Estimating the skew angle of scanned document through background area information

Angélica A. Mascaro and George D. C. Cavalcanti

Center of Informatics, Federal University of Pernambuco
AiLeader Technologies
Recife, PE, Brazil
{aam3,gdcc}@cin.ufpe.br

Abstract

The skew correction of scanned document images is an important step toward an automatic recognition system. Several techniques have been developed to estimate the skew angle of a scanned document. However, most of these techniques have the problem of expensive computational costs. A variation of a fast method based on parallelograms covering is presented in this article. The objective is to obtain lower computation time and provide a way to work well over noisy images and also over different layouts containing non-text areas with no decrease in performance. Experimental study with different databases achieved results that overcome previous techniques.

1. Introduction

In document automatic recognition systems, the quality of the input image is crucial to the final performance. There are a variety of interfering effects such as noise and skewing that appear during the scanning process. These components can damage the image and decrease the performance of the recognizer.

Skew correction plays an important role in the image preprocessing. A small inclination in the document image can interfere in the layout analysis and consequently in the rest of the process.

To solve this problem, a large amount of techniques were developed to estimate the skew angle of scanned document images. A part of it is based on Hough's Transform [1], where occurs a change in the coordinates to find the skew angle. The problem with the Hough Transform is its expensive cost and its high sensitive to noise and non-text areas.

Ishitani proposed a skew detection method based on maximum variance of transition-counts [2] to deal with images containing a mixture of text areas, photographs, figures, charts, and tables.

There are also the projection based approaches [3], which is commonly used. Other techniques were developed based on cross-correlations [4]. Great part of these solutions (like the Hough's Transform and projection based approach) suffers with high computational costs. In contrast with this, the automatic recognition systems need low time consumption.

The present proposal is based on a variation of the piecewise covering method for skew angle estimation proposed by Chou *et al.* [5]. This method follows the idea that a document is composed by a combination of rectangular objects, such as text lines, forms, figures, tables, etc. The main idea of the method is to construct parallelograms at all angles and decide the one which best fits the objects in the image. Our objective is to show a variation of the Chou's method changing the criterion to evaluate the best angle – now based on the background area. With this variation it was possible to develop an effective way to reduce the search space for the skew angle – instead of looking at all angles. We also propose a variation that works well with noisy images and different layouts including non-text areas.

In the next sections we describe details about the proposed approach. Section 2 presents the Chou's original method and our proposal. Section 3 shows an experimental study with synthetic rotated and real scanned document images. In Section 4 we discuss the limitations of the method and expose some conclusions.

2. Skew angle estimation method

The parallelograms covering method proposed by Chou *et al*. [5] to estimate the skew angle of a scanned document is based on drawing scan lines over the document at various angles. The skew angle is estimated by constructing parallelograms with these scan lines that cover objects in the image.

The process starts by drawing the parallel lines at a certain angle θº. A scan line is a row with 1 pixel width that crosses the image horizontally. Then the image is vertically divided into regions with fixed size, called *slabs* – see Figure 1a. After, the lines are divided into as many *sections* as the number of slabs. Chou suggested using 450 pixels as the size of the slab. As expected, the last slab in the right will be smaller if the width of the document is not a multiple of 450 pixels.

In the parallelograms construction phase, each section of the scan lines is examined by a certain angle. If this section contains at least one black pixel, it is turned to gray; otherwise, it stays white. Adjacent gray sections form parallelograms. Figure 1a shows a document skewed at -6º and Figure 1b shows the parallelograms constructed with the scan lines at -6º.

After that, the scan lines are skewed at various angles and the size of the region covered by parallelograms is calculated. As exposed in Figure 1a-c, when the lines are drawn in the same angle of the document inclination, the number of white sections is the largest one.

