skew due to human factors. Skew document images cause inconvenience to subsequent image processing, and may even lead to wrong results, so skew detection and correction are important steps in image preprocessing. Several methods have been proposed for skew detection and correction of document. The most popular methods are projection profile (PP), Hough transform (HT), and nearest neighbor (NN) methods. Besides, some other methods based on the textual characteristics of document have also been proposed [5–8]. As far as we know, processing speed is important when high volumes of scanned documents have to be processed using optical character recognition, especially in some systems with high real-time requirements. So, in addition to the accuracy of angle detection, the speed of angle detection also needs to be concerned. However, some methods require extensive computation, such as Hough transform-based methods. Some methods such as PP-based methods can only deal with small skew angles because they have high computational cost of exhaustive search. And some other methods sacrifice the accuracy in order to increase the calculation speed, such as the axes-parallel bounding box method in [9].

In this paper, we propose a novel method for skew angle detection. On one hand, our method uses bounding boxes and probability model to calculate the slopes of the document, which has the
advantage of low computation cost compared with other methods. On the other hand, we combine the Dixon’s Q test and PP method to find a more accurate skew angle to improve the performance of the algorithm.

The structure of this paper is as follows. In Section 2, we present a survey of previous works dedicated to the skew detection subject. Section 3 describes our proposed skew angle detection method in detail. The experimental results and comparative study are presented in Section 4. Finally, we conclude the paper and illustrate our future work in Section 5.

2. Related Work

Over the last few years, various methods have been developed by researchers for skew angle detection, which can be mainly divided into three types: the projection profile analysis method, the Hough transform method, and the nearest neighbor method [10]. In addition to the above three commonly used methods, some novel methods can also be found which are based on the features of the documents.

Postl [11] first proposed the projection profile method. In his method, histograms of the number of the black pixels along horizontal line through the document for a range of angles are calculated. In 1995, Bloomberg [12] proposed an efficient, accurate, and robust method for measuring document image skew and orientation. He used the sample image instead of the whole document image to calculate the skew angle, which greatly increases the speed of the image skew angle calculation. In 2014, Jain [13] presented two algorithms, which were vertical projection profile analysis and horizontal projection profile analysis, and his experiment results show that the horizontal projection profile had better performance than the vertical projection profile. Moreover, the horizontal profile technique could be used for skew correction with noisy images. Although projection profile methods are easy to implement and relatively intuitive, they have high computational complexity and a small range of angle estimation limitation. Additionally, these methods are very sensitive to diagrams, graphs, or noise.

The Hough transform is a well-known skew detection method detecting lines and curves in digital images. However, in this method, every black pixel of document image needs to be transferred from Cartesian space to Hough space, which makes Hough transform method computationally expensive. The Hough transform method was initially proposed by Hough [14] in 1962. In order to improve the calculation speed, in 1990, Xu proposed randomized Hough transform (RHT) method [15]. He randomly picked n pixels and mapped them into one point in the parameter space. Boukharouba [16] used randomized Hough transform method to detect the lower baseline of the text lines of an Arabic document, but this method requires a line of text with a distinct bottom baseline. In 2000, Matas [17] presented the progressive probabilistic Hough transform (PPHT). They minimized the amount of computation needed to detect lines by exploiting the difference in the fraction of votes needed to reliably detect lines with different numbers of supporting points. Boudraa [18] introduced morphological skeleton method to remove the redundant pixels and only retain the central curves of the input document image, which achieved high accuracy but reduced the speed of angle detection.

Nearest neighbor (NN) method is based on finding the connected components of a document. Hashizume et al. [19] first proposed NN method. They computed all the orientations of connected components (CC), then calculated a histogram of all the above orientations, in which the peak value indicates the skew angle of the document. In 2003, Lu and Tan [20] improved the NN method. They proposed a nearest-neighbor chain (NCC)-based approach. In the NCC method, they extracted all the eligible NCCs, calculated the slope of each NCC, and then used the slope of the most number of times to represent the slope of the true document skew. The NCC method achieved an improved accuracy for estimating skew angle and applied to a variety of different languages. Fabrizio [21] proposed a simple and accurate method, which first used a KNN clustering to preprocess the input document image and then estimated the skew angle in the frequency domain.
In addition to the three commonly used methods mentioned above, some other innovative methods have also been found. For instance, [9] presented a novel approach in skew detection of a document by minimizing the area of the axis-parallel bounding box. Chou et al. [22] proposed a fast and robust skew detection method by using piecewise covering of objects, such as text lines, figures, or tables. Finally, the overview of existing skew detection methods is summarized in Table 1.

| Category          | Advantages                          | Disadvantages                                           |
|-------------------|-------------------------------------|---------------------------------------------------------|
| Projection profile | Simplicity, efficiency, and straightforward | Sensitive to noise and tables, very slow   |
| Hough transform   | High accuracy                        | High computation cost, large memory, and limitation in handling curved text-lines |
| Nearest neighbor  | Robust, fast, and not limited to range of skew angle | Design for special script, very sensitive to noise |

In statistics, Dixon’s Q test [23] (also named Q test) is a method to find outliers in very small datasets, which are usually defined as somewhere between 3 and 10 items. It is commonly used in chemistry, where datasets sometimes include one suspect observation that is much lower or higher than the average. Researchers often use Dixon’s Q test to check whether the suspect observation is an outlier. Dixon’s Q test usually includes the following steps: (1) sort all observations from small to large; (2) calculate the maximum difference \( \text{Range} \) between all observations; (3) find the absolute difference \( \text{Gap} \) between the suspect observation and its closest observation; (4) calculate the ratio \( Q \) of \( \text{Gap} \) to \( \text{Range} \); (5) compare the \( Q \) value with the \( Q_{\text{crit}} \) value in Table 2. If \( Q \) is larger than \( Q_{\text{crit}} \), the suspect observation is an outlier and needs to be removed. Table 2 shows critical values \( Q_{\text{crit}} \) at different numbers of observations (NO) and different confidence levels (CL).

