SKEW CORRECTION IN DOCUMENTS WITH SEVERAL DIFFERENTLY SKEWED TEXT AREAS

P. Saragiotis and N. Papamarkos
Electric Circuits Analysis Laboratory
Department of Electrical and Computer Engineering
Democritus University of Thrace, 67100 Xanthi, Greece

Keywords: Skew Correction, Text Area Localization, Connected Component Analysis, Linear Regression, Optical Character Recognition.

Abstract: In this paper we propose a technique for detecting and correcting the skew of text areas in a document. The documents we work with may contain several areas of text with different skew angles. In the first stage, a text localization procedure is applied based on connected components analysis. Specifically, the connected components of the document are extracted and filtered according to their size and geometric characteristics. Next, the candidate characters are grouped using a nearest neighbour approach to form words, in a first step, and then text lines of any skew, in a second step. Using linear regression, two lines are estimated for each text line representing its top and bottom boundaries. The text lines in near locations with similar skew angles are grown to form text areas. These text areas are rotated independently to a horizontal or vertical plane. This technique has been tested and proved efficient and robust on a wide variety of documents including spreadsheets, book and magazine covers and advertisements.

1 INTRODUCTION

Optical Character Recognition has become an increasingly important technology in the office automation software and lots of commercial document analysis systems are available to the end users. Document layout analysis is used in such systems to improve their capabilities in dealing with complicated document layouts and diverse scripts. An efficient and accurate method for determining document image skew is an essential need, which can simplify layout analysis and improve character recognition. Most document analysis systems require a prior skew detection before the images are forwarded for processing by the subsequent layout analysis and character recognition stages.

Document skew is a distortion that often occurs during scanning or copying of a document or as a design feature in the document's layout. This mainly concerns the orientation of text lines, where a zero skew occurs when the lines are horizontal or vertical, depending on the language and page layout. Skew estimation and correction are therefore required before the actual document analysis is done. Inaccurate de-skew will significantly deteriorate the subsequent processing stages and may lead to incorrect layout analysis, erroneous word or character segmentation, and misrecognition. The overall performance of a document analysis system will thereby be severely decreased due to the skew.

In addition, automatic skew detection and correction also have practical value in improving the visual appearance for facsimile machines and duplicating machines. Ideally, a skewed input could be automatically corrected to produce a desirable output in the machines for more pleasant reading.

In general, there can be three types of skew within a page: a global skew, when all text areas have the same orientation; a multiple skew, when certain text areas have a different slant than the others; and a nonuniform text line skew, when the orientation fluctuates within a line, e.g., a line is bent at one or both of its ends, or a line has a wave-like shape. In this work we focus on documents with multiple skew, which is also the novelty of the proposed technique.

This work is co-funded by the European Social Fund and National Resources - (EPEAEK-II) ARXIMHDHS 1, TEI Serron.

Saragiotis P. and Papamarkos N. (2007).
SKEW CORRECTION IN DOCUMENTS WITH SEVERAL DIFFERENTLY SKEWED TEXT AREAS.
In *Proceedings of the Second International Conference on Computer Vision Theory and Applications - IFP/IA*, pages 85-92

1.1 Related Work

A number of methods have previously been proposed for identifying document image skew angles. A survey was reported by Hull (1998) and an extended reference is made by O. Okun et al. (1999).

The main methods proposed in the literature may be categorized into the following groups: methods based on projection profile analysis, methods based on nearest-neighbor clustering, methods based on Hough transform, methods based on cross-correlation and methods based on morphological transform.

The projection profile is a histogram of text index pixels or representative (fiducial) points of characters such as centers of connected components bounding boxes, for example, along a given direction. The points are projected in multiple directions and the variation in the obtained projection profile is calculated for each direction. The angle corresponding to the maximum variation is the desired skew.

The Hough transform is another popular technique for skew detection. This transform is often applied to a number of representative points of characters such as the lowermost pixels or centres of gravity. Each representative point (x,y) is mapped from the Cartesian space to the points (ρ,θ) in the Hough space by forming a set of lines coming through (x,y) with a slope θ and distance ρ from the origin. The skew corresponds to the angle associated with a peak in the Hough space.

