J. King Saud Univ., Vol. 15, *Comp. & Info. Sci.*, pp. 67-86 (A.H. 1423/2003)

Skew Correction of Textural Documents

Ibrahim S. I. Abuhaiba
Department of Electrical and Computer Engineering,
Islamic University of Gaza, Gaza, P. O. Box 1276, Palestine
E-mail: isabuhaiba@hotmail.com

(Received 01 January 2002; accepted for publication 30 April 2002)

Abstract. Two algorithms for accurate skew detection and correction of textual documents are presented. They depend on finding a horizontal RLSA image of the skewed document. The average skew of selected black connected components in the RLSA image is considered as the skew angle for the whole document which is finally rotated in the opposite direction by that amount to obtain the final corrected image. The first algorithm is a single-pass skew detection and correction algorithm, i.e., the input image is either not rotated or rotated only once. The second algorithm is a multi-pass skew detection and correction algorithm, i.e., the input image is either not rotated or rotated at least once. A total of 50 pages were used to test the algorithms. The maximum absolute measured error after correction is 1.53° and 0.36° for the first and second algorithms, respectively.

Keywords: Printed documents, Skew correction, Layout analysis.

1. Introduction

A document originally has zero skew, but when a page is manually scanned or photocopied, nonzero skew may be introduced. The skew may cause problems in text baseline extraction and document layout analysis techniques which assume the Manhattan layouts, that is, the layouts whose blocks are separable by vertical and horizontal cuts. For example, in the recursive projection histogram splitting method [1], document area segmentation will not be possible if the text lines are rotated so that no significant valleys can be detected in the projection histogram. Therefore, it is often necessary to determine the skew angle before structural analysis.

Several methods have been developed by many researchers for skew angle detection. In [2-5,] methods using the projection histogram are proposed. A restriction of

these methods is that they are limited to documents that have fairly small skews that are typically less than ±10°. In [3], a method based on Fourier transform, which can be time consuming for a large image, is proposed. In some cases, the skew angle can be determined from text margins if they exist in the scanned image [6]. The Hough transform-based methods have been used by several people [7-12]. One limitation of this approach is that if text is sparse it may be difficult to choose correctly a peak in Hough space. The skew angle can also be determined from the centers of character boxes using a least-squares line fitting technique [13]. In [14], the skew angle is determined using cross-correlation between lines at a fixed distance. Methods based on nearest-neighbour clustering are described in [15, 16]. In these methods, noise, subparts of characters (dot on "i"), and between-line connections can reduce their accuracy. In [17], a method based on the detection of headlines of document words is presented for skew angle detection of Indian script documents. As reported by the authors, the method is faster than the conventional Hough transform method.

In this paper, skew of textual documents is detected with high accuracy. The new method depends on finding a horizontal RLSA image of the skewed document. The average skew of selected black connected components in the RLSA image is considered as the skew angle for the whole document which is finally rotated in the opposite direction by that amount to obtain the final corrected image. Our method has advantages over others:

- The method is not limited, in contrast to methods using the projection histogram [2-5], to documents that have fairly small skews that are typically less than ±10°.
- Peak-finding problems do not exist in our method as compared to the Hough transform-based methods [7-12].
- Noise, subparts of characters (dot on "i"), and between-line connections can reduce the accuracy of the methods based on nearest-neighbour clustering [15, 16]. Such problems don't exist in our method.

The paper is organized as follows. A one-pass skew correction algorithm is presented in Section 2. A more accurate multi-pass skew correction algorithm is introduced in Section 3. Experimental results are reported in Section 4. Finally, conclusions are given in Section 5.

2. One-pass Skew Correction

A block segmentation technique, called RLSA (run length smoothing algorithm), was proposed in [18]. The RLSA algorithm is applied to a binary sequence in which white pixels are represented by 0's and black pixels by 1's. The algorithm transforms a binary sequence x into an output sequence y according to the following rules: (1) 0's in x

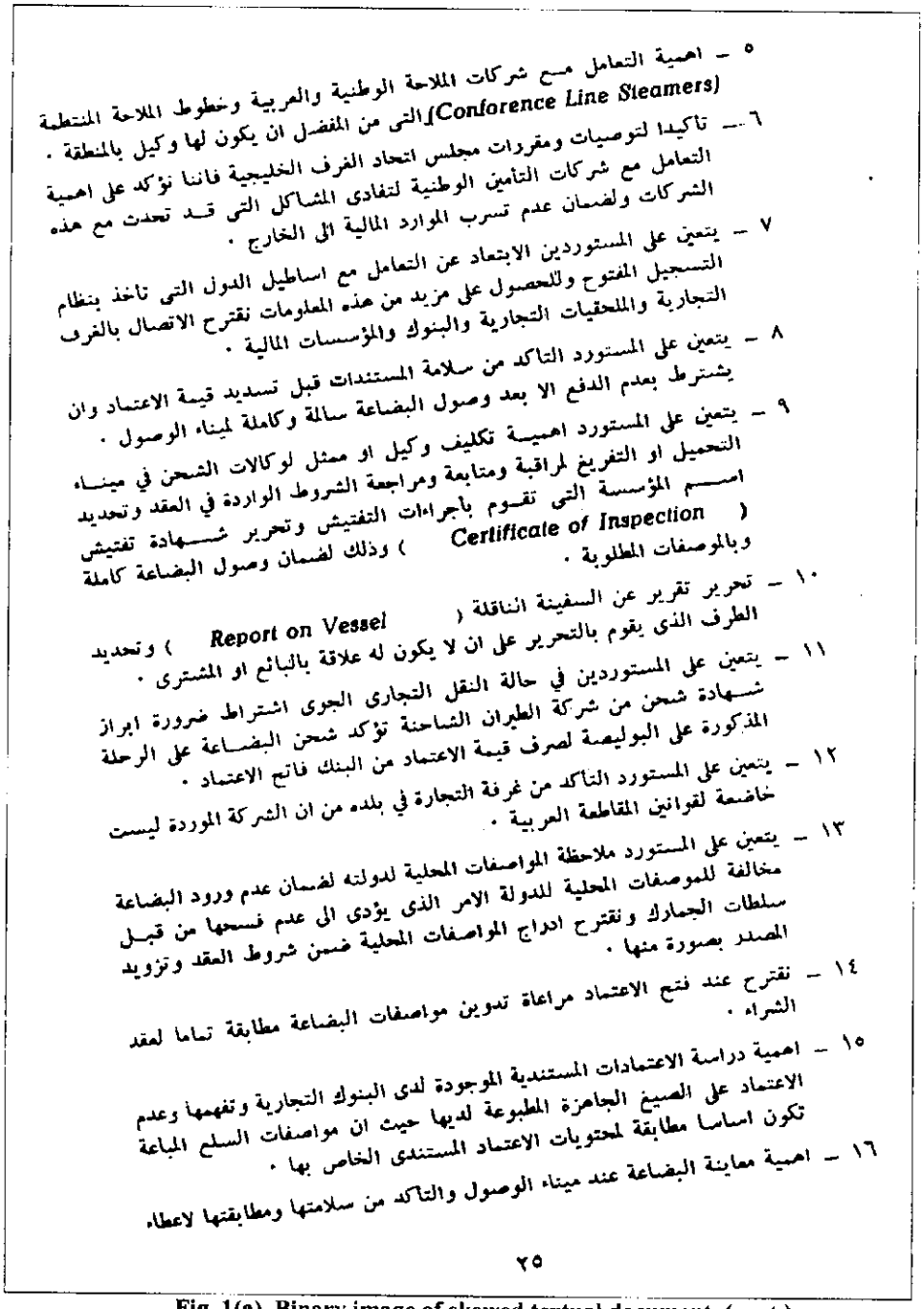
are changed to 1's in y if the number of adjacent 0's is less than or equal to a certain limit r, and (2) 1's in x are unchanged in y. For example, if the sequence x equals 00011000001100100001 and the value of r is 3, then the output sequence y is 11111000001111100001.