3. Proposed Method

In this part, we propose a method based primarily on connected components (CC). The CC mainly exists in the binary image, which is a small area unit composed of same intensity pixels, where the pixel value is generally 1. The types of CC can be divided into 4-adjacency connectivity and 8-adjacency connectivity according to different adjacency modes [24]. Our method uses 8-adjacency connectivity method to search the document image. Hashizume [19] first proposed a nearest neighbor-based method. In their method, the connected components are detected first. They use the direction vector of connected components to indicate the skew of the document. Some previous research works also use the connected component as the basic processing unit to calculate the skew angle of the document [6,25–27].

Different from other CC-based methods, our proposed method randomly selects two different CCs rather than the nearest CC to calculate the document skew slope. In order to improve the calculation speed, we only consider some eligible connected components (ECC) instead of all the CCs of the document. Bounding boxes are used to pick out the ECCs from the document and calculate the locations of the corresponding ECCs [28]. After detecting several skew slope values, Dixon’s Q test or PP method is performed to find the optimal skew angle. Then, the skew document is rotated with the nearest neighbor interpolation method. The approach of skew document normalization in this paper is generally divided into three steps: (1) Detect and obtain the skew slopes of the document. (2) Select the most accurate slope value to calculate the final skew angle. (3) Rotate the document image to achieve the proposed correction. From the above three steps we can see that the skew slope detection and angle calculation are the core steps of the whole technology. And the detailed flowchart of our proposed method is shown in Figure 1.
3.1. Document Image Preprocessing

In this procedure, we mainly preprocess the input document, which including three steps: (1) document image binarization; (2) noise processing; (3) image morphology operation. These processes are necessary for skew angle detection. On one hand, they can improve the accuracy of the results. On the other hand, they speed up the calculation by removing some redundant pixels.

3.1.1. Image Binarization

Since the input document is usually multicolor or gray, we first perform binarization on the document. The existing document image binarization methods are classified in two categories, namely global and local [29]. Global thresholding methods use a single threshold for the whole document, while local methods find a local threshold based on window operation. Considering the obvious difference between the text-lines and the background in the document, we apply Otsu’s method [30] to the input gray image. In comparison with other local thresholding methods, Otsu’s method also has the advantage of simple computation. After image binarization with Otsu’s method, a binary image is output where the relevant information (text-lines) is set to 0 (black intensity), then we need to invert the binary image, because morphological operation deals with white intensity pixels.

![Flowchart of proposed method](image-url)

**Figure 1.** Flowchart of proposed method. ECC: eligible connected components.

| i = 0 | i = i+1 | i <= 8*N |
|-------|---------|----------|
| NO    | YES     | NO       |

**Note:**

- N: number of ECCs
- TR: threshold of the range
- ECC: eligible connected component
3.1.2. Noise Removal

When performing a connected component search, some punctuation (comma and dot) and isolated noises can be identified as connected components, which seriously affect the result whether the current CC is an ECC or not. Therefore, we need to remove the noise components from the document. Since the input document image is a binary image, we do not use common image denoising methods [31,32], such as median filtering [33], mean filtering [34], or fast fourier transform (FFT) methods [35], and so on.

In this stage, a CC with less than 25 pixels is considered a punctuation mark or noise and will be removed. We consider the threshold value of 25 after conducting a number of experiments on the scanned document. Experiment shows that some non-text content such as dots, common, isolate pixels, and noise components are removed.

3.1.3. Morphological Operation

In our method, aimed to get accurate skew angle, we hope that every Chinese character or English letter only has one CC. But under actual circumstances, as shown in Figure 2a, a Chinese character is recognized as upper and lower two connected components which are surrounded by bounding boxes. As can also be seen from Figure 2b, several English letters connect to each other so that they are considered to be one CC.

In order to solve the above problems, we first apply erode method to the document image by considering a 1 × 2 line-shaped structuring element, this erode operation can separate connected letters, as shown in Figure 2d. Then, we use dilation operation to connect the Chinese character’s upper part and lower part to form a connected region by considering a circular structuring element with radius 1, and the result is shown in Figure 2c.

![Figure 2](image_url)

**Figure 2.** The morphological operation: (a) the original Chinese character with two CCs; (b) the original English word; (c) the Chinese character after dilation operation; (d) the English word with erode operation.