Most of the above methods have their inherent weakness, because most of them actually are tailor-made algorithms that are applicable to a particular document layout. As a result, some of them may fail to estimate skew angles of documents containing complicated layouts with multiple font styles and sizes, arbitrary text orientation and script, or high proportion of non-text regions such as graphics and tables. Moreover, they estimate a single or the dominant skew angle of the document and fail to recognize multiple skew angles.

Messelodi S. and Modena (1999) showed that projection profiles in combination with a clustering procedure based on simple heuristics may overcome the problem of the limited angle range. Although this method can detect multiple skew and small interline spacing, it was only tested on small-sized (512x512 pixels) images of book covers containing a few text lines. Gatos et al. (1997), uses an interline cross-correlation for two or more vertical lines located at a fixed distance d for skew estimation. The cross-correlation function is computed for an entire image to obtain the documents skew angle. This can, however, be time consuming and the presence of graphics degrades the accuracy. Y. Lu, and C. L. Tan (2003) folowed a nearest-neighbor chain based approach developing a skew estimation method with a high accuracy and with language-independent capability. Their approach detects only a dominant skew for the document.

2 PROPOSED TECHNIQUE

The proposed technique aims in correcting the skew in documents that contain several areas with text bend in different slopes and be robust enough to handle a great variety of printed documents, including book and magazine covers, spreadsheets among with regularly layout documents in any language. This assumes that the text is supposed to be correctly oriented in either vertical or horizontal alignment. It is also assumed that the document can contain from several down to a single text area slopes.

To achieve this goal a bottom-up approach is applied which is better suited to the specific problem. First, the document is prepossessed to identify the text colour index. Then, the connected components of the document are retrieved using a simple serial labelling algorithm and their bounded rectangles are constructed. A filtering procedure is applied to the connected components to discard non-text connected elements according to their geometrical characteristics and indicate the candidate characters. These candidate characters are grouped using a nearest neighbour approach to form words. The words are grouped, based on a rough slope calculation, to form lines of text. Using linear regression on the edge pixels of the connected components bounding rectangles, a set of straight lines is estimated for each text line representing its top and bottom boundaries. The text lines in near locations with similar skew angles are grown to form text areas and their slope is defined according to the slope of the text line boundary lines. The connected components that have been filtered or failed to construct words, included in a text area are supposed to be part of that text area. Finally, each text area is rotated to a horizontal or vertical plane taking measures to avoid the possibility of overlapping.

The result of the technique is a single binary image ready to be processed by the layout analysis module of an OCR system. Next, analysis of the main stages of the proposed technique is given.

2.1 Preprocessing

The proposed technique is applied to grey scale document images that have a resolution high enough to retain separated elements. For example, the scanning resolution of book cover could be as low as 50dpi but a dense written document must be scanned with 300dpi. The document is filtered with a Gaussian filter to remove noise and strengthen the connectivity of its elements. The document is binarized using the well known Otsu's threshold selection method (Otsu N., 1979).

2.2 Connected Component Analysis

After preprocessing, the index colour of the binary image that represents text is chosen and a simple serial algorithm is used to label the elements of the document based on its 4-th neighbourhood connectivity. Each document image line is scanned looking for text index pixels. When found its upper and left pixels are checked for labels. If they are labelled with the same label the current pixel gets that label. Otherwise the equality of the labels is marked on an equality array. A second scan of the document image lines is preformed and the equal labels are replaced with a single one.

For each connected element a bounding rectangle is constructed. We have no knowledge for the nature of the connected element; it could be a character, connected characters or even images from the document. These bounding rectangles form the skeleton for all future analysis on the page. The position and dimensions of the bounding rectangle are marked as well as the points where the text index pixels touches its edges, called now on boundary rectangle edge pixels.

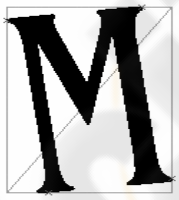

Figure 1: Bounding rectangle with its edge pixel marked.