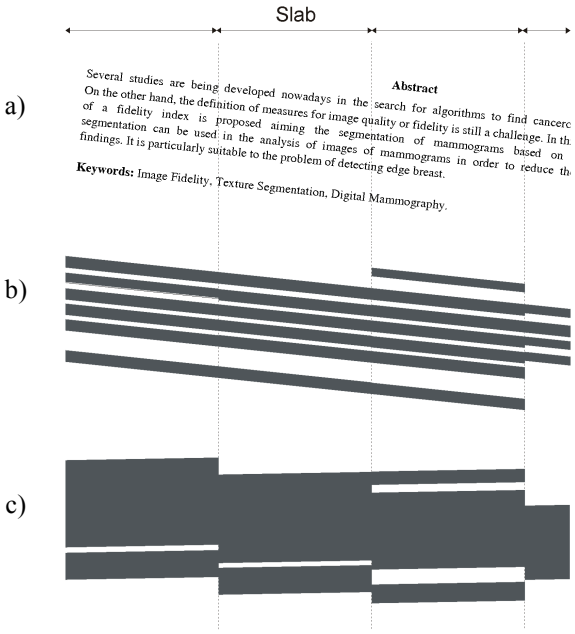


Figure 1. (a) An image with -6º skew angle. (b) Parallelograms constructed at -6º. (c) Parallelograms constructed at 1º.

By doing so, the angle with the largest white region is considered as the skew angle of the document. Figure 1b shows the parallelograms constructed with scan lines drawn at -6º, which is the same angle of the document skew. Figure 1c shows the parallelograms constructed with scan lines at 1º. As we can see, the white region at -6º is larger than at 1º.

2.1 Reasoning about the scan lines

To measure the size of the white region at an angle θº, Chou *et al*. proposed to count the number of white sections at that angle. However, this approach has a problem when the value of the angle increases (in positive or negative directions): the total number of sections covering the whole image at that angle is also changed.

As illustrated in Figure 2a, at 0º we will have *n* scan lines, which is also the number of rows (the height) of the image. In Figure 2b, the scan lines are rotated and the total number of scan lines will be *n* + a, where *a* is the number of extra small lines.

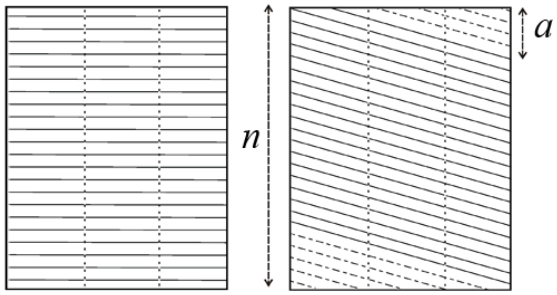


Figure 2. (a) A document with scan lines drawn at 0º. (b) The scan lines drawn at an angle different from 0º. The total number of scan lines is increased with *a* extra small lines.

Aiming to make a fair evaluation, one will need to specify the total number of sections for every angle. In other words, if you have 1000 sections at angle θº, you will need 1000 sections at every angle.

As shown in Figure 2b, part of the image must be ignored to assure that the total number of scan lines is constant, so only the lower (or upper) part of the smaller lines will be considered. This is a problem because the part ignored can contain valuable information to estimate the skew angle.

Even setting up the number of scan lines to a constant, a different number of sections per angle is produced. This happens because, in the upper and lower parts of the image, the lines are smaller than the ones in the middle of the document. For example, in Figure 2a, there are 3 slabs, so the total number of sections is 3×n. In Figure 2b the total number is changed because of the smaller lines. Some sections disappear and others have a smaller size.

Another problem with the Chou's approach is: counting the number of white sections will give the same weight for different number of pixels. Thus, small sections which contain only few pixels will have the same "importance" in the analysis as sections with the full slab width. This happens in the lower (or

upper) part of the image and with the sections in the last slab in the right.

Based on that, instead of using the number of white sections, it is here proposed a more efficient alternative to measure the white sections: based on its area. So instead of counting the number of white sections, we will count the number of white pixels (i.e. the number of pixels that is not covered by parallelograms). To avoid constructing the image with the parallelograms, an efficient way to make this measure is through the operation:

if no black pixel is detected in the section
then white counter += size of the section

This strategy naturally gives a weight to a section proportional to its size. In other words, small sections will contribute to increase the counter of white sections only with its corresponding number of pixels.