3.2. Eligible Connected Components Selection

After document image preprocessing, we find all the CCs in the document. In general, documents not only contain text-lines, but also contain tables, images, symbols, and so on. Different from the text-line CCs, these table or image CCs, named non-text CCs, are usually very large or extremely small and randomly distributed, which seriously affects the result, meaning they need to be removed. Therefore, we need to filter these non-text CCs by analyzing their bounding boxes’ size.

In our method, we achieve the locations of CCs using the centers of their bounding boxes. In an English word, it is difficult to ensure that the centers of all the letters are on a line, so some researchers used least squares method to estimate the text-line orientation, which is a little bit more complicated [10]. In order to reduce the computation, in our algorithm, an English letter whose center of its corresponding bounding box is offset from the text-line needs to be remove. As shown in Figure 3a, the 26 English
letters can be classified into three types: A type such as the letter “i”, which is on the upper two lines; B type such as letter “e” is in the middle line; C type such as letter “q” is located in the lower two lines. Figure 3b shows the bounding boxes of all the letters in the Figure 3a. In our method, we need to save the bounding boxes of B type letters while removing the bounding boxes of A and C type letters. The result is as in Figure 3c.

In summary, after searching for CCs in a document, we pick out the ECCs by following two steps: (1) use adaptive area threshold to remove non-text CCs, and (2) remove A and C type CCs by analyzing their bounding boxes’ aspect ratio.

3.2.1. Non-text CCs Removal

In this procedure, we mainly analyze the size of bounding boxes of non-text CCs instead of non-text CCs themselves, because the bounding boxes can more accurately represent the spatial position of the CCs. Then, we filter these non-text CCs by analyzing their bounding boxes’ size. Let us denote a set of bounding boxes in the binarized document as \( \{B_i\} \). Then, we remove the CCs that violate

\[
B_{\text{max}} > B_i > B_{\text{min}},
\]

where \( B_{\text{max}} \) and \( B_{\text{min}} \) are thresholds for the size-based filtering. Both of them are document-based adaptive parameters. They can be defined by using following formulas:

\[
B_{\text{max}} = (W_{\text{ave}} \times H_{\text{ave}}) \times 4,
\]

\[
B_{\text{min}} = (W_{\text{ave}} \times H_{\text{ave}}) \times 0.25,
\]

where \( W_{\text{ave}} \) is the mean width value of all the bounding boxes, and \( H_{\text{ave}} \) is the mean height value of all the bounding boxes. Through the above processing, most of the non-text CCs will be filtered out and not involved in subsequent processes.

3.2.2. A and C Type CCs Filtering

As we can see from Figure 3a,b, the bounding boxes of A type letters and C type letters are generally higher than the average bounding box height of letters in an English word. And the bounding boxes’ centers of them are also away from the text-line. Therefore, we impose constraints on the bounding boxes’ aspect ratios to remove A and C type CCs. We denote the widths and heights of bounding boxes as \( w(B_i) \) and \( h(B_i) \), we remove CCs that violate

\[
a < w(B_i) / h(B_i) < \beta,
\]

where \( a \) and \( \beta \) are two constants 0.6 and 2, respectively. These two parameters were determined by our extensive experiments, and Figure 3c shows the ECCs after removing A and C type CCs by using bounding boxes.

With the help of the above two constrains, our approach can process complex documents with pictures and charts, which shows the robustness of our algorithm.
3.3. Skew Slopes Calculation

After image preprocessing and non-text CCs filtering, we get the ECCs to calculate the document skew slope values. Figure 4b is a part of Figure 4a which is surrounded by a red rectangular frame. From Figure 4b we can see that the ECCs is picked out, and these ECCs’ centers are marked in Figure 4c. The position coordinates of centers in Figure 4c are used to indicate the positions of ECCs of Figure 4b in our method.

In this procedure, we refer to the classical probability model to calculate the slope of the skewed document. From Figure 4c, we can find that these centers are regularly distributed. They form several parallel lines because they are extracted from parallel text-lines. Through mathematical analysis, all the points of Figure 4c have an equal probability to be taken. If we randomly select two different centers from Figure 4c to form a line segment, this line segment has the highest probability of being approximately parallel to the text-line. The classical probability model indicates that when the number of samples is large enough, we can find that there are the most line segments parallel to the text-line.

Therefore, we first randomly select two ECCs of a document to calculate its slope, and put the resulting slope into a histogram. Then, we repeat the centers selection and slope calculation until the number of selections reaches a predefined threshold. Generally, the peak of histogram gives the slope of skew document. Considering that some documents with a small number of ECCs need repeated centers selection to increase the accuracy of the results, but too many times of repetition will undoubtedly increase the amount of computation, we chose a value of eight times the number of ECCs as the adaptive threshold value with the help of our experiments.