2.3 Filtering

At this stage each connected component accompanied by its bounding rectangle represents a region of text index pixels without any further knowledge of its contents. We assume that the set of connected components will contain all text components mixed with several non-text ones. We will try to identify the non-text or unreadable text components by analyzing their geometrical structure (Y. Zhong et al., 1995; W.Y. Chen, S.Y. Chen, 1998) and discard them. We specify some size and geometrical filters in order to discard components which are likely to correspond to non-textual objects. Filters selection is based on the analysis of some internal features, i.e. inherent to the single connected component, and possibly to its neighbourhood. The internal features that are used in order to eliminate connected components that have low probability to be text objects are

- Area. The area of a component is defined as the number of its bounding rectangle pixels divided by the scanning resolution of the document. Connected components with area less than 2 mm^2 are suppressed in the proposed technique, as we supposed they correspond to noisy connected components. The punctuation marks are removed at this stage. This fact is acknowledged and used later on the processing. Connected Components with area larger than 100 mm^2 are removed also as they probably represent large non-text objects.
- Density. It is defined as the ratio between the component area and the area of its convex hull. It permits to detect sparsely filled or compact connected components. In the proposed technique we remove connected components with density less than 10%, as we suppose they are line art, or greater than 70% as they are probably images or frames.
- Width to Height Ratio. The width to height ration permits to detect long components or narrow components. Width to height ratio is a filter that regards the orientation of text. Thus, width to height ration in horizontal aligned text is height to width ratio in vertical aligned text. In the proposed technique we prevent connected components with width to height ration more than 150% or less than 30% to form horizontal text lines, as most characters proportions are between those limits.

In all three filter rules the proposed technique uses, a small probability exists of classifying text connected components as non-text. This possibility is being acknowledged and connected components that have been removed will be included in the text area growing step of the proposed technique.

The connected components that remain after the filtering stage are considered candidate characters that will form words and text lines.

2.4 Word Grouping

At this stage of the proposed technique, the candidate characters are grouped to form words aligned either horizontally or vertically. The grouping is based on the Euclidean distance of the candidate characters bounding rectangles (C. Strouthopoulos et al, 1997). First we group the horizontal aligned candidate characters. Let WH be an ordered group of connected components:

$$WH : \{C_1, C_2, \cdots, C_n\} \tag{1}$$

And $WH_{averageWidth}$ is the average width of its members bounding rectangles:

$$WH_{averageWidth} = \frac{\sum_{i=1}^{n} C_1^{Width}}{n} \tag{2}$$

For each candidate character C_l that is not a member of any ordered group or it is the first member of an ordered group WH_k the Euclidean distance $d_{l\to n} = \left|\overline{C_l^{ml} - C_n^{mr}}\right|$ between its middle left BR edge point and the middle right BR edge point of the C_n is calculated. If this distance is less than the maximum of $WH_{averageWidth}$ and C_l^{Width} multiplied by a factor *f* chosen to be equal to 1.0:

$$d_{l\to n} < \max\left(WH_{averageWidth}, C_l^{Width}\right) \cdot f \tag{3}$$

C_l is candidate next character of the ordered group WH . A list with all candidate members C_l of WH is formed and the C_l with the smallest Euclidean distance $d_{l\to n}$ will be the next character of the ordered group WH . If the chosen candidate character C_l is the first member of the ordered group WH_k then the two groups are merged to create the new group WH' .

$$WH' = WH \cup WH_k \tag{4}$$

This procedure is repeated until no candidate next character can be found for any ordered group WH . The algorithm ensures that all possible grouping has been done in the less possible repetitions.

To group the vertically aligned candidate characters in WV ordered groups, we use the same reparative algorithm considering the widths in the vertical direction and calculating and comparing the Euclidean distance $d_{l\to n} = \left|\overline{C_l^{mb} - C_n^{mt}}\right|$ between the middle BR top edge point of C_n and the middle bottom BR edge point of the C_l .

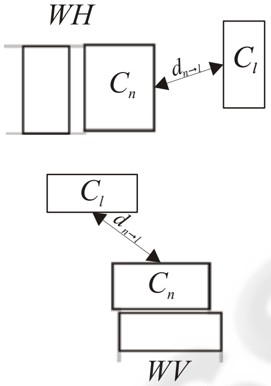

Figure 2: Construction of the Horizontal and Vertical Words using the distance between the connected components.

Either horizontal or vertical groups with less than 5 members are discarded. This may seem that these words won't be aligned at the result of this technique, but that is not the case. It has been observed that small groups create problems in the text line identification stage. Furthermore, small words like those are most often part of a greater text area with the some alignment. So, the proposed technique includes those words in the text area growing stage.

The next step of the word grouping stage is the revocation of the ambiguity of a candidate character C_l belonging to both a horizontal WH and a vertical WV word. This is achieved efficiently with a simple rule. If the members of WV are more than those of WH then the WH word is destructed and C_l remains a member of WV , else the WV word is destructed.