Figure 1 (a) shows a skewed image of a textual document. By applying the RLSA row-by-row, the bit map is obtained in Fig. 1(b). Our idea for skew angle detection depends on finding the average angle of elongated smeared black blocks. For simplicity, let us assume that Fig. 2 represents the RLSA block of one skewed line of text, where ULC and LRC are the upper left corner and lower right corner of the minimum horizontal bounding rectangle which surrounds the block. The angle of this block can be determined as follows:

1. Find the uppermost black point of the block, which lies at a distance d to the right of ULC, point A in Fig. 2.
2. Find the lowest black point of the block, which lies at a distance d to the right of ULC, point B in Fig. 2.
3. Find the middle point, U, of the two points found in Steps 1 and 2.
4. Find the uppermost black point of the block, which lies at a distance d to the left of LRC, point C in Fig. 2.
5. Find the lowest black point of the block, which lies at a distance d to the left of LRC, point D in Fig. 2.
6. Find the middle point, V, of the two points found in Steps 4 and 5.
7. The angle of the straight line which connects the two middle points, U and V, constitutes the skew angle of the block and hence the skew angle of the corresponding line of text.

For a textual document such as that in Fig. 1(a), these steps are repeated for every selected block in the RLSA image. The average angle of all blocks constitutes the skew angle of the whole document. The accuracy of this method depends on:

1. How much the minimum bounding rectangles of the RLSA blocks are elongated. In other words, the more elongated such rectangles are the more accurate skew angle is. As shown in Fig. 1(b), the RLSA blocks of adjacent lines of text may merge together yielding one block with a less elongated rectangle that will give a less accurate skew estimation. This negative effect can be minimized by controlling the smearing threshold, r, as follows. Consider Fig. 3 which shows the skewed minimum bounding rectangles of two adjacent lines of text. Here, we assume that the skewed minimum bounding rectangle of a line encloses everything that belongs to that line and only to that line. Let us assume that g_{min} represents the minimum vertical gap between any two adjacent rectangles in the document. Let A be a black point that lies on the upper side of the rectangle of the lower line. Similarly, let point B be a black point that lies on the lower side of the rectangle of the upper line such that the straight line



٥ – اهمية التعامل مع شركات الملاحة الوطنية والعربية وخطوط الملاحة المنتظمة (Conforence Line Steamers) التى من المفضل ان يكون لها وكيل بالمنطقة .

٦ – تأكيدا لتوصيات ومقررات مجلس اتحاد الغرف الخليجية فاننا نؤكد على اهمية التعامل مع شركات التأمين الوطنية لتفادى المشاكل التى قد تحدث مع هذه الشركات ولضمان عدم تسرب الموارد المالية الى الخارج .

٧ – يتعين على المستوردين الابتعاد عن التعامل مع اساطيل الدول التى تاخذ بنظام التسجيل المفتوح وللحصول على مزيد من هذه المعلومات نقترح الاتصال بالغرف التجارية والملحقيات التجارية والبنوك والمؤسسات المالية .

٨ – يتعين على المستورد التاكد من سلامة المستندات قبل تسديد قيمة الاعتماد وان يشترط بعدم الدفع الا بعد وصول البضاعة سالمة وكاملة لميناء الوصول .

٩ – يتعين على المستورد اهمية تكليف وكيل او ممثل لوكالات الشحن في ميناء التحميل او التفريغ لمراقبة ومتابعة ومراجعة الشروط الواردة في العقد وتحديد اسم المؤسسة التى تقوم باجراءات التفتيش وتحرير شهادة تفتيش (Certificate of Inspection) وذلك لضمان وصول البضاعة كاملة وبالموصفات المطلوبة .

١٠ – تحرير تقرير عن السفينة الناقلة (Report on Vessel) وتحديد الطرف الذى يقوم بالتحرير على ان لا يكون له علاقة بالبائع او المشترى .

١١ – يتعين على المستوردين في حالة النقل التجارى الجوى اشتراط ضرورة ابراز شهادة شحن من شركة الطيران الشاحنة تؤكد شحن البضاعة على الرحلة المذكورة على البوليصة لصرف قيمة الاعتماد من البنك فاتح الاعتماد .

١٢ – يتعين على المستورد التاكد من غرفة التجارة في بلده من ان الشركة الموردة ليست خاضعة لقوانين المقاطعة العربية .

١٣ – يتعين على المستورد ملاحظة المواصفات المحلية لدولته لضمان عدم ورود البضاعة مخالفة للموصفات المحلية للدولة الامر الذى يؤدى الى عدم فسحها من قبل سلطات الجمارك ونقترح ادراج المواصفات المحلية ضمن شروط العقد وتزويد المصدر بصورة منها .

١٤ – نقترح عند فتح الاعتماد مراعاة تدوين مواصفات البضاعة مطابقة تماما لعقد الشراء .

١٥ – اهمية دراسة الاعتمادات المستندية الموجودة لدى البنوك التجارية وتفهمها وعدم الاعتماد على الصيغ الجاهزة المطبوعة لديها حيث ان مواصفات السلع المباعة تكون اساسا مطابقة لمحتويات الاعتماد المستندى الخاص بها .

١٦ – اهمية معاينة البضاعة عند ميناء الوصول والتاكد من سلامتها ومطابقتها لاعطاء

٢٥

Fig. 1(a). Binary image of skewed textual document, (cont.).