In order to improve the accuracy of our method, we select several slope values with the most accumulations instead of the one corresponding to the peak in histogram. On the one hand, in reality, there are several slope values that approximate the text-line slope due to the position errors, so we need to choose the best slope value to represent the skew of the document. On the other hand, some complex documents have multidirectional text-lines, and the slope value corresponding to peak of histogram may vary greatly at different times.
3.4. Skew Angle Estimation

In this section, we combine the Q test and PP method to find the optimal skew angle value from the detected slope values of previous section (Section 3.3). Here, we assume that we detect T slope values in Section 3.3, and these slope values are $S_1, S_2, S_3 \cdots S_T$. We can find the optimal slope value for the above slope values using the Algorithm 1:

### Algorithm 1. A procedure to find the optimal skew angle value.

**Input:** Slope values $S_1, S_2, S_3 \cdots S_T$ after slope calculation (Section 3.3), original image $img$

**Output:** The optimal angle $\theta$

1. array $SS \leftarrow \text{sort}(S_1, S_2, S_3 \cdots S_T)$;
2. find range $R$ in array $SS ([SS(T) - SS(1)])$;
3. calculate each angle $\theta_i(^\circ)$ corresponding to each slope of $SS$;
4. if ($R \leq TR$)
5. use $Q$ test to find the optimal slope value and calculate the skew angle $\theta$
6. else
7. for $i = 1 : T$
8. $img \leftarrow \text{imrotate}(img, \theta_i)$;
9. project image $img$ horizontally or vertically and count the number ($N_i$) of blank rows or columns;
10. end
11. find the maximum $N_{max}$ value of $N_1 \sim N_T$, and analysis it to calculate the optimal angle $\theta$
12. end

Figure 4. Document skew detection and correction: (a) original document image; (b) the document image with bounding boxes; (c) the positions of ECCs; (d) the corrected document image.
In this procedure, we mainly use Q test to find a more accurate result. But in some cases, the document has more than one text-line direction, and using classical probability model may result in very different slope values. Q test can only find outliers between some approximations, and if we continue to use Q test method, it may lead to serious errors. Therefore, if detected slope values are very different, we choose to use PP method to pick out the best slope value. Since we only need to rotate the original image T times based on the detected slope values, there is no significant increase in calculation complexity. The detailed steps will be described in following part.

3.4.1. Skew Angle Calculation with Dixon’s Q Test

In statistics, Dixon’s Q test (also named Q test) is used to test if one single value is an outlier in a sample size between 3 and 10. In this procedure, we use Q test to check whether each slope value is an outlier by the following steps (assume there five detected slope values):

Step 1: Arrange five slope values in ascending order (smallest to largest). Assume

\[ S_2 < S_1 < S_4 < S_3 < S_5, \]  

(5)

Step 2: Find the difference between the maximum value and the minimum value Range.

\[ \text{Range} = S_5 - S_2, \]  

(6)

Step 3: Find the absolute difference \textit{Gap} between the suspected outlier (assume \( S_5 \)) and its closest number.

\[ \text{Gap} = |S_5 - S_3|, \]  

(7)

Step 4: Calculate the experiment Q value (\( Q_{\text{exp}} \)).

\[ Q_{\text{exp}} = \frac{\text{Gap}}{\text{Range}} = \frac{|S_5 - S_3|}{S_5 - S_2}, \]  

(8)

Step 5: Find the Q critical value \( Q_{\text{crit}} \) in the Q table, where Q table is a reference value corresponding to the sample size and confidence level. As shown in Table 2, with four observations and at 90% confidence, the Q critical value is 0.765.

Step 6: Compare \( Q_{\text{exp}} \) with \( Q_{\text{crit}} \). If \( Q_{\text{exp}} \) is larger than \( Q_{\text{crit}} \), this observation is considered to be an outlier, and mark it.

Step 7: Repeat step 3–step 6 to check whether each slope value is an outlier, then delete the marked outliers in step 6, and choose the mean of the remaining slope values to be the optimal slope value \( S_{\text{optm}} \).

According to above slope value correction, we finally get the optimal slope value \( S_{\text{optm}} \). The skew angle can be calculated using the following formula:

\[ \theta = \arctan(S_{\text{optm}}) \times \frac{180^\circ}{\pi}, \]  

(9)

where, \( S_{\text{optm}} \) is the slope value of document. Here, the right part of above equation is appended to have degree representation (\(^\circ\)), instead of the radian representation (\( \pi \)). The \( \theta \) is the document skew angle.

3.4.2. Skew Angle Detection with PP Method

As described in Algorithm 1 if the range \( R \) of the detected slope values is larger than \( TR \), which means the document may have more than one text-line direction, we will use the PP method to find the most accurate skew angle. This method consists of the following steps:
Step 1: Use Equation (9) to calculation the angle $\theta_i$ for each detected slope value, $i = 1, 2, 3 \cdots T$.
Step 2: Rotate the document $\theta_i$ counterclockwise if $|\theta_i| \leq 45^\circ$, else rotate $90^\circ - \theta_i$ clockwise.
Step 3: Project the document horizontally if $|\theta_i| \leq 45^\circ$, else project it vertically, and count the number of blank rows or columns in the document.
Step 4: Select another angle that has not been rotated, repeat step 2 and step 3 until all the above $T$ angles have been processed.
Step 5: Find the angle $\theta$ corresponding to the maximum number of blank rows or columns, which is the skew angle of the document.