A benefit when using the background area is that: it is not necessary to care about fixing the total number of sections; this information is preserved by the total area (number of pixels) of the document at all angles. Thus, the entire image can be used to estimate the skew angle and no part is ignored.

2.2 Reducing the search space

To search for the skew angle of a document, Chou limits its search within [-15º,+15º]. This is a practice commonly adopted by other techniques with no decrease in performance because real scanned documents usually have a small amount of skew.

To save time, instead of search trough all angles, Chou proposed to search for the skew angle of the document as follows:

1) Search for the best skew angle β within [-15º,+15º] with a step size of 2º;
2) Search for the best skew angle γ, within [β-1, β+1], with a step size of 1º;
3) Finally, search for the best skew angle δ within [γ - 1, γ +1] with a step size of 0.1º.

Where δ is the estimated skew angle of the document and the best skew angle at each step is the one that achieves the largest number of white sections (in our approach, the largest white area).

A good way to speed up the first of these three steps is to watch the behavior of the white region measure at each angle. This measure should increase as the angle of the scan lines is closer to the real skew of the document and should decrease as the angle of the scan lines is distant from the real skew angle.

However, this evaluation is not always true using the number of white sections to measure the size of the white region. On the other hand, the area is an alternative to this.

In Figure 3 we show the behavior of measuring the white region in the first step of the algorithm – where we look for the β angle from -15º to +15º with step size of 2º. Figure 3a shows a scanned document with a skew angle close to 0º. Figure 3b shows the behavior of the white counter using the number of sections as proposed by Chou. We see that Chou's original method will fail in finding a correct skew angle close to 0º because the angles close to 15º and -15º have larger values. This is caused by a problem already discussed: the size and the number variation of the sections.

Figure 3c shows the behavior of measuring the white region through the area, as we propose here. We can see that close to 0º the white counter assumes its largest value and decreases when the scan lines goes away from 0º. Figure 3d shows the curve of measuring the white area of the same document now with -10º skew angle.

Thus, the behavior of the curve can be helpful in early stopping the search for the best skew angle β. That is, if we detect that the value of the white counter is decreasing, we can stop the search for β and move to the next step, refining the search.

Another point to consider in real cases is that most of scanned images have 0º skew angle or close to this. So, we propose to start our search evaluating 0º and observe the behavior of the curve

We can define $B(\theta^o)$ as the size of the white region. So, we propose a new way to speed up the search for the skew angle as follows:

1) Search for the best skew angle within {0º, 2º, -2º}.
   If $B(0^o)$ has the largest value
   Then assign 0º to β and move to the next step.
   Else continue searching for the best skew angle β in the direction of the largest $B(\theta^o)$