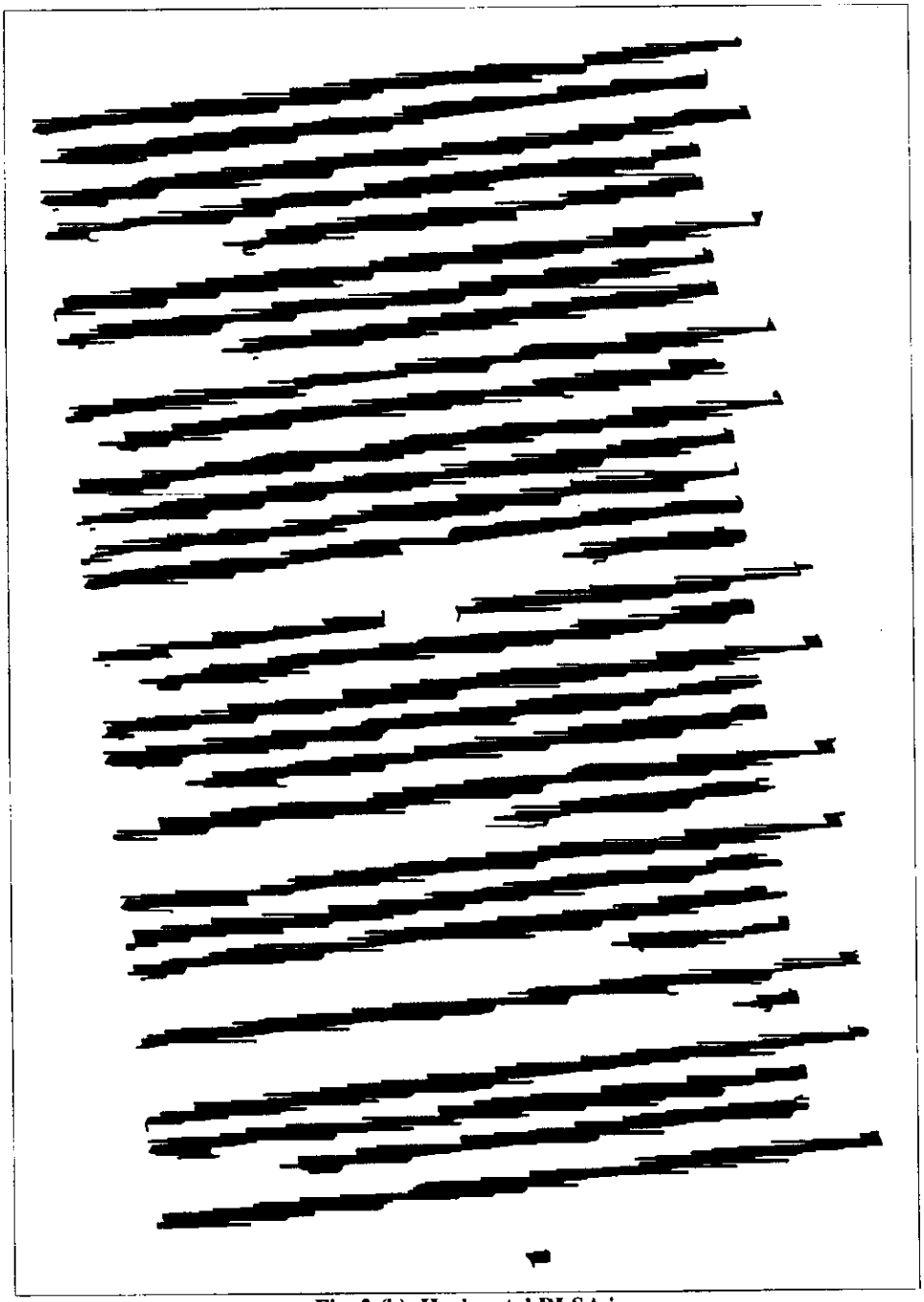


Fig. 2 (b). Horizontal RLSA image.

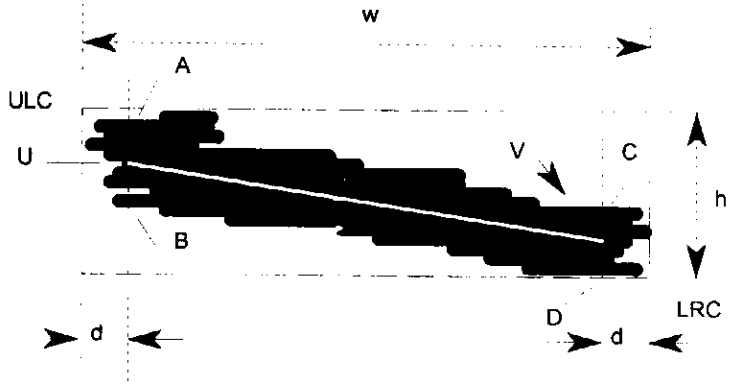


Fig. 2. RLSA block of one skewed line of text.

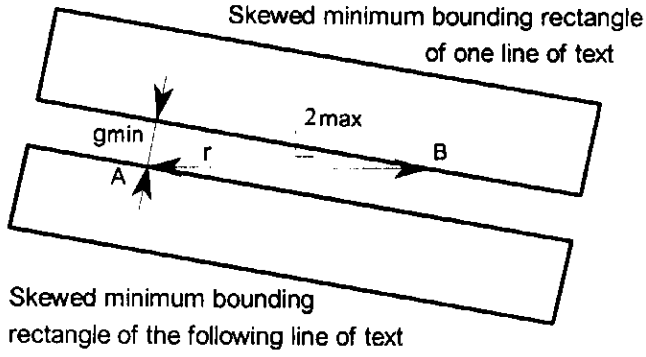


Fig. 3. Skewed minimum bounding rectangles of two adjacent lines of text.

connecting A and B, the length of which is r, is horizontal. If the maximum expected skew angle of the input document is θ_{max}, then a condition for the RLSA blocks of these lines not to merge into one block is $r < g_{min} / \sin(\theta_{max})$.

2. The distance d should not be equal to zero for better accuracy to increase the chance of having the points A, B, and C, D lie on the upper and lower curved sides, and not on the left and right curved sides, of the RLSA block, respectively. This results in U and V being in the middle of the block, and hence, a more accurate representative straight line, UV, is obtained.

Our algorithm for skew correction assumes that the following parameters are predefined:

1. Maximum expected skew angle, θ_{max}, in either clockwise or counterclockwise directions,
2. Minimum expected gap, g_{min},
3. Minimum width, w_{min}, of the bounding rectangle of a smeared black block,
4. Minimum width to height ratio, α_{min}, of the bounding rectangle of a smeared black block,
5. The distance d, see Fig. 2, and
6. Maximum allowed skew angle, $\theta_{allowed}$, in the output image, I_c,

See Section 4 on how to predefine these parameters to make the algorithm, accurate, robust, and flexible to various conditions. According to the earlier discussion, the overall basic algorithm for skew angle detection and correction is formally described in the appendix. The basic algorithm is applied to the image of Fig. 1(a) to obtain the corrected image of Fig. 4.

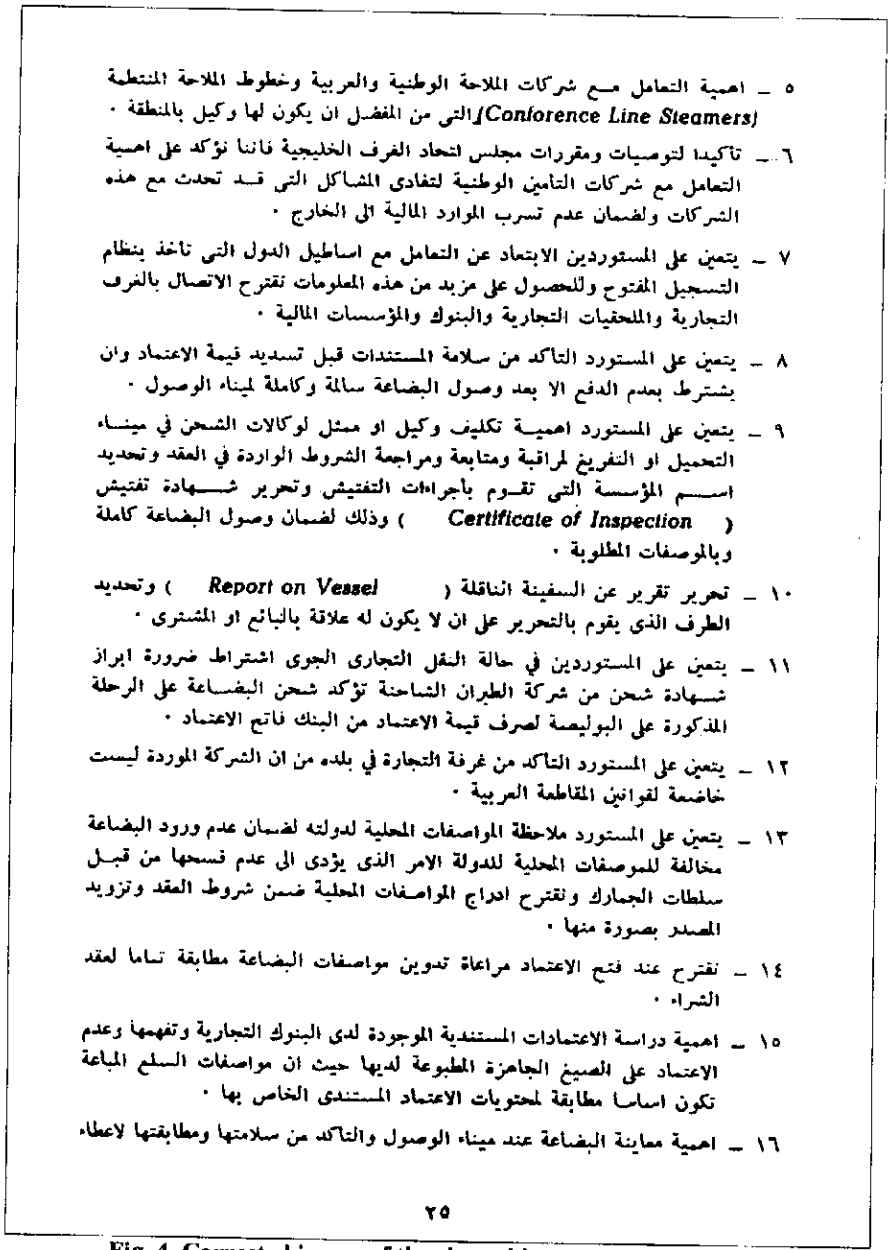
3. Multi-pass Skew Correction

The basic algorithm was designed to handle documents with the maximum expected skew angle $\theta_{max} = 10°$. For larger skew angles, it gives inaccurate results. Fig. 5(a) shows an image which has a measured skew of 16.70° before correction. Fig. 5(b) shows the corresponding RLSA image. Note that most of the lines are connected together due to large initial skew. The only smeared block which satisfies the conditions of Step 3 of the basic algorithm is the uppermost block. This block consists of two text lines. Since it is the only satisfying block, the angle of its UV line, 14.53°, determines the angle of the whole document, which is the source of the big error, 2.17°. Fig. 5(c) shows the corrected image the measured and calculated skews of which are 2.15° and 2.22°, respectively.