Additionally, if the documents have different text orientations (i.e., vertical and horizontal), the skew angle $\theta$ detected by above two methods requires further analysis using the following formula:

$$\Theta = \begin{cases} 
\theta + 90^\circ & \text{if } \theta \leq -45^\circ \\
\theta - 90^\circ & \text{if } \theta \geq +45^\circ \\
\theta & \text{Otherwise}
\end{cases}$$

This extension is for some special cases, for instance, some traditional Chinese or Japanese documents are written vertically from top to bottom and starting at the right side of them.

| Q$_{crit}$ | NO 1 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|-----------|-----|---|---|---|---|---|---|---|----|
| (CL: 90%) | 0.941 | 0.765 | 0.642 | 0.560 | 0.507 | 0.468 | 0.437 | 0.412 |
| (CL: 95%) | 0.970 | 0.829 | 0.710 | 0.625 | 0.568 | 0.526 | 0.493 | 0.466 |
| (CL: 99%) | 0.994 | 0.926 | 0.821 | 0.740 | 0.680 | 0.634 | 0.598 | 0.568 |

1 NO indicates the number of observations (also the sample size). 2 CL indicates the confidence level.

3.5. Skew Document Correction

The final step is to rectify this inclination, the original document image is rotated by the calculated angle $\Theta$. It is advised to use interpolation in image rotation. The interpolation method used in this paper is nearest neighbor interpolation [36], because it is the simplest and least time-consuming compared with other interpolation methods. In addition, the phenomenon of contour jaggies caused by the nearest neighbor interpolation has little effect on the reading of the document. Figure 4d illustrates an example of skew detection and image rotation using our proposed method.

4. Experiments and Evaluation

In order to demonstrate the performance and accuracy of our proposed method, we conducted simulation experiments on publicly available datasets. We tested our experiments on the well-known dataset from DISEC’13 [37] contest, which includes an experimental dataset and a benchmarking dataset with a total of 1750 images, ranging from $-15^\circ$ to $15^\circ$. In our experiment, we selected 900 document images of the benchmarking dataset to tune our proposed algorithm, while we selected 200 of them to compare the performance and efficiency with other methods. And, we also used four common evaluation criteria (as shown in Section 4.1) to evaluate the performance of our algorithm. These document images were performed on a PC equipped with a microprocessor Core i7 1.8-GHz and 8 GB RAM under Windows 10. And the application was developed with MATLAB R2018b.
4.1. Evaluation Criteria

In Boudraa’s method [38], to validate the inclination angle estimation performance, they used some indicators to evaluate and compare their experiment results. In our experiment, we used the same indicators. The indicators mainly include average error deviation (AED), average error deviation of the top 80% (ATOP80), variance of error estimation (VEE), and variance of error estimation of the top 80% (VTOP80), and their concepts are as follows:

- **Average error deviation (AED):** This norm calculates the mean error deviation, as in the next equation:
  \[
  AED = \frac{\sum_{j=1}^{N} E(j)}{N},
  \]
  where \( E(j) \) symbolizes the unsigned difference rounded to the third 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 in the following equation:
  \[
  ATOP80 = \frac{\sum_{j=1}^{0.8N} E_{80}(j)}{0.8 \times N},
  \]
  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
  \[
  VEE = \frac{\sum_{j=1}^{N} (E(j) - AED)^2}{N},
  \]

- **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
  \[
  VTOP80 = \frac{\sum_{j=1}^{0.8N} (E_{80}(j) - ATOP80)^2}{0.8 \times N},
  \]

4.2. Experiment Results

As we know, various types of documents were used in DISEC’13 [37], including newspapers, literature, scientific and course books, dictionaries, travel and museum guides, and even comics. The document images of the dataset are written mainly in English, Chinese, and Greek languages, while there are several documents written in Japanese, Russian, Danish, Italian, and Turkish. These documents cover various types of data and some particular challenges, for instance, vertical and horizontal text-lines, figures, charts, block diagrams, architectural plans, and equations. Besides, every document image with the same content was arbitrarily pivoted in ten distinct angles, sorted out from \(-15^\circ\) to \(+15^\circ\).

Figure 5 shows some complex documents of different type, including cartoons, newspapers, and so on. These skew documents can be accurately corrected using our method, and the average error of the detected skew angle is only 0.085°. Figure 5d has the maximum skew angle detection error of 0.16°, while Figure 5h has the minimum skew angle detection error of 0.01°, so it can be seen that our method has high accuracy even for complex documents like cartoons.
Figure 5. The rectification results of different type of document images with distinct angles: (a,d,g,j) are original images with skew angles of $-12.78^\circ$, $8.24^\circ$, $7.30^\circ$, and $-9.52^\circ$ respectively; (b,e,h,k) are the parts of (a,d,g,j), respectively, which are surrounded by red rectangles; (c,f,i,l) are the corresponding correction result of (b,e,h,k) using our method, and the detected skew angles are $-12.84^\circ$, $8.08^\circ$, $7.29^\circ$, and $-9.41^\circ$, respectively.
As described in Section 3.4, we use several detected slope values to find the optimal result. In this part, 900 different documents from the benchmarking dataset were taken into experiments to find the proper number of slopes and range threshold. We tested the number of slopes from three to seven, and tested the range threshold from 0.005 to 0.02 with steps of 0.005. In our method, we selected to use 90% confidence level. Table 3 shows the ATOP80 values at different slope numbers and different range thresholds. Table 4 shows the average runtimes of different slope numbers and different range thresholds. Considering the balance between efficiency and performance, we chose to detect four slopes with the range threshold of 0.015.