2.5 Text Line Identification

To identify text lines in a document we must merge sets of identified words and candidate characters that failed to construct words in the previous steps. These words and candidate characters are separated by large gaps (otherwise they would have been merged in the previous stage of the proposed technique). The first step in doing so is to calculate a rough angle

ϑ_{WH} for each horizontal and vertical word (ϑ_{WV}) as the average slope between the central points of the BR edges of consecutive characters C_i :

$$\vartheta_{WH} = \frac{\sum_{i=1}^{n-1} \tan\left(C_i^{rm} \to C_{i+1}^{lm}\right)}{n} \qquad (5)$$

Then a line is extended from each side of a word with length $WH_{averageWidth} \cdot f_{word}$ and angle ϑ, for the horizontal words first. The angle ϑ is measured from the *x* axis and the right side of the word. At the left side of the word the angle is $-\vartheta$. This line may cross the vertical edges of several candidate characters. The candidate character C_l that is crossed by the line with the smaller distance $d_{l \to n}$ will be the last character or first, depending on the line, of the ordered group WH . Again, if the chosen candidate character C_l is the first or last member of the ordered group WH_k and the rough calculated angles of the two groups ϑ, ϑ_k differ less than a predefined ϑ_{rd} then the two groups are merged to create the new group WH'. The value of ϑ_{rd} has been calculated to ensure both that errors in the rough calculation of the angle will not restrict two adjacent words to create a new group and unaligned words won't be grouped. The value used is 20^o.

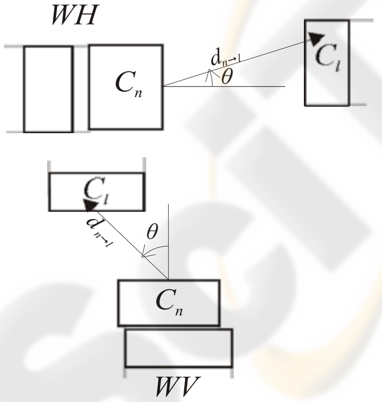

Figure 3: Construction of the Horizontal and Vertical Words using the distance between the connected components.

Next, the same step is repeated for the vertical aligned words. The angle ϑ_{WV} is measured from the *y* axis and the top side of the word. At the bottom side of the word the angle is $-WV^{\vartheta}$

At the end of this stage, the text lines have been identified and are represented as ordered groups of characters WH and WV .

2.6 Text Line Skew Calculation

In order to determine the skew angle of a text line, we first estimate its lower and top base line. This approach is similar to the procedure used by U.-V. Marti and H. Bunke (2000). Specifically, the position of the bottom edge pixel from the characters C_i of each identified horizontal text line WH is used to construct the set P_1 of pixels. The set P_2 of pixels is constructed from the position of the top edge pixels. The sets P_1, P_2 of pixels approximate the lower and upper contour of the text lines. Formally, it can be represented as follows:

$$P_1 = \{p_i = (x_i, y_i) | \text{ bottom edge of } C_i\} \qquad (6)$$

$$P_2 = \{p_i = (x_i, y_i) | \text{ top edge of } C_i\} \qquad (7)$$

Assume that each set P of P_1, P_2 has k entries. On this set of points a linear regression can be applied. The final goal of skew detection is to find the parameters a and b of a straight line expressed by the following equation:

$$x = ax + b \qquad (8)$$

For this purpose the mean values of the two variables x and y have to be computed:

$$\mu_x = \frac{1}{k}\sum_{i=1}^{k} x_i \text{ , } \mu_y = \frac{1}{k}\sum_{i=1}^{k} y_i \qquad (9)$$

Then, the line parameters a and b can be obtained using the following two formulas:

$$a = \frac{\sum_{i=1}^{k} x_i y_i - k\mu_x \mu_y}{\sum_{i=1}^{k} x_i^2 - k\mu_x^2} \text{ , } b = \mu_y - a\mu_x \qquad (10)$$

Linear regression minimizes the error between the line and the given set of points. A problem with this approximation is, however, that outliers and punctuation marks in the set P may influence the result disturbing the calculated line. The punctuation marks have been removed by filtering. For the task of baseline estimation, descender characters can be regarded as outliers. The same will be considered for capital characters in a line of lowercase characters or ascender characters in top line estimation. To reduce their influence on the regression line, the summed square error between the line and the set P is computed by:

$$e = \sum_{i=1}^{k} \left(ax_i + b - y_i \right)^2 \qquad (11)$$

If the total error e is larger than a predefined threshold t_e, the point p_i with the largest amount in the sum is eliminated from set P. This procedure is repeated until the error e is smaller than threshold t_e. The procedure is also stopped if $t_{\max removed}$ percentage of points has been removed from the set. This is an indication of a poor result and it is taken into account in the area growing stage.