The above problem can be overcome by implementing a multi-pass skew correction algorithm where the skew is iteratively corrected until a certain criteria is satisfied. The same parameters predefined for the basic algorithm are assumed to be predefined for the multi-pass algorithm. In addition, the maximum number of passes, N_{max}, to correct skew is also predefined. See Section 4 on how to predefine these parameters to make these algorithms accurate, robust, and flexible to various conditions.

The multi-pass algorithm was applied to the image in Fig. 5(a). The images of the first and second passes are shown in Figs.5(c,d), respectively. For the image in Fig. 5(d),

the measured and calculated skews are –0.36° and 0.03°, respectively which is adequately accurate to segment the textual lines with white cuts.



٥ ـ اهمية التعامل مـع شركات الملاحة الوطنية والعربية وخطوط الملاحة المنتظمة (Conforence Line Steamers) التى من المفضل ان يكون لها وكيل بالمنطقة ٠

٦ ـ تاكيدا لتوصيات ومقررات مجلس اتحاد الغرف الخليجية فاننا نؤكد على اهمية التعامل مع شركات التامين الوطنية لتفادى المشاكل التى قـد تحدث مع هذه الشركات ولضمان عدم تسرب الموارد المالية الى الخارج ٠

٧ ـ يتعين على المستوردين الابتعاد عن التعامل مع اساطيل الدول التى تاخذ بنظام التسجيل المفتوح وللحصول على مزيد من هذه المعلومات نقترح الاتصال بالغرف التجارية والملحقيات التجارية والبنوك والمؤسسات المالية ٠

٨ ـ يتعين على المستورد التاكد من سلامة المستندات قبل تسديد قيمة الاعتماد وان يشترط بعدم الدفع الا بعد وصول البضاعة سالمة وكاملة لميناء الوصول ٠

٩ ـ يتعين على المستورد اهميـة تكليف وكيل او ممثل لوكالات الشحن في ميناء التحميل او التفريغ لمراقبة ومتابعة ومراجعة الشروط الواردة في العقد وتحديد اسـم المؤسسة التى تقـوم باجراءات التفتيش وتحرير شـهادة تفتيش (Certificate of Inspection) وذلك لضمان وصول البضاعة كاملة وبالموصفات المطلوبة ٠

١٠ ـ تحرير تقرير عن السفينة الناقلة (Report on Vessel) وتحديد الطرف الذى يقوم بالتحرير على ان لا يكون له علاقة بالبائع او المشترى ٠

١١ ـ يتعين على المستوردين في حالة النقل التجارى الجوى اشتراط ضرورة ابراز شـهادة شحن من شركة الطيران الشاحنة تؤكد شحن البضـاعة على الرحلة المذكورة على البوليصة لصرف قيمة الاعتماد من البنك فاتح الاعتماد ٠

١٢ ـ يتعين على المستورد التاكد من غرفة التجارة في بلده من ان الشركة الموردة ليست خاضعة لقوانين المقاطعة العربية ٠

١٣ ـ يتعين على المستورد ملاحظة المواصفات المحلية لدولته لضمان عدم ورود البضاعة مخالفة للموصفات المحلية للدولة الامر الذى يؤدى الى عدم فسحها من قبـل سلطات الجمارك ونقترح ادراج المواصـفات المحلية ضمن شروط العقد وتزويد المصدر بصورة منها ٠

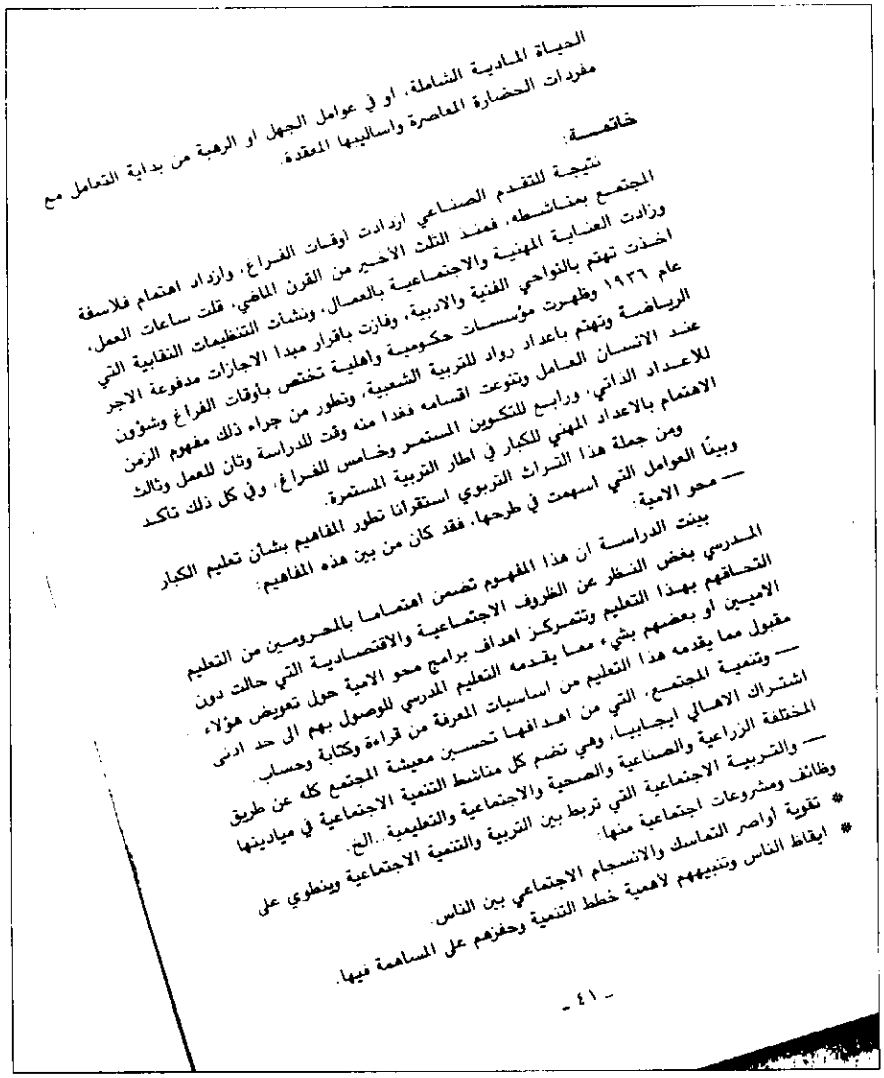
١٤ ـ نقترح عند فتح الاعتماد مراعاة تدوين مواصفات البضاعة مطابقة تماما لعقد الشراء ٠

١٥ ـ اهمية دراسة الاعتمادات المستندية الموجودة لدى البنوك التجارية وتفهمها وعدم الاعتماد على الصيغ الجاهزة المطبوعة لديها حيث ان مواصفات السلع المباعة تكون اساسا مطابقة لمحتويات الاعتماد المستندى الخاص بها ٠

١٦ ـ اهمية معاينة البضاعة عند ميناء الوصول والتاكد من سلامتها ومطابقتها لاعطاء

٢٥

Fig. 4. Corrected image of the skewed image shown in Fig. 1(a).