**Table 3.** ATOP80 values of different slope numbers and different range thresholds.

| TR  | 3   | 4   | 5   | 6   | 7   |
|-----|-----|-----|-----|-----|-----|
| 0.005 | 0.080 | 0.082 | 0.076 | 0.077 | 0.078 |
| 0.01 | 0.087 | 0.083 | 0.080 | 0.081 | 0.076 |
| 0.015 | 0.087 | 0.082 | 0.081 | 0.083 | 0.079 |
| 0.02 | 0.085 | 0.085 | 0.084 | 0.082 | 0.082 |

1 Number of detected slopes. 2 The range threshold. ATOP80, average error deviation of the top 80%.

**Table 4.** Average runtimes of different slope numbers and different range thresholds.

| TR  | 3   | 4   | 5   | 6   | 7   |
|-----|-----|-----|-----|-----|-----|
| 0.005 | 0.390 | 0.428 | 0.481 | 0.551 | 0.651 |
| 0.01 | 0.370 | 0.359 | 0.403 | 0.417 | 0.504 |
| 0.015 | 0.365 | 0.346 | 0.422 | 0.437 | 0.471 |
| 0.02 | 0.362 | 0.376 | 0.398 | 0.432 | 0.468 |

4.3. Comparison of Accuracy

Here, we compare our approach with the existing related skew correction methods, which are standard Hough transform (SHT) method in [39], projection profile analysis method [11], and Boudraa’s method [38]. The SHT method is applied on extracted contours of objects using Canny filter. Boudraa proposed a skew angle detection and correction method based on morphological skeleton and progressive probabilistic Hough transform (PPHT). Boudraa’s experimental result shows their method has the advantages of accuracy and effectiveness. We also 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) [37].

In our experiment, we also tested 200 binary document images of dataset of ICDAR2013 using the SHT method, PP method, and our proposed method based on the same platform and application. Since Boudraa’s method is only implemented based on QT/C++ platform, we directly refer the experimental result from his paper. Besides, as we used the same dataset, the comparison results are reliable. Based on the above indicators described in Section 4.1, the results of our experiments are shown in Table 5.

**Table 5.** Absolute Degree (°) of the seven tested methods applied to the document dataset.

| Method | Our Method | SHT 1 | PP 2 | Boudraa 3 | LRDE-EPITA-a | Ajou-SNU | LRDE-EPITA-b |
|--------|------------|-------|------|-----------|--------------|----------|--------------|
| AED    | 0.230      | 6.120 | 0.290| 0.078     | 0.085        | 0.085    | 0.097        |
| ATOP80 | 0.081      | 4.374 | 0.195| 0.051     | 0.051        | 0.051    | 0.053        |
| VEE    | 0.564      | 22.781| 0.065| 0.004     | 0.003        | 0.01     | 0.001        |
| VTOP80 | 0.003      | 12.752| 0.014| 0.001     | 0.001        | 0.001    | 0.002        |

1 Standard Hough transform method. 2 Projection profile analysis method. 3 Boudraa’s method. AED, average error deviation. VEE, variance of error estimation. VTOP80, variance of error estimation of the top 80%. 
From the Table 5, we find that our method has higher accuracy than the standard Hough transform method and the projection profile Analysis method. The results show that the SHT method has a large detection error because we tested various types of documents, and some documents are not suitable for the SHT method. As for Boudraa’s method, he uses morphological skeleton methods to improve the accuracy of detecting skew angle, so his method performs very well, but the calculation is also very complicated. Experimental results show that our method has high precision and can ensure the error within 0.1°.

### 4.4. Comparison of Runtime

Table 6 shows the runtime of three methods based on the same platform. Since we use the probability model, we only need to perform a simple division calculation on the selected two points, and then save the results. This method results in a short runtime compare with the other two methods. As we know, the PP method needs to rotate the document image many times to find the peak number of black pixels, so it has the longest runtime. The SHT method also takes a lot of time to convert pixels from image space to parameter space. As can be seen from Table 6, our method increases the computing speed by 5.04 and 9.94 times compared with SHT method and PP method.

#### Table 6. Mean and standard deviation of computational time (in seconds) using our proposed method when compared with the other methods.

| Runtime(s) | Our Method | SHT   | PP    |
|------------|------------|-------|-------|
| Mean 1     | 0.345      | 2.085 | 3.776 |
| SD 2       | 0.267      | 1.108 | 1.393 |

1 Mean of Computational time. 2 Standard Deviation of Computational time. SHT, standard Hough transform. PP, projection profile.

Figure 6 shows that the ATOP80 value of our method is 0.081, which is much smaller than that of the SHT method and PP method, and is only 0.3 larger than that of the-state-of-art method (Boudraa’s method). In addition, from the experimental results of [38], we can see that the computational time of Boudraa’s method is greater than that of SHT method and similar to the computational time of PP method. From Figure 6, we can see that our method is much faster than SHT method and PP method. So, it can be inferred that our method is at least five times faster than Boudraa’s method. These comparisons show that our approach can greatly improve the computational efficiency at the cost of little accuracy.