To estimate the top and baseline of the vertical aligned text lines we use the sets of pixels P_1, P_2:

$$P_1 = \left\{ p_i = \left(x_i, y_i \right) \middle| \text{ right edge of } C_i \right\} \qquad (12)$$

$$P_2 = \left\{ p_i = \left(x_i, y_i \right) \middle| \text{ left edge of } C_i \right\} \qquad (13)$$

The skew angle of the text line is $\vartheta = \tan^{-1}(a)$. Where a is selected from Table 1. The result of this stage can be seen in Figure 4.

Table 1: Selection of a value used to calculate the skew angle of a text line based on the regression result of the set of fitted lines.

Accepted Result on	Selected a
P_1	a_1
P_2	a_2
none	a_1

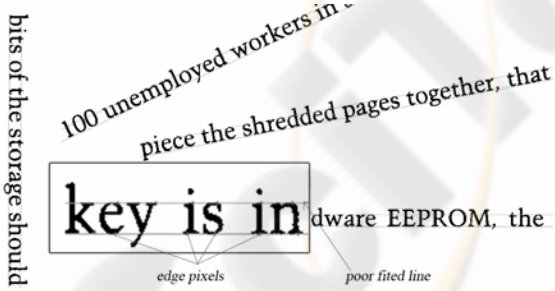
bits of the storage should
100 unemployed workers in
piece the shredded pages together, that
key is in
dware EEPROM, the
edge pixels
poor fited line

Figure 4: Fitting of a pair of lines in horizontal and vertical aligned text. In the zoomed part of the Figure edge pixels can identified as well as poor fit indication.

2.7 Text Area Growing

The next stage of the proposed technique is the construction of the text areas. The seed in this procedure are the text lines and their estimated skew. A text area is created for each identified text line. The text area is a rectangle rotated by an angle ϑ from the *x* axis containing all the text index pixels of characters C_i members of the text line as shown in Figure 5. This is done efficiently by using only the edge pixels of the characters bounding boxes shown on Figure 1.

Two text areas A_j, A_k are grown into an area $A_l = A_j \cup A_k$ when the following two conditions hold:

- There is no candidate character C_i member of an area A_m contained in the result rectangle of area A_l.
- The two areas rectangle rotation angles ϑ_j, ϑ_k differ only be a few degrees ϑ_d. The chosen value for ϑ_d is 5°, a value estimated from the experiments that will allow most adjacent text areas with the same skew to be joined. It has also been observed that adjacent text areas, differently aligned in a document have skew angles many times greater than the selected value.

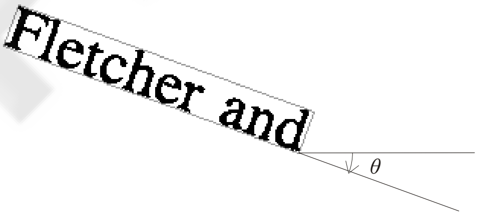

Figure 5: Construction of a text area from a single identified text line.

The resultant's area A_l rectangle will contain all the text index pixels of characters C_i members of both areas and its skew angle ϑ_l will be calculated as a weighted average:

$$\vartheta_l = \frac{\vartheta_j m_j + \vartheta_k m_k}{m_j m_k} \qquad (14)$$

Where m_j, m_k are the C_i member counts for each text area. If the rotation angle ϑ_k of text area A_k, in the previous stage, is a result of two poor fitted lines,