الحياة المادية الشاملة، او في عوامل الجهل او الرهبة من بداية التعامل مع
مفردات الحضارة المعاصرة واساليبها المعقدة.

خاتمة:

نتيجة للتقدم الصناعي ازدادت اوقات الفراغ، وازداد اهتمام فلاسفة
المجتمع بعناشطه، فعند الثلث الاخير من القرن الماضي، قلت ساعات العمل،
وزادت العناية المهنية والاجتماعية بالعمال، ونشأت التنظيمات النقابية التي
اخذت تهتم بالنواحي الفنية والادبية، وقارنت باقرار مبدا الاجازات مدفوعة الاجر
عام ١٩٣٦ وظهرت مؤسسات حكومية واهلية تختص بأوقات الفراغ وشؤون
الرياضة وتهتم باعداد رواد للتربية الشعبية، وتطور من جراء ذلك مفهوم الزمن
عند الانسان العامل وتنوعت اقسامه فغدا منه وقت للدراسة وثان للعمل وثالث
للاعداد الذاتي، ورابع للتكوين المستمر وخامس للفراغ، وفي كل ذلك تاكد
الاهتمام بالاعداد المهني للكبار في اطار التربية المستمرة.

ومن جملة هذا التراث التربوي استقرانا تطور المفاهيم بشأن تعليم الكبار
وبينا العوامل التي اسهمت في طرحها. فقد كان من بين هذه المفاهيم:

— محو الامية:

بينت الدراسة ان هذا المفهوم تضمن اهتماما بالمحرومين من التعليم
المدرسي بغض النظر عن الظروف الاجتماعية والاقتصادية التي حالت دون
التحاقهم بهذا التعليم وتتمركز اهداف برامج محو الامية حول تعويض هؤلاء
الاميين او بعضهم بشيء مما يقدمه التعليم المدرسي للوصول بهم الى حد ادنى
مقبول مما يقدمه هذا التعليم من اساسيات المعرفة من قراءة وكتابة وحساب.

— وتنمية المجتمع: التي من اهدافها تحسين معيشة المجتمع كله عن طريق
اشتراك الاهالي ايجابيا، وهي تضم كل مناشط التنمية الاجتماعية في ميادينها
المختلفة الزراعية والصناعية والصحية والاجتماعية والتعليمية..الخ.

— والتربية الاجتماعية: وظائف ومشروعات اجتماعية التي تربط بين التربية والتنمية الاجتماعية وينطوي على

* تقوية أواصر التماسك والانسجام الاجتماعي بين الناس.
* ايقاظ الناس وتنبيههم لاهمية خطط التنمية وحفزهم على المساهمة فيها.

- ٤١ -

Fig. 5 (a). Binary image of skewed textual document, (cont.).

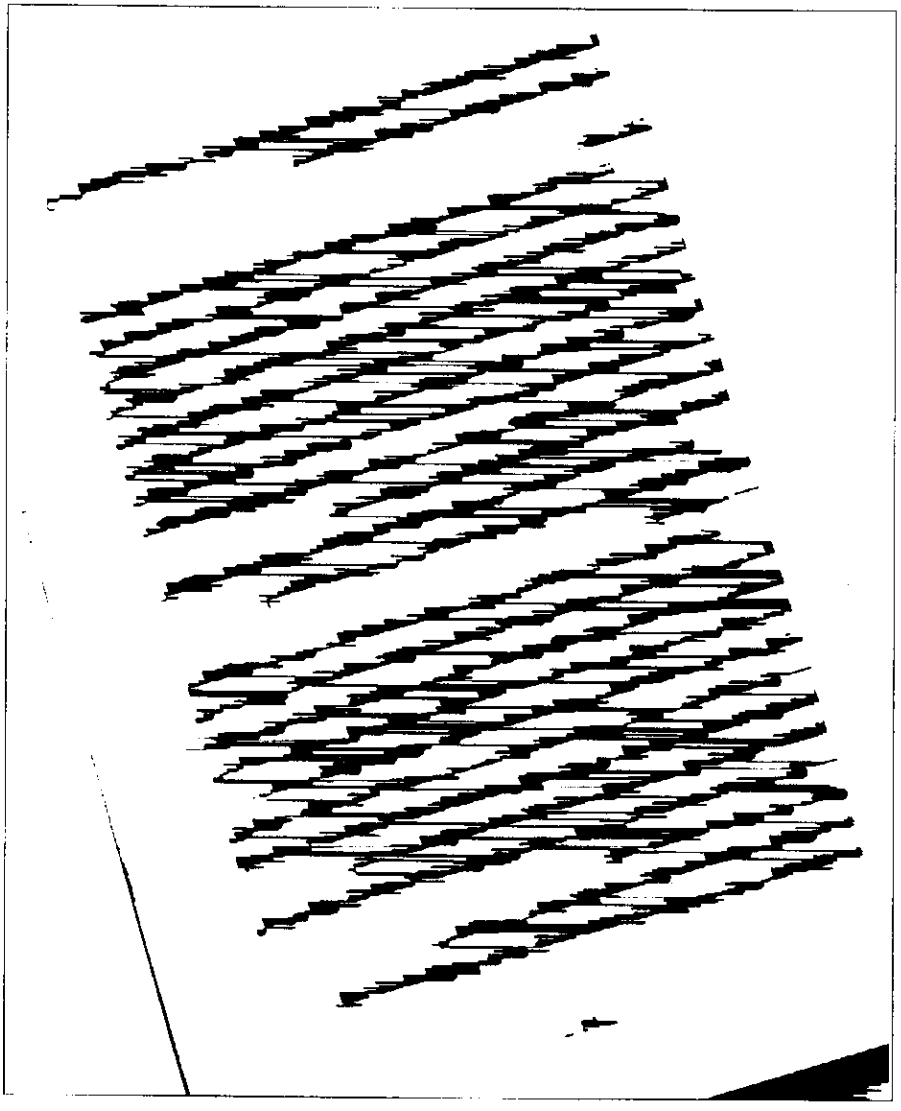
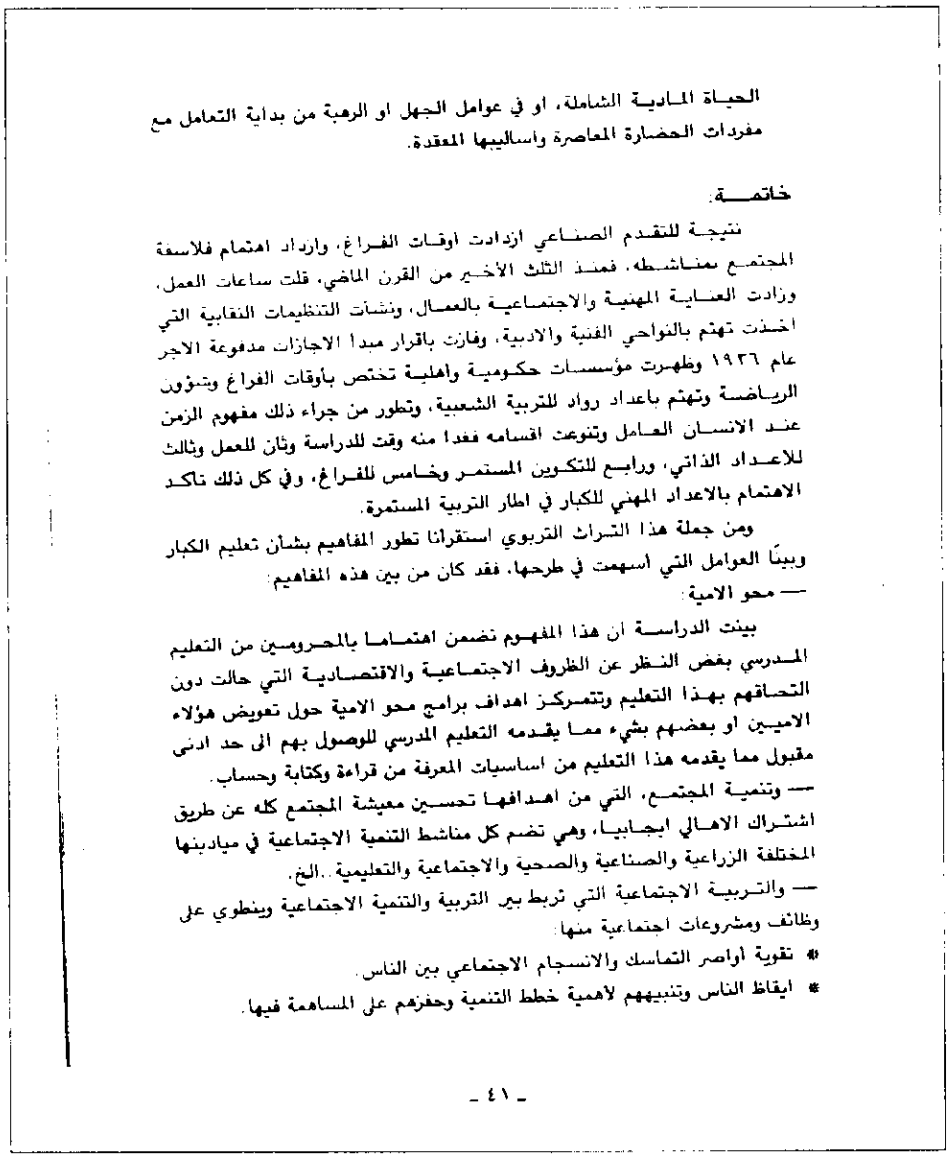