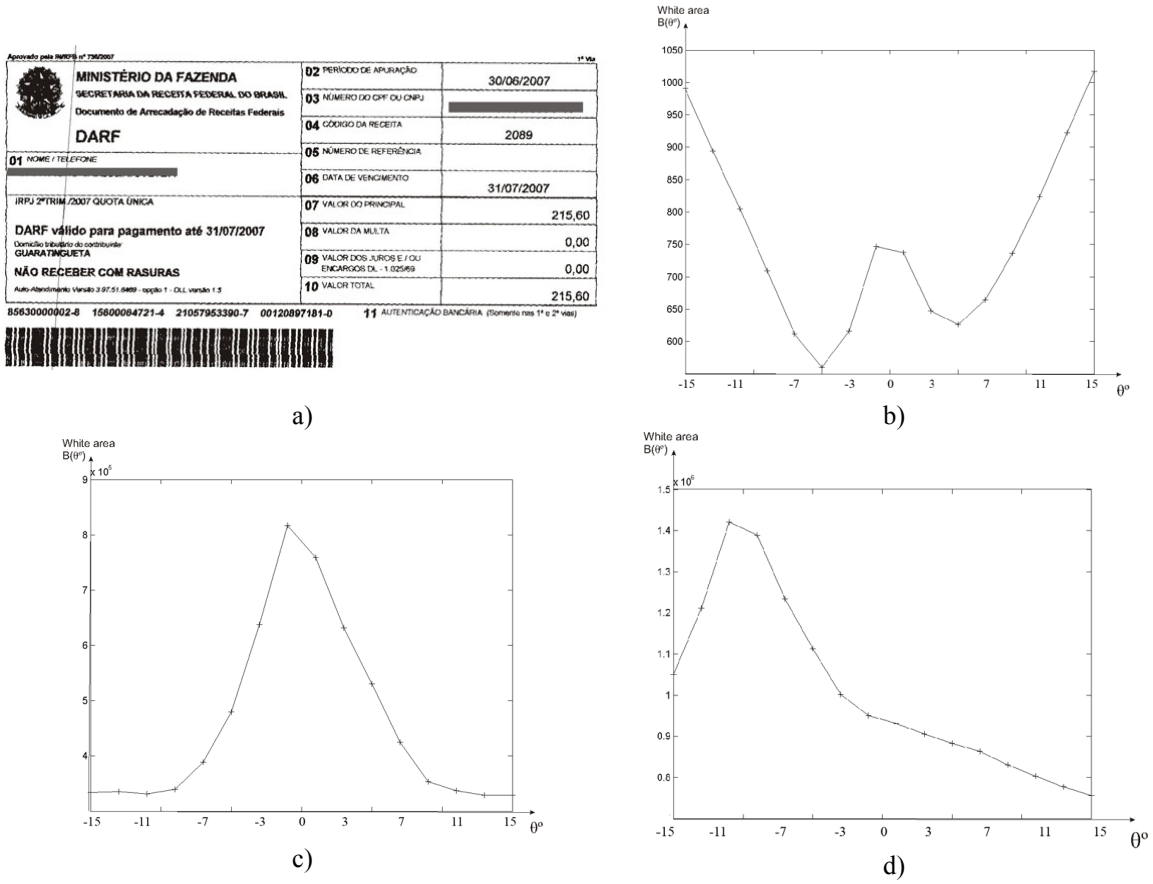


Figure 3. The behavior of the white region quantity – looking for β angle. (a) A scanned document with an angle close to 0º. (b) Measuring the white region of the document in (a) trough the number of white section as proposed by Chou. (c) Measuring the white region of the document in (a) trough the area. (d) Measuring the white region of the document in (a) rotated by -10º using the area.

with a step size of 2º. Stop when *B(θº)* starts to decrease.

2) Search for the best skew angle γ, within [β-1, β+1], with a step size of 1º;
3) Finally, search for the best skew angle δ within [γ -0.6, γ +0.6] with a step size of 0.1º.

Using the strategy proposed above, it is possible to note the reduction in the search space for the skew angle β. The best case occurs when *B(0º)* is larger than *B(2º)* and larger than *B(2º)*, so β is set to 0º. Knowing that most of the document images in real problems has low skew angles this will greatly affect the computation time. The worst case occurs when β is +15º or -15º. Even in these cases, the search space will be reduced by half in comparison with the original search described by Chou, because only one path (positive or negative) need to be explored.

The third step also reduces the search space. We only search for the skew angle from γ-0.6 to γ+0.6 instead of γ-1 to γ+1. This is done aiming to avoid redundancy with the previous step.

To avoid local minimum, we suggest to stop the search for the best skew angle β only when the *B(θº)* decreases after two iterations.

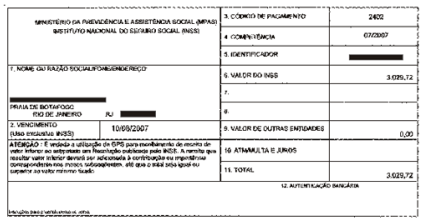
2.2 Minimizing the interference of noise and vertical separators

The presented method is vulnerable to images with noise in the background. As it turns to gray every section that has at least one black pixel it can become difficult to evaluate the inclination angle of noisy image documents.