![Figure 6](image_url). The average errors deviations over all tested images and over the best 80% of images, and the average runtime of them.
5. Conclusions

In this paper, a novel fast and accurate method for skew angle detection and correction is proposed. The main novelty of our approach is that we combine the probability model, Q_test, and PP method to achieve a good balance between computational complexity and accuracy. Comprehensive experiments have demonstrated the advantage of our method in terms of efficiency, robustness, and accuracy. Especially in terms of high efficiency, our approach greatly reduces the runtime by randomly selecting two center points to calculate the slopes. In addition to the above advantages, our method also attains high accuracy in skew estimation over the most popular dataset involving various document types: books, papers, newspapers, letters, forms, and even cartoons, and so on. This dataset also includes diverse linguistic writings such as English, Chinese, Japanese, and Greek.

Moreover, with the aim of developing our work, some further research may need to be done. For instance:

- Improve our approach by replacing the fixed number of random selections in the slope calculation step with an adaptive value, which can further improve the speed of the algorithm.
- Develop a new extension of our method allowing to detect documents without characters or words, such as design drawings and so forth.

Author Contributions: Z.C. conceived and developed the ideas behind the research, performed the experiments and wrote the paper under the supervision of K.H. and M.Y. Authors X.Y. and A.Y. provided guidance and key suggestions. K.H. and X.Y. supervised the research and finalized the paper. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Science and Technology Major Project grant number 2017ZX01030-102-002.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Clark, P.; Mirmehdi, M. Rectifying perspective views of text in 3D scenes using vanishing points. Pattern Recognit. 2003, 36, 2673–2686. [CrossRef]
2. Zhang, Y.; Zhong, J.; Yu, H.; Kong, L. Research on deskew algorithm of scanned image. In Proceedings of the 2018 IEEE International Conference on Mechatronics and Automation (ICMA), Changchun, China, 5–8 August 2018; pp. 397–402.
3. Dance, C.R. Perspective estimation for document images. In Proceedings of the Document Recognition and Retrieval IX, San Jose, CA, USA, 19–25 January 2002; International Society for Optics and Photonics: San Jose, CA, USA, 2001; Volume 4670, pp. 244–254.
4. Gatos, B.; Pratikakis, I.; Perantonis, S.J. Adaptive degraded document image binarization. Pattern Recognit. 2006, 39, 317–327. [CrossRef]
5. Panwar, S.; Nain, N. A novel approach of skew normalization for handwritten text lines and words. In Proceedings of the 2012 Eighth International Conference on Signal Image Technology and Internet Based Systems, Naples, Italy, 25–29 November 2012; IEEE: Piscataway, NJ, USA, 2012; pp. 296–299.
6. Jiang, X.; Bunke, H.; Widmer-Klijajo, D. Skew detection of document images by focused nearest-neighbor clustering. In Proceedings of the Fifth International Conference on Document Analysis and Recognition. ICDAR’99 (Cat. No. PR00318), Bangalore, India, 22–22 September 1999; IEEE: Piscataway, NJ, USA, 1999; pp. 629–632.
7. Al-Khatatneh, A.; Pitchay, S.A.; Al-qudah, M. A review of skew detection techniques for document. In Proceedings of the 2015 17th UKSim-AMSS International Conference on Modelling and Simulation (UKSim), Cambridge, UK, 25–27 March 2015; IEEE: Piscataway, NJ, USA, 2015; pp. 316–321.
8. Papandreou, A.; Gatos, B.; Perantonis, S.J.; Gerardis, I. Efficient skew detection of printed document images based on novel combination of enhanced profiles. Int. J. Doc. Anal. Recognit. (IJDAR) 2014, 17, 433–454. [CrossRef]
9. Shafii, M.; Sid-Ahmed, M. Skew detection and correction based on an axes-parallel bounding box. Int. J. Doc. Anal. Recognit. (IJDAR) 2015, 18, 59–71. [CrossRef]
10. Brodić, D. The evaluation of the initial skew rate for printed text. *J. Electr. Eng.* 2011, 62, 134–140. [CrossRef]
11. Postl, W. Detection of linear oblique structures and skew scan in digitized documents. In Proceedings of the International Conference on Pattern Recognition, Paris, France, 27–31 October 1986; pp. 687–689.
12. Bloomberg, D.S.; Kopec, G.E.; Dasari, L. Measuring document image skew and orientation. In Proceedings of the Document Recognition II, San Jose, CA, USA, 6–7 February 1995; International Society for Optics and Photonics: San Jose, CA, USA, 1995; Volume 2422, pp. 302–316.
13. Jain, B.; Borah, M. A comparison paper on skew detection of scanned document images based on horizontal and vertical projection profile analysis. *Int. J. Sci. Res. Publ.* 2014, 4, 1–4.
14. Hough, P.V.C. Method and Means for Recognizing Complex Patterns. U.S. Patent 3,069,654, 18 December 1962.
15. Xu, L.; Oja, E.; Kultanen, P. A new curve detection method: Randomized Hough transform (RHT). *Pattern Recognit. Lett.* 1990, 11, 331–338. [CrossRef]
16. Boukharouba, A. A new algorithm for skew correction and baseline detection based on the randomized Hough Transform. *J. King Saud Univ. Comput. Inf. Sci.* 2017, 29, 29–38. [CrossRef]
17. Matas, J.; Galambos, C.; Kittler, J. Robust detection of lines using the progressive probabilistic hough transform. *Comput. Vis. Image Underst.* 2000, 78, 119–137. [CrossRef]
18. Boudraa, O.; Hidouci, W.K.; Michelucci, D. An improved skew angle detection and correction technique for historical scanned documents using morphological skeleton and progressive probabilistic hough transform. In Proceedings of the 2017 5th International Conference on Electrical Engineering-Boumerdes (ICEE-B), Boumerdes, Algeria, 29–31 October 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 1–6.
19. Hashizume, A.; Yeh, P.S.; Rosenfeld, A. A method of detecting the orientation of aligned components. *Pattern Recognit. Lett.* 1986, 4, 125–132. [CrossRef]
20. Lu, Y.; Tan, C.L. A nearest-neighbor chain based approach to skew estimation in document images. *Pattern Recognit. Lett.* 2003, 24, 2315–2323. [CrossRef]
21. Fabrizio, J. A precise skew estimation algorithm for document images using KNN clustering and fourier transform. In Proceedings of the 2014 IEEE International Conference on Image Processing (ICIP), Paris, France, 27–30 October 2014; IEEE: Piscataway, NJ, USA, 2014; pp. 2585–2588.
22. Chou, C.H.; Chu, S.Y.; Chang, F. Estimation of skew angles for scanned documents based on piecewise covering by parallelograms. *Pattern Recognit.* 2007, 40, 443–455. [CrossRef]
23. Dean, R.B.; Dixon, W.J. Simplified statistics for small numbers of observations. *Anal. Chem.* 1951, 23, 636–638. [CrossRef]
24. Gonzalez, R.C.; Woods, R.E. *Digital Image Processing*, 3rd ed.; Prentice-Hall, Inc.: Upper Saddle River, NJ, USA, 2007.
25. O’Gorman, L. The document spectrum for page layout analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* 1993, 15, 1162–1173. [CrossRef]
26. Liolios, N.; Fakotakis, N.; Kokkinakis, G. Improved document skew detection based on text line connected-component clustering. In Proceedings of the 2001 International Conference on Image Processing (Cat. No. 01CH37205), Thessaloniki, Greece, 7–10 October 2001; IEEE: Piscataway, NJ, USA, 2001; Volume 1, pp. 1098–1101.
27. Lu, Y.; Tan, C.L. Improved nearest neighbor based approach to accurate document skew estimation. In Proceedings of the Seventh International Conference on Document Analysis and Recognition, Edinburgh, UK, 6 August 2003; IEEE: Piscataway, NJ, USA, 2003; pp. 503–507.
28. Smith, R. A simple and efficient skew detection algorithm via text row accumulation. In Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, Canada, 14–16 August 1995; IEEE: Piscataway, NJ, USA, 1995; Volume 2, pp. 1145–1148.
29. Nitrogiannis, K.; Gatos, B.; Pratikakis, I. A combined approach for the binarization of handwritten document images. *Pattern Recognit. Lett.* 2014, 35, 3–15. [CrossRef]
30. Otsu, N. A threshold selection method from gray-level histograms. *IEEE Trans. Syst. ManCybern.* 1979, 9, 62–66. [CrossRef]
31. Motwani, M.C.; Gadiya, M.C.; Motwani, R.C.; Harris, F.C. Survey of image denoising techniques. In Proceedings of the GSPX, Santa Clara, CA, USA, 27–30 September 2004; pp. 27–30.
32. Jain, P.; Tyagi, V. A survey of edge-preserving image denoising methods. *Inf. Syst. Front.* 2016, 18, 159–170. [CrossRef]
33. Sethian, J.A.; Sethian, J.A. *Level Set Methods: Evolving Interfaces in Geometry, Fluid Mechanics, Computer Vision, and Materials Science*; Cambridge University Press: Cambridge, UK, 1996.