Fig. 5(b). Horizontal RLSA image, (cont.).



الحياة المادية الشاملة، او في عوامل الجهل او الرهبة من بداية التعامل مع مفردات الحضارة المعاصرة واساليبها المعقدة.

خاتمة:

نتيجة للتقدم الصناعي ازدادت اوقات الفراغ، وازداد اهتمام فلاسفة المجتمع بمناشطه. فمنذ الثلث الاخير من القرن الماضي، قلت ساعات العمل، وزادت العناية المهنية والاجتماعية بالعمال، ونشأت التنظيمات النقابية التي اخذت تهتم بالنواحي الفنية والادبية، وفازت باقرار مبدا الاجازات مدفوعة الاجر عام ١٩٣٦ وظهرت مؤسسات حكومية واهلية تختص بأوقات الفراغ وتشؤون الرياضة وتهتم باعداد رواد للتربية الشعبية، وتطور من جراء ذلك مفهوم الزمن عند الانسان العامل وتنوعت اقسامه فغدا منه وقت للدراسة وثان للعمل وثالث للاعداد الذاتي، ورابع للتكوين المستمر وخامس للفراغ، وفي كل ذلك تاكد الاهتمام بالاعداد المهني للكبار في اطار التربية المستمرة.

ومن جملة هذا التراث التربوي استقرانا تطور المفاهيم بشأن تعليم الكبار وببنّا العوامل التي اسهمت في طرحها. فقد كان من بين هذه المفاهيم:

— محو الامية:

بينت الدراسة ان هذا المفهوم تضمن اهتماما بالمحرومين من التعليم المدرسي بغض النظر عن الظروف الاجتماعية والاقتصادية التي حالت دون التحاقهم بهذا التعليم وتتمركز اهداف برامج محو الامية حول تعويض هؤلاء الاميين او بعضهم بشيء مما يقدمه التعليم المدرسي للوصول بهم الى حد ادنى مقبول مما يقدمه هذا التعليم من اساسيات المعرفة من قراءة وكتابة وحساب.

— وتنمية المجتمع، التي من اهدافها تحسين معيشة المجتمع كله عن طريق اشتراك الاهالي ايجابيا، وهي تضم كل مناشط التنمية الاجتماعية في ميادينها المختلفة الزراعية والصناعية والصحية والاجتماعية والتعليمية..الخ.

— والتربية الاجتماعية التي تربط بين التربية والتنمية الاجتماعية وينطوي على وظائف ومشروعات اجتماعية منها:

* تقوية أواصر التماسك والانسجام الاجتماعي بين الناس.
* ايقاظ الناس وتنبيههم لاهمية خطط التنمية وحفزهم على المساهمة فيها.

- ٤١ -

Fig. 5(c). Corrected image using basic algorithm and first pass of the multi-pass algorithm, (cont.).

الحياة المادية الشاملة، او في عوامل الجهل او الرهبة من بداية التعامل مع مفردات الحضارة المعاصرة واساليبها المعقدة.

خاتمة:

نتيجة للتقدم الصناعي ازدادت اوقات الفراغ، وازداد اهتمام فلاسفة المجتمع بمناشطه، فمنذ الثلث الاخير من القرن الماضي، قلت ساعات العمل، وزادت العناية المهنية والاجتماعية بالعمال، ونشأت التنظيمات النقابية التي اخذت تهتم بالنواحي الفنية والادبية. وفازت باقرار مبدأ الاجازات مدفوعة الاجر عام ١٩٣٦ وظهرت مؤسسات حكومية واهلية تختص بأوقات الفراغ وتشؤون الرياضة وتهتم باعداد رواد للتربية الشعبية، وتطور من جراء ذلك مفهوم الزمن عند الانسان العامل وتنوعت اقسامه فغدا منه وقت للدراسة وثان للعمل وثالث للاعداد الذاتي، ورابع للتكوين المستمر وخامس للفراغ، وفي كل ذلك تأكد الاهتمام بالاعداد المهني للكبار في اطار التربية المستمرة.

ومن جملة هذا التراث التربوي استقرانا تطور المفاهيم بشأن تعليم الكبار وبينّا العوامل التي اسهمت في طرحها، فقد كان من بين هذه المفاهيم:

— محو الامية:

بينت الدراسة ان هذا المفهوم تضمن اهتماما بالمحرومين من التعليم المدرسي بغض النظر عن الظروف الاجتماعية والاقتصادية التي حالت دون التحاقهم بهذا التعليم وتتمركز اهداف برامج محو الامية حول تعويض هؤلاء الاميين او بعضهم بشيء مما يقدمه التعليم المدرسي للوصول بهم الى حد ادنى مقبول مما يقدمه هذا التعليم من اساسيات المعرفة من قراءة وكتابة وحساب.

— وتنمية المجتمع، التي من اهدافها تحسين معيشة المجتمع كله عن طريق اشتراك الاهالي ايجابيا، وهي تضم كل مناشط التنمية الاجتماعية في ميادينها المختلفة الزراعية والصناعية والصحية والاجتماعية والتعليمية..الخ.

— والتربية الاجتماعية التي تربط بين التربية والتنمية الاجتماعية وينطوي على وظائف ومشروعات اجتماعية منها:

* تقوية اواصر التماسك والانسجام الاجتماعي بين الناس.
* ايقاظ الناس وتنبيههم لاهمية خطط التنمية وحفزهم على المساهمة فيها.

Fig. 5(d). Final corrected image using the multi-pass algorithm.