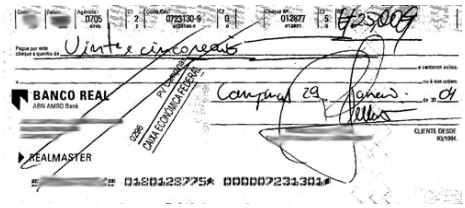
This vulnerability can also happen in images of documents with large amount of vertical separators (vertical bars), such as tables, forms and other non-textual elements. In these cases, all the sections covered by a single vertical separator will be turned to gray. This problem also happens with handwritten documents. Figure 4 illustrates these cases.

If the document image has, at least, one of these components, the skew angle estimation can be

damaged. To overcome this vulnerability we use a threshold T to decide if the section will be turned to gray or not.



a)



b)

Figure 4. Examples of documents with non-textual elements that can mislead the skew angle estimation proposed by Chou. (a) A form containing vertical separators. (b) A Brazilian Bank Check with noise, handwritten components and a stamp (data was hidden for identity preserving purposes).

Based on that, black pixels coming from vertical bars or background noise will not interfere in the skew angle estimation. After preliminary experiments the threshold T was set to 5 pixels.

3. Experimental Study

We evaluated our proposal over a collection of images. The original Chou's approach was implemented and tested for comparison. We computed the error of skew angle estimation for an image as the difference between the estimated angle and the target angle. This is called the *estimation error*.

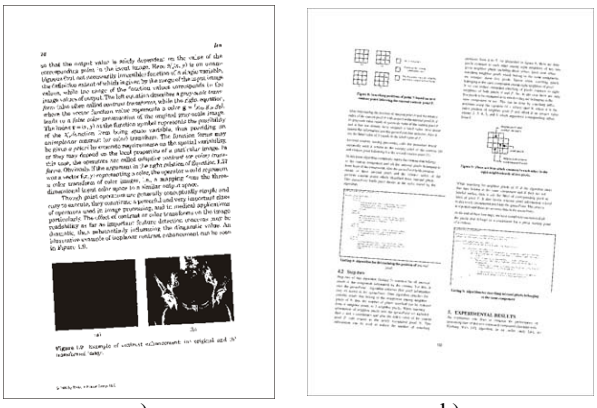
Chou compared his technique with the following ones: a projection-based method [3], a maximum variance of transitions counts method [2] and a cross-correlations method [4]. As conclusion, Chou showed that his method achieved the best results in the estimation of the skew angle and also in the computation time. Based on that, in this section we will only show the comparison between our method and the original Chou's estimation skew angle.

For all the experiments the slab width was set to 450 pixels for both approaches. The threshold of our approach was fixed to 5 pixels.

3.1. Databases

To examine the performance of the proposed method we constructed two disjoint databases with document images: one artificial and one real.

The first one was used to compute the estimation error of skew angles. To ensure that each document in this database had 0º skew angle we generated images from electronic documents. After that, the images were artificially rotated with an angle within [-15º, +15º]. For this database we collected a variety of documents composed of mixture of textual and non-textual components (like forms, tables, pictures, etc.) with 200dpi. Figure 5 shows examples from the documents used in this database. We collected 644 images in this manner.



a) b)

c) d)

