Skew Detection Using Directional Profile Analysis

Stephen W. Lam and Victor C. Zandy
Center of Excellence for Document Analysis and Recognition (CEDAR)
State University of New York at Buffalo

The UB Commons, Suite 202
520 Lee Entrance
Amherst, NY 14228-2567, USA

Abstract

Skew detection is an important stage of document image processing. It is desirable to have a fast, all-purpose skew detector. This paper describes a fast skew detector using directional profile analysis. It can quickly and accurately detect skew on a wide variety of document images, including images with multiple text alignments and non-text regions. The complexity of the skew detection process directly depends on the document layout complexity.

1. Introduction

One of the problems in document image understanding is that the document image is not always digitized correctly. In some cases, the document is not aligned properly on the scanner bed, causing the image to be skewed. This presents a difficulty because most document analysis processes, such as page segmentation and text recognition, require the document image to be in proper alignment. Since proper alignment cannot be guaranteed during image capture, a separate routine is needed to determine the skew angle.

A useful skew detection routine should be capable of processing simple document images which consist only of text, as well as documents with complex layout which, in addition to text, contain figures, photographs, tables, and other non-text elements. It should also be able to process document images that contain both horizontal and vertical text (*e.g.*, Japanese documents). Finally, it should be able to do its processing quickly and reliably without specialized hardware. Several skew detection algorithms had been proposed [1, 2, 3]. The problem with most of them is that they are not able to process an image if it has a complex layout. They are also generally slow, requiring specialized hardware to run in reasonable amount of time.

A fast algorithm [4] was recently proposed. It is capable of detecting skew in complex images, including those with both horizontal and vertical text, without special hardware. The algorithm introduced two new ideas. One is that skew can be detected in complex images by examining suitable local regions. Suitable local regions are regions of the image expected to contain lines of text (Figure 1). The algorithm is able to isolate such regions automatically. The skew angle of a suitable local region is defined to be the overall document image skew angle. The second idea is that a complexity variance profile can be used to first confirm that a local region is suitable, and then to determine the region's skew angle. Complexity variance, V, depends upon the orientation angle, θ, at which it is computed and is defined as follows:

$$V(\theta) = \frac{1}{n}\sum_{i=1}^{n}(N_i - M)^2$$

where

$$M = \frac{1}{n}\sum_{i=1}^{n} N_i$$

n is the number of scan lines, and N_i is the number of transitions from white to black pixels on scan line i.

The algorithm determines the region's suitability and skew by examining complexity variance over a range of angles (Figure 2).

Figure 1: Skew detection on a complex layout document requires local text regions.

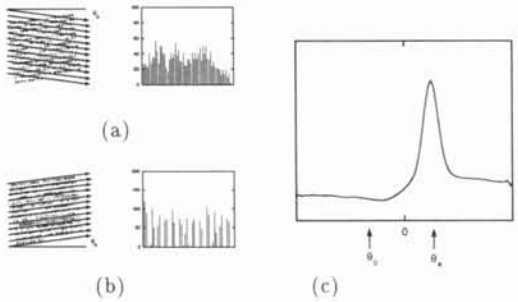


Figure 2: (a) Complexity variance is low when θ is away from skew. (b) Complexity variance is high when θ is close to skew. (c) Skew estimated to be angle which maximizes complexity variance.

This paper describes the development of a skew detector based on this approach. In the course of its development, two aspects of the approach were found to be difficult in ensuring its robustness. The first difficulty is the automatic selection of candidate local regions and the

second one is the profile analysis. The suitability of local regions cannot reliably be decided by analyzing profiles on an individual basis. The proposed new skew detector, `detect-skew`, circumvents these difficulties by using a new criteria for local region selection, and then requiring a consensus from several local regions before making a final skew angle estimation. This new approach is believed to be more reliable because of its statistical nature.

2. Background

Several skew detection methods using Hough Transform were reported [3, 5]. However, Hough Transform is inherently slow in software implmentation. It is also required the documents contain mainly text. An approach using energy alignment measure was proposed by Baird [1]. Although it is faster than the Hough Transform approach, it is still incapable of processing complex layout documents. A skew detection method called left margin search was proposed by Dengel [2]. It is done by forming a straight line between the left most pixels of the image rows. However, this method will fail when the document contains irregular indentation objects (such as graphics and figures) at the left hand side of the page.