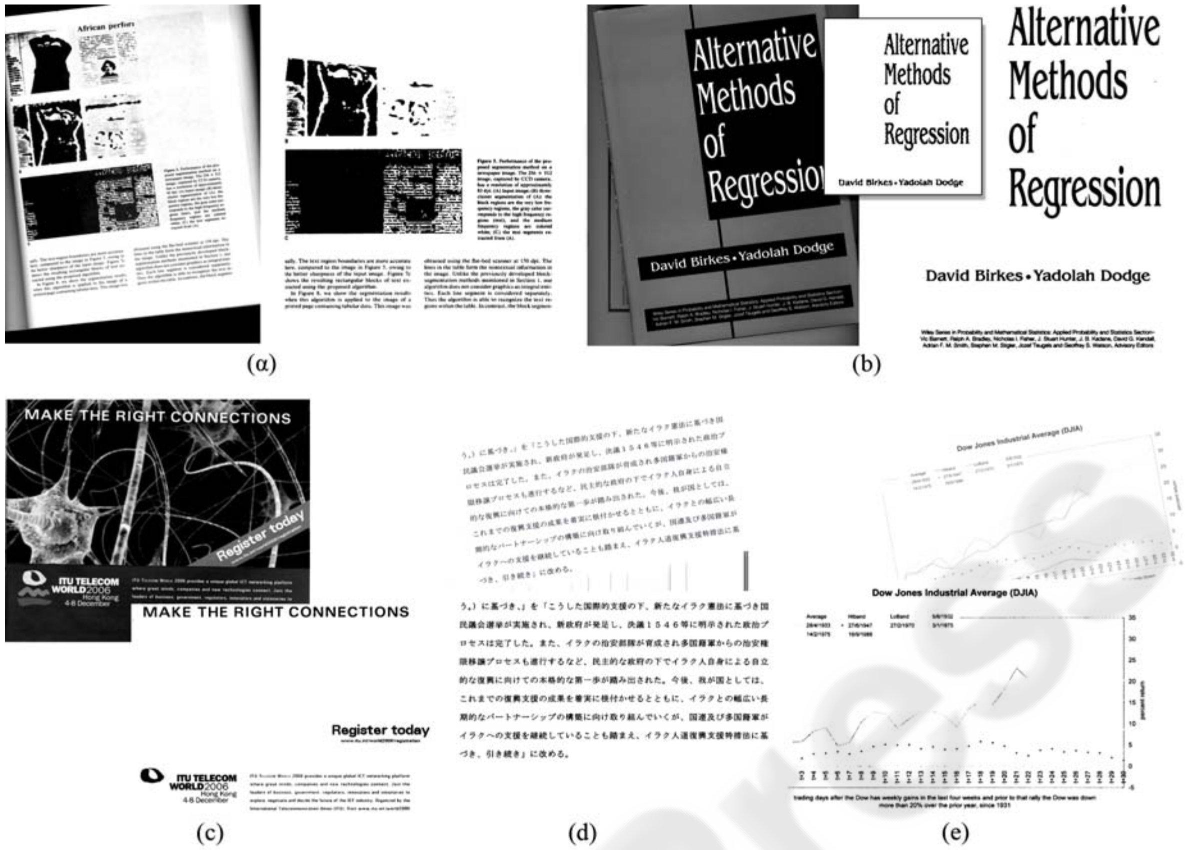
African perfor
(α)
Alternative Methods of Regression
David Birkes • Yadolah Dodge
(b)
MAKE THE RIGHT CONNECTIONS
Register today
ITU TELECOM WORLD 2006
Hong Kong
4-8 December
(c)
(d)
Dow Jones Industrial Average (DJIA)
(e)

Figure 6: Experimental result. (a) skewed magazine page, (b) book cover scanned with 200dpi and 50dpi (in centre), (c) advertisement with multiple skewed text areas, (d) skewed paragraph in Japanese and (e) spreadsheet with multiple skewed text areas.

then the rotation angle ϑ_l of the resultant text area A_l will be ϑ_j. With this measure the proposed technique eliminates the skew error that results from poorly fitted regression lines.

The step of area growing by joining two areas is repeated until no more areas can be joined. The next step of this stage of the proposed technique is the inclusion of candidate characters that failed to form words and the connected components that have been filtered, into text areas.

The candidate characters C_l that failed to form words are joined with its nearest text area by adding them to the area. Text areas are not ordered groups like words and text lines, so the adding of a member isn't difficult. The text area's rectangle is grown to include the text index pixels of the newly added member. The rotation angle does not change.

The connected components that have been filtered, i.e. punctuation marks and non-text elements, whose bounding box has common areas with a text area, are considered parts of that area and will be rotated by the text areas rotation angle ϑ. Connected components outside the text areas are considered noise, or frames.

At the end of this stage, the text of the document has been localized and its skew rotation is measured.