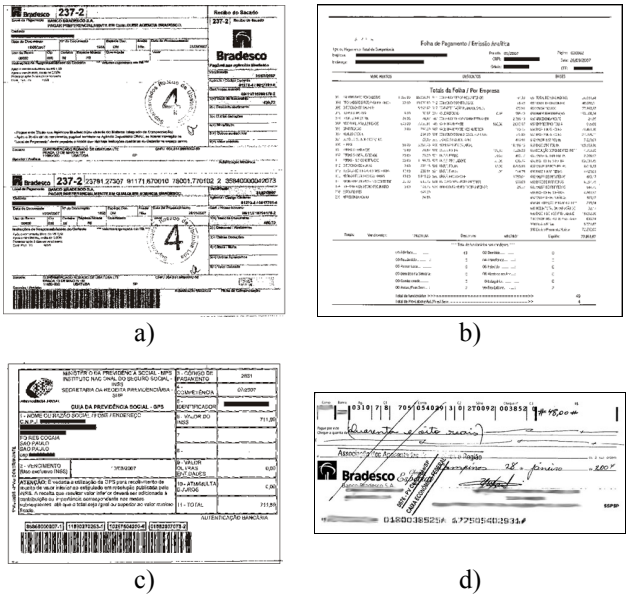
Figure 5. Examples of the first database composed of electronic documents converted to images. Documents rotated at (a) 3º, (b) -4º, (c) 7º and (d) -13º.

In order to evaluate the difference in performance using the threshold described in Section 2.2, we simulate noisy images by adding salt and pepper noise in the first database. The error over images with the salt and pepper noise was computed apart.

The second database was used to evaluate the performance of the method over real scanned images. We collected 3,268 images including Brazilian bank checks, forms, payrolls and bank payment slips distributed in the categories shown in Table 1. Figure 6 illustrates some examples of images that compound this second database.

Table 1. Distribution of images in the real scanned image database

Document type	Number of images
Bank payment slip	54
Payroll	167
Forms	303
Brazilian bank checks	2,744



a) b) c) d)

Figure 6. Examples of the second database composed with real scanned images of documents. (a) Bank payment slip, (b) Payroll, (c) Forms and (d) Brazilian bank checks (data was hidden for identity preserving purposes).

As can be seen, the documents contain handwritten components, vertical separators, non-text areas and noise. As we do not know the real skew angle of the images, the error was not calculated over this database and we only made a visual analysis of the results.

3.2 Results with synthetic rotated images

Table 2 presents the results of estimating the skew angle of the images from the database of synthetic rotated documents. The column labeled as "Chou" represents the error computed using the original Chou's method. The column labeled as "Proposed" shows the results from our approach.

As we can see, the proposed method achieved better results. This improvement occurred especially in images with a larger amount of vertical separators. Figure 7 shows an example of an image containing a table rotated at 8º. Chou's method estimated wrongly (-17º). On the other hand, our approach estimated correctly (8º).

Table 2. Error rates in degrees (º) from the images of the synthetic database.

Synthetic database	Chou	Proposed
Average estimation error	0.805	0.020
Standard deviation	2.894	0.042

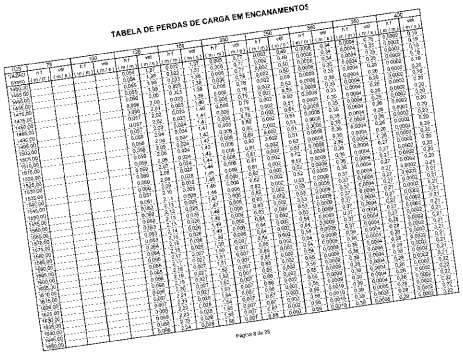


Figure 7. Image rotated at 8º. Chou's original method had a great sensitivity to the large amount of vertical separators and estimated a skew angle of -17º. The proposed approach succeeded in finding 8º skew angle.

Aiming to verify the performance using the threshold proposed in Section 2.2 we added salt and pepper noise with different densities in images of the first database. As shown in Table 3, the error over the Chou's approach was greatly increased, especially with larger density (d) noise. On the other hand, the proposed approach continued to work well. Visually examining the results for d = 0.02 and d = 0.03, we saw that Chou's estimation algorithm strongly confused the skew angle estimation of all the images in the database.

Table 3. Error rates in degrees (º). Images from the synthetic database with salt and pepper noise.