4. Experiments

To make the algorithms accurate, robust, and flexible to various conditions, we predefine their parameters as follows:

1) The distance d, see Fig. 2, should not be equal to zero to increase the chance of having the points A, B, and C, D lie on the upper and lower curved sides, and not on the left and right curved sides, of the RLSA block, respectively. This results in U and V being in the middle of the block, and hence, a more accurate representative straight line, UV, is obtained. Empirically, d was found to be 3 mm.
2) When a textual document is smeared using the RLSA, blocks the width of which does not fall below 7.5 cm can be easily found. Blocks with smaller width decrease the accuracy and are not allowed to share in skew estimation. Thus, we set $w_{min} = 7.5$ cm.
3) The more elongated the minimum bounding rectangles of the RLSA blocks are the more accurate the skew angle is. How much a rectangle is elongated can be measured by its width to height ratio, α. Empirically, blocks with height to width ratio not less than $\alpha_{min} = 3.0$ are considered long enough to share in skew estimation.
4) By setting the allowed skew angle, $\theta_{allowed}$, in the output image, I_c, to 0.5°, the vertical gab, g, between the skewed minimum bounding rectangles of two adjacent lines is $\sin(\theta_{allowed} = 0.5°)$ times the page width. This width is typically 21 cm for an A4 page which results in $g = 21 \times \sin(0.5°) = 1.83$ mm. Thus we take $g_{min} = 2$ mm, which is sufficient to segment lines with white cuts and agrees with interline spacing even with very small fonts.
5) RLSA blocks with minimum bounding rectangles of high width to height ratio increase the accuracy of the algorithms, which can be achieved by increasing the run-length, r. However, r can't be increased without a limit since this results in the merge of blocks of adjacent lines degrading accuracy. A suitable value of r was found to be 10 mm which guarantees that the blocks of two adjacent lines don't merge if the maximum expected skew angle, θ_{max}, does not exceed $\sin^{-1}(g_{min} / r) = \sin^{-1}(2 / 10) = 11.5°$. Thus, we set $\theta_{max} = 10°$.
6) In the multi-pass algorithm, the parameter, N_{max}, was set to 3 passes which is sufficient to correct the skew.

In summary, the parameters are predefined as follows: $d = 3$ mm, $w_{min} = 7.5$ cm, $\alpha_{min} = 3.0$, $\theta_{allowed} = 0.5°$, $g_{min} = 2$ mm, $\theta_{max} = 10°$, and $N_{max} = 3$ passes.

The basic algorithm was tested against 50 pages of printed text. Fifty books were randomly selected. Then, one page was randomly selected and copied from each page. Binary images of copied pages, which were manually skewed, were captured using a

scanner with a 300 dpi resolution. The approximate skew for the first 45 pages was in the range ±10° and for the last five pages it was outside this range.

Table 1 displays the result of our test of the basic algorithm. The measured skew, columns 2 and 5, of Table 1, is found manually. The calculated skew, column 3, is found using the basic algorithm (Steps 1 to 3). The skew error before correction is the difference between the measured and calculated skews before correction. The measured skew after correction, column 5, is the measured skew of the output image, I_c.

For the first 45 images in Table 1 (skew is within ±10°), the following figures can be drawn. The maximum absolute skew error is 1.40°. The average absolute skew error is 0.24° which is very small and shows the high accuracy of the basic algorithm for skew angle calculation.

For the last five images, Nos. 46 to 50, the measured and calculated skews were greater than 10°. The basic algorithm could successfully correct pages 46 and 47. However, for the last three images, Nos. 48 to 50, the measured skew was –5.00°, 2.15°, and 2.15°, respectively, which is not adequate to segment the lines with white cuts.

The multi-pass algorithm was tested against the last five images, pages 46 to 50. Table 2 shows detailed results of this test. It is noticed that for pages 46 and 47 only one pass of rotation was sufficient to correct the document. For the pages 48 to 50, two passes were needed.

The algorithms were implemented using Visual C++ 6.0 on a Pentium III compatible PC, with 866MHz clock and 128Mbytes RAM. To get an idea of how much processing time is needed, the multi-pass algorithm was tested on ten binary images of textual documents (Algorithm parameters were the same as earlier except that N_{max} was set to one pass). All images were 6×8 inches^2, (1800×2400 pixels^2). Table 3 illustrates the results indicating an average processing time of 4.05 seconds. The algorithms will run faster on more recent computers and using some programming and code optimisation techniques.

5. Conclusion

Two algorithms for accurate skew detection and correction of textual documents were presented. They rely on finding a horizontal RLSA image of the skewed document. The average skew of selected black connected components is considered as the skew angle for the whole document which is finally rotated in the opposite direction by that amount to obtain the final corrected image.

Table 1. Results of basic algorithm

Page No.	Skew before correction (degrees)			Measured skew after correction (degrees)
	Measure	Calculated	Error	
1	–4.64	–4.45	–0.19	–0.19
2	5.44	5.55	–0.11	–0.39
3	6.92	6.79	0.13	0.00
4	6.96	7.35	–0.39	–0.45
5	–4.50	–4.35	–0.15	–0.23
6	7.35	7.16	0.19	0.37
7	–5.01	–5.17	0.16	0.00
8	5.19	5.39	–0.20	–0.19
9	–5.60	–5.56	–0.04	–0.19
10	5.86	5.67	0.19	0.19
11	6.62	6.82	–0.20	–0.18
12	–7.42	–7.36	–0.06	–0.20
13	6.53	6.08	0.45	0.00
14	–6.40	–6.24	–0.16	–0.39
15	5.36	5.15	0.21	0.00
16	–4.66	–4.35	–0.31	–0.20
17	4.73	4.60	0.13	–0.22
18	–6.00	–6.13	0.13	0.00
19	–8.88	–8.92	0.04	–0.18
20	–4.64	–4.37	–0.27	–0.39
21	6.84	6.71	0.13	–0.38
22	–7.35	–7.14	–0.21	–0.18
23	6.25	6.50	–0.25	0.00
24	–8.26	–8.04	–0.22	0.00
25	7.96	7.95	0.01	0.20

Error = Measured skew before correction – calculated skew before correction

[Contd. ...

Table 1. Results of basic algorithm (contd.)

Page No.	Skew before correction (degrees)			Measured skew after correction (degrees)
	Measured	Calculated	Error	
26	–8.07	–7.58	–0.49	–0.20
27	7.08	7.20	–0.12	–0.19
28	–9.83	–11.23	1.40	1.53
29	7.13	7.03	0.10	0.00
30	–7.65	–7.13	–0.52	–0.38
31	6.77	6.92	–0.15	–0.18
32	–7.17	–6.89	–0.28	0.00
33	7.62	7.66	–0.04	0.00
34	–9.93	–9.66	–0.27	–0.18
35	7.33	7.20	0.13	0.19
36	–8.44	–8.18	–0.26	–0.18
37	6.71	6.68	0.03	0.00
38	–7.03	–6.77	–0.26	0.00
39	7.08	6.96	0.12	0.00
40	–8.46	–8.26	–0.20	–0.19
41	6.55	6.49	0.06	0.19
42	–7.50	–7.23	–0.27	0.00
43	6.34	5.99	0.35	0.00
44	–7.91	–7.81	–0.10	0.00
45	7.90	6.75	1.15	1.05
46	–12.00	–11.45	–0.55	–0.18
47	–12.17	–12.01	–0.16	0.00
48	–15.21	–10.10	–5.11	–5.00
49	15.71	13.43	2.28	2.15
50	16.70	14.53	2.17	2.15

Error = measured skew before correction – calculated skew before correction

Table 2. Results of multi-pass algorithm

Page No.	SBC		SAFC		SASC		FMSAC
	M	C	M	C	M	C	
46	–12.00	–11.45	–0.18	–0.13	—	—	–0.18
47	–12.17	–12.01	0.00	0.14	—	—	0.00
48	–15.21	–10.10	–5.00	–5.07	0.00	–0.05	0.00
49	15.71	13.43	2.15	2.17	0.00	0.07	0.00
50	16.7	14.53	2.15	2.22	–0.36	0.03	–0.36

SBC: skew before correction, SAFC: skew after first correction, SASC: skew after second Correction, FMSAC: final measured skew after correction; in degrees, M: measured, C: calculated.

Table 3. Processing time of the multi-pass algorithm

Page No.	Calculated skew before correction (degrees)	Calculated skew after correction (degrees)	Processing time (seconds)
1	–8.47	–1.39	3.29
2	–7.72	–0.17	3.40
3	–5.76	–0.12	4.40
4	–3.85	–0.03	4.18
5	–1.88	–0.01	5.39
6	2.08	0.03	4.66
7	4.06	0.05	4.78
8	5.91	0.21	3.71
9	8.30	–0.16	3.57
10	10.85	–0.71	3.11

The algorithm was implemented using Visual C++ 6.0 on a Pentium III compatible PC, with 866MHz clock and 128Mbytes RAM.