34. Benesty, J.; Chen, J.; Huang, Y. Study of the widely linear Wiener filter for noise reduction. In Proceedings of the 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, Dallas, TX, USA, 14–19 March 2010; IEEE: Piscataway, NJ, USA, 2010; pp. 205–208.

35. Gopinathan, S.; Kokila, R.; Thangavel, P. Wavelet and FFT Based Image Denoising Using Non-Linear Filters. *Int. J. Electr. Comput. Eng.* (2088–8708) **2015**, *5*, 1018–1026. [CrossRef]

36. Jiang, N.; Wang, L. Quantum image scaling using nearest neighbor interpolation. *Quantum Inf. Process.* **2015**, *14*, 1559–1571. [CrossRef]

37. Papandreou, A.; Gatos, B.; Louloudis, G.; Stamatopoulos, N. Icдар 2013 document image skew estimation contest (disec 2013). In Proceedings of the 2013 12th International Conference on Document Analysis and Recognition, Washington, DC, USA, 25–28 August 2013; IEEE: Piscataway, NJ, USA, 2013; pp. 1444–1448.

38. Boudraa, O.; Hidouci, W.K.; Michelucci, D. Using skeleton and Hough transform variant to correct skew in historical documents. *Math. Comput. Simul.* **2020**, *167*, 389–403. [CrossRef]

39. Mukhopadhyay, P.; Chaudhuri, B.B. A survey of Hough Transform. *Pattern Recognit.* **2015**, *48*, 993–1010. [CrossRef]

© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).