Density	Estimation error	Chou	Proposed
0.01	Average error	4.0358	0.0599
0.01	Standard Deviation	5.2709	0.0808
0.02	Average error	9.0026	0.3673
0.02	Standard Deviation	4.0196	1.9440
0.03	Average error	9.0639	1.9034
0.03	Standard Deviation	4.0157	4.9746

3.3 Results with real scanned images

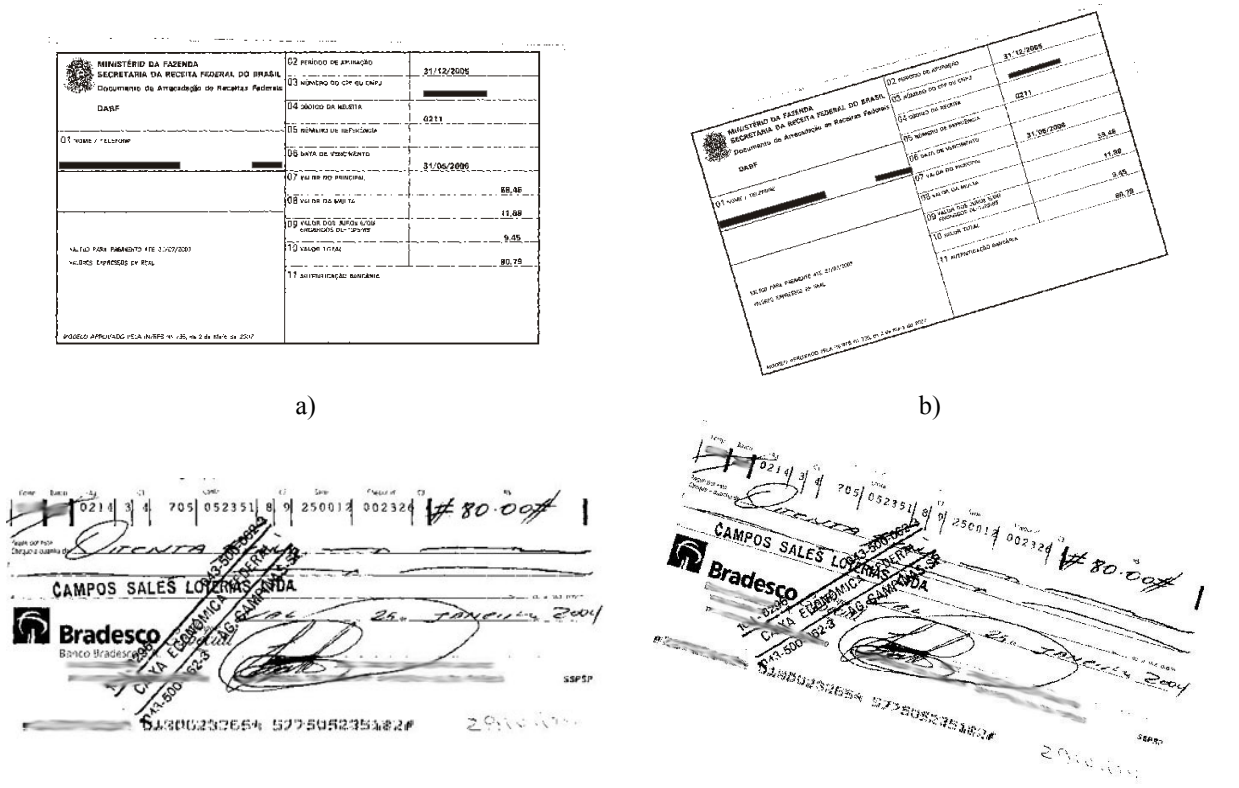
Aiming to observe the skew correction in a practical environment, the new approach was tested over real scanned images from the second database. Figure 8 shows some results examples.

As we decided not to measure the error rate in this database, the results were visually examined.

The proposed approach achieved very satisfactory results for all the images in the real scanned images, but the original's Chou method just confused a great part of the skew angles.