The basic algorithm is a single-pass skew detection and correction algorithm. It gives an average absolute measured skew after correction equal to 0.20° when the real skew angles are within ±10° and the interline spacing is less than 1mm.

The second algorithm is a multi-pass skew detection and correction algorithm where the angle is iteratively computed and corrected. It can be used for correcting skew angles that are not within ±10°. In our experiments, the measured skew after correction did not exceed 0.36°. We emphasize again that our algorithms have the following advantages over others:

- The multi-pass algorithm is not limited, in contrast to methods using the projection histogram [2-5], to documents that have fairly small skews that are typically less than ±10°.
- Peak-finding problems do not exist in our algorithms as compared to the Hough transform-based methods [7-12].
- Noise, subparts of characters (dot on "i"), and between-line connections can reduce the accuracy of the methods based on nearest-neighbour clustering [15, 16]. Such problems don't exist in our algorithms.

The average processing time of the multi-pass algorithm, when run on an 866MHz Pentium III PC, is 4.05 seconds for 1800 × 2400 pixels^2 images. The algorithms will run faster on more recent computers and using some programming and code optimisation techniques.

References

[1] Wang, D. and Srihari, S.N. "Classification of Newspaper Image Blocks Using Texture Analysis". *Computer Vision, Graphics, and Image Processing*, 47 (1989), 327-352.

[2] Ciardiello, G., Scafuro, G., Degrandi, M.T., Spada, M.R. and Roccotelli, M.P. "An Experimental System for Office Document Handling and Text Recognition". *Proc. 9th Int. Conf. on Pattern Recognition*, 1988, 739-743.

[3] Postl, W. "Detection of Linear Oblique Structures and Skew Scan in Digitized Documents". *Proc. 8th Int. Conf. on Pattern Recognition*, Paris, France, 1986, 687-689.

[4] Akiyama, T. and Hagita, N. "Automated Entry System for Printed Documents". *Pattern Recognition*, 23 No.11, (1990), 1141-1153.

[5] Pavlidis, T. and Zhou, J. "Page Segmentation by White Streams". *Proc. 1st Int. Conf. Document Anal. Recogn. (ICDAR)*, St. Malo, France, 1991, 945-953.

[6] Dengel, A. ANASTASIL: "A System for Low-level and High-level Geometric Analysis of Printed Documents". In: Baird, H.S. Bunke, H. and Yamamoto, Y. (Eds.). *Structured Document Image Analysis*, (Berlin: Springer-Verlag , 1992, 70-98.

[7] Baird, H.S. "The Skew Angle of Printed Documents". *Proc. SPSE 40th Symp. Hybrid Imaging Systems*, Rochester, NY, 1987, 21-24.

[8] Srihari, S.N. and Govindaraju, V. "Analysis of Textual Images Using the Hough Transform". *Mach. Vision Appl.*, 3, 1989, 141-153.

[9] Nakano, Y., Shima, Y., Fujisawa, H., Higashino, J. and Fujinawa, M. "An Algorithm for the Skew Normalization of Document Image." *Proc.10th Int. Conf. on Pattern Recognition*. Atlantic City, NJ: 1990, 8-11.

[10] Le, D.S., Thoma, G.R. and Weschler, H. "Automated Page Orientation and Skew Angle Detection for Binary Document Images". *Pattern Recognition*, 27 No.10 (1994), 1325-1344.

[11] Schurmann, J., Bartneck, N., Bayer, T., Franke, J., Mandler, E. and Oberlander, M. "Document Analysis – from Pixels to Contents". *Proc. IEEE* 80, 1992, 1101-1119.

[12] Hinds, S.C., Fisher, J.L. and D'Amato, D.P. "A Document Skew Detection Method Using Run-length Encoding and the Hough Transform". *Proc. 10th Int. Conf. Pattern Recognition (ICPR)*, NJ, Atlantic City, 1990, 464-468.

[13] Chen, S., Hardick, R. M. and Phillips, I. T. "Automatic Text Skew Estimation in Document Images". *Proc. 3rd ICDAR*, Montreal, Canada, 1995, 1153-1156.

[14] Yan, H. "Skew Correction of Document Images Using Interline Cross-Correlation". *CVGIP: Graphical Models and Image Processing*, 55 No. 6, (1993), 538-543.
[15] Hashizume, A., Yeh, P.S. and Rosenfeld, A. "A Method for Detecting the Orientation of Aligned Components, *Pattern Recognition Letters*, 4, No.2 (1986), 125-132.
[16] O'Gorman, L. "The Document Spectrum for Page Layout Analysis". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15, No.11 (1993), 1162-1173.
[17] Chaudhuri, B. B. and Pal, U. "Skew Angle Detection of Digitized Indian Script Documents". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19, No.2 (1997), 182-186.
[18] Wong, K.Y., Casey, R.G. and Wahl, F.M. "Document Analysis System, *IBM J. Res. Develop.*, 26, No.6 (1982), 647-656.
[19] Danielsson, Per-Erik. An Improved Segmentation and Coding Algorithm for Binary and Nonbinary Images". *IBM J. Res. Develop.*, 26, No.6 (1982), 698-707.

Appendix

Basic Algorithm

Use: Skew angle detection and correction

Input: Binary image, I_s, of skewed document

Output: Binary image, I_c, of corrected document

Procedure:

Step 1. Let the run length r be equal to g_{min} / sin (θ_{max}). Apply the RLSA in the horizontal direction to the image, I_s, to obtain the RLSA image, I_r.

Step 2. Extract the connected components, i.e., the smeared black blocks, of the image I_r. A segmentation algorithm for this purpose can be found in [19]. For each block, record the x and y coordinates of ULC and LRC. Find the width, w, and height, h, of each block where $w = x_{LRC} - x_{ULC}$, and $h = y_{ULC} - y_{LRC}$.

Step 3. (a) Initialize the variables: SumOfAngles = 0, Counter = 0. For each segmented block which satisfies the following conditions: (1) $w \geq w_{min}$, and (2) $w / h \geq \alpha_{min}$:

i- Use the ULC and LRC points and the distance d to find the points U and V, see Fig. 2.

ii- Find the angle, θ, of the straight line which connects the points U and V.

iii- Let SumOfAngles = SumOfAngles + θ, and increase Counter.

(b) If Counter > 0 then the skew angle, θ_s, of the document equals SumOfAngles / Counter. Otherwise, $\theta_s = 0$.

Step 4. If $|\theta_s| > \theta_{allowed}$ then rotate the image I_s by the amount $-\theta_s$ to get the output image I_c. Otherwise, $I_c = I_s$.

تصحيح الانحراف في المستندات النصّية

إبراهيم سليمان إبراهيم أبو هيبة
قسم هندسة الكهرباء والحاسوب، الجامعة الإسلامية،
ص ب ١٢٧٦، غزة، فلسطين

(قدّم للنشر في ٢٠٠٢/٠١/٠١م؛ وقبل للنشر في ٢٠٠٢/٠٢/٣٠م)

ملخص البحث. نقدم خوارزميتين للاكتشاف الدقيق للانحراف و تصحيحه في المستندات النصّية، و يعتمدان على حساب صورة RLSA أفقية. نعتبر متوسط انحراف أجزاء متصلة سوداء مختارة في صورة RLSA زاوية الانحراف لكل المستند الذي يدار في الاتجاه المعاكس بمقدار تلك الزاوية لنحصل على الصورة المصححة. في الخوارزمية الأولى إما أن تدار الصورة مرة واحدة فقط أو لا تدار، وفي الخوارزمية الثانية إما أن تدار الصورة مرة واحدة على الأقل أو لا تدار. تم استخدام خمسين صفحة لفحص الخوارزميتين. بلغ الحد الأقصى للخطأ المقاس ١,٥٣° و ٠,٣٦° في الخوارزميتين الأولى و الثانية، على الترتيب.