2.8 Text Area Rotation

The last stage of the proposed technique is the rotation of the identified text areas to a vertical or horizontal orientation. To avoid the possibility of overlapping target rectangles, they are moved left as needed. The pixels of each area members are projected to the target rectangle using bilinear interpolation. Bilinear interpolation is used to minimize the effect of artefacts produced by simply rotating binary images.

The result of this stage is a single layered binary image ready to be processed by the layout analysis module of an OCR system.

3 EXPERIMENTAL RESULTS

The proposed technique has been tested to a great variety of complex documents with a single or multiple skews, images and graphics that where scanned with a variety of resolutions. The documents tested were magazine and book pages, spreadsheets, book covers and advertisements. Some of the results, emphasizing on the ability of the proposed technique to handle different types of documents with single or multiple skews, can be seen in Figure 6.

To test the proposed technique's accuracy we scanned with 200dpi a document paragraph taken from a magazine column with great effort in aligning the paragraph correctly. Then we rotated the paragraph in multiple angles and constructed a document with all these rotated paragraphs. The results for the detected skews from the proposed technique are reported on Table 2.

Table 2: Results indicating the accuracy of the proposed technique. Rotation is the angle by which a text area was rotated and Estimated Skew is the result of the proposed technique.

Rotation	**Estimated Skew**	**Difference**
35	35.02	0.02
25	24.96	0.04
0	0.01	0.01
-5	-4.99	0.01
-15	-14.98	0.02
-65	-65	0.00
-75	-75.04	0.04
-85	-85.01	0.01
Average Error		0.01875

The experiments run on an Intel Pentium 4 CPU running at 2.8 GHz. The processing of the documents depends on its complexity as the technique uses connected component analysis. For most of the documents tested processing, without the pre-processing stage, took less than a second.

4 CONCLUSIONS

In this paper, a new technique is proposed for skew correction in documents with several differently skewed text areas. The technique shown to be robust in handling a variety of documents such as magazine and book pages, spreadsheets, book covers and advertisements in multiple languages and unlimited rotation angles.

The main contribution of the proposed technique is its ability to correct the skew in documents with several differently skewed text areas. The novelties of our approach are the vertical and horizontal grouping of candidate characters, which improves the ability to handle a great range of skew angles, the calculation of two skew angles for each identified text line, which improves the techniques accuracy, and the area growing technique which allows handling of unclassified connected components and punctuation marks.

REFERENCES

W.Y. Chen, S.Y. Chen, *Adaptive page segmentation for color technical journal's cover images,* Image and Vision Computing 16, pp. 855-877, 1998.

B. Gatos, N. Papamarkos and C. Chamzas, *Skew detection and text line position determination in digitized documents*, Pattern Recognition, Vol. 30, No. 9, pp. 1505-1519, 1997.

J.J. Hull, *Document image skew detection: survey and anotated bibliography. In: Hull, J.J., Taylor, S.L. (Eds.)*, Document Analysis Systems II. World Scientific, pp. 40–64, 1998.

Y. Lu, and C. L. Tan, *A nearest-neighbor chain based approach to skew estimation in document images*, Pattern Recogn. Lett. 24, 14, pp. 2315-2323, 2003.

U.-V. Marti, H. Bunke, *Using a statistical language model to improve the performance of an HMM-based Cursive Handwriting Recognition System*, Internat. J. Pattern Recognit. Artificial Intell. 15 (1), 65—90, 2000.

S. Messelodi, C.M. Modena, *Automatic identication and skew estimation of text lines in real scene images*, Pattern Recognition 32:5, 791-810, 1999.

O. Okun, M. Pietikainen, and J. Sauvola, *Document skew estimation without angle range restriction*, International Jurnal on Document Analysis and Recognition, pp. 132-144, 1999.

N. Otsu, *A Threshold selection method from gray-level histograms*, IEEE Tran. on System Man and Cybernetics, SMC-9 (1), pp. 62-69, 1979.

C. Strouthopoulos, N. Papamarkos and C. Chamzas, *Identification of text-only areas in mixed type documents*, Engineering Applications of Artificial Intelligence, Vol. 10, No. 4, pp. 387-401, 1997.

Y. Zhong, K. Karu, A.K. Jain, *Locating text in complex color images*, Pattern Recognition, 28 (10), pp. 1523-1535, 1995.