For the payroll and bank payment slip dataset, both the approaches achieved very satisfactory results. However for the bank checks and forms datasets, the Chou's approach confused almost all the images. Despite of this, the proposed approach still presented very satisfactory results. Figure 8 shows examples.

This happened because our new approach makes a better evaluation about the white space (the background) using the area to measure it. As the images contain components that are not plain text, (like handwritten components, noise, vertical separators, etc.) the proposal of using a threshold brought the ability to deal with this kind of features.



a) b)

c) d)

Figure 8. Examples of the skew correction of real scanned images. (a) Scanned document of the category "Forms". (b) Result of applying the correction to the form based on the Chou's estimation – Chou's method detected 17º of skew. Our proposed approach detected 0º skew angle. (c) A scanned brazilian bank check. (d) Result of applying the correction to the check based on the Chou's estimation – Chou's method detected 16.5º of skew. Our proposed approach detected -0.1º skew angle.

3.4 Computational time results

Now we make an analysis about reduction in computation time using our approach. Recall that Chou look for the skew angle β in the full interval [-15º,+15º] using 2º as the search step size and the search for the δ angle occurs in the interval [-1,+1] with a step size of 0.1º.

Chou concluded that its method was the best, in terms of computational time, when compared with the techniques in [2-4]. With the reduced search space proposed here, this computation time is also reduced.

Table 4 shows the average difference (in %) in the computation time between both approaches. We show how much our approach accelerated the original Chou's method. In the first row we show that the new

approach is approximately 12% faster than the Chou's original one for the synthetically rotated images.

Table 4. Gain in computation time (in %) using the reduced search space.

Average difference	%
Images synthetically rotated in various angles	12.35
Real scanned images	51.30

And as we want to show how we accelerate the real cases (where the images have skew angles close to 0º) we present in the second row the results for the second database, composed with the real scanned images. We can see that reducing the search space gave approximately 50% reduction in the computational time. As an example, if the Chou method needs 1 second to find the skew angle of a document with a 0º skew angle, our approach requires only 0.5 seconds, in average.

It is important to mention that the Chou's method was faster than a projection-based method [3], a maximum variance of transitions counts method [2] and a cross-correlations method [4]. Based on that, we can say that the approach present here is faster than these techniques too.

4. Final Remarks

Using parallelograms to fit the objects of a document is useful to estimate its skew angle. In this paper we proposed a variation of Chou's method based on parallelogram covering.

Through experimental tests over synthetic rotated images we showed that the proposal achieved better results over noisy images and documents containing vertical separators, like tables and forms. Through the real scanned images database we saw that our new approach can make a better evaluation of the background in the document and also works better with printed images containing handwritten components, noise and vertical separators.

We also showed an efficient way to reduce the search space thus reducing the computational time of the algorithm. About the time consumption, the proposed approach saves more time when the images have skewing angles close to 0º. That is because we assumed that most real cases has this inclination. As the proposed algorithm starts searching at 0º, the time consumption is proportional to the skew angle of the image.

One last consideration is about a limitation of this approach. When scanning a document, sometimes the image can acquire more than one angle at different parts of the document. As the proposed technique was developed to find a global skew angle for an image, it is not suitable to correct the skew of a document with various angles. We leave this as future work to evolve the present approach to deal with this kind of skew.

Acknowledgments

The authors would like to thank Fu Chang and Chien-Hsing Chou for their valuable comments and explanations.

References

[1] Amin, A.; Fisher, S."A Document Skew Detection Method Using the Hough Transform", Pattern Analysis & Applications 3 (3) 2000 pp. 243-253.

[2] Ishitani, Y. "Document skew detection based on local region complexity", Proceedings of the 2nd International Conference on Document Analysis and Recognition, 1993, pp. 49-52.

[3] Postl, W. "Detection of linear oblique structures and skew scans in digitized documents", Proceedings of the 6th International Conference on Pattern Recognition, 1986, pp. 687-689.

[4] Avanindra, S. "Robust detection of skew in document images", IEEE Transactions on Image Processing. 6 (2) 1997. pp. 344-349.

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