The rest of this section gives a brief overview of the approach described in [4], which is the basis of the skew detector discussed in this paper.

2.1 Local Region Selection

Ishitani [4] described the process by which regions expected to contain text lines are identified in the image prior to skew detection. A circle with a fixed radius is moved intermittently from the top left corner to the bottom right corner. At each stop, the complexity is computed for all angle resolutions. The regions with high complexity will be retained for further analysis. However, this method has been demonstrated to be impractical on a wide variety of images for the following reasons:

1. Regions with high black pixel density (such as photographs) are often among the most complex regions in an image.
2. Since some scan lines in a text region may have very low complexity, as is necessary for the algorithm's success, the total complexity of a text region may not be high enough.
3. Image noise increases the complexity of a region; poor candidates may be chosen in low-quality images.

2.2 Profile Analysis

The key step in Ishitani's algorithm is deciding whether a local region's complexity variance profile "has a sharp peak". Attempts to develop a reliable routine revealed several key problems:

1. Peak height and sharpness is dependent upon the density of the text, making it difficult to form a general parameterization
2. Complexity variance is rarely constant at points away from the peak, making it difficult to compare flat and sharp features
3. Profiles for non-text regions have unpredictable shapes, including sometimes sharp peaks

The algorithm is able to end local region complexity analysis as soon as it finds a single profile with a sharp peak. However, it is risky to rely on just one suitable local region. The alternative approach taken by `detect-skew` is described in the following section

3. Proposed Algorithm

`detect-skew` determines the skew of an input document image with the following loop:

1. Compute the coordinates of candidate local regions on the image, as described in Section 3.1.
2. Compute the complexity variance, $V(\theta)$, and local skew estimate for the next region, as described in Section 3.2.
3. Decide if there is a consensus in local skew estimates, as described in Section 3.3. If there isn't and there are still unexamined local regions, goto step 2.
4. Estimate overall document image skew, as described in Section 3.4

The steps of this algorithm are detailed below.

3.1 Region Selection

`detect-skew` does not have a mechanism for selecting suitable local regions before performing complexity analysis. Consequently, several, if not many, unsuitable local regions will have to be processed. Since these at best will only waste processing time, it is desirable to find other criteria that will reduce the probability of analyzing an unsuitable region. Two criteria used in `detect-skew` are absolute and relative location of local regions.

Absolute Location of Local Regions

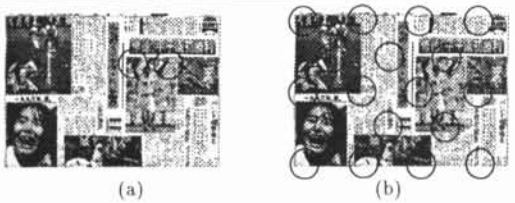
Skewed images contain excessive whitespace around their borders, as illustrated in Figure 3. Skew estimates based on regions located along these borders can be misleading because they don't contain enough information to make useful complexity variance profiles. Thus such regions should be avoided. Most of the valuable complexity information is found within the inner subimage indicated in Figure 3. `detect-skew` draws an imaginary border inside the image defining this subimage, and selects local region candidates only from within this subimage. This then reduces the probability of picking an unsuitable local region.

Figure 3: Misleading information in a local region.

Relative Location of Local Regions

Nonoverlapping regions adjacent to each other will generally each contain portions of a document feature which is common to all of them (Figure 4(a)). When these features

are textual, each region will yield similar "good" skew estimation results. When these features are non-textual, each region will yield comparably "bad" skew estimation results. In either case, clearly it is unnecessary to consider all of the regions in a subarea of the image. To locate suitable regions, it is best to distribute the search evenly over the entire image, without spending too much time looking at regions adjacent to previously considered regions. In `detect-skew`, experiments with strategies for distributing the search have been conducted. Presently, `detect-skew` looks at regions in a "checkerboard" pattern, similar to the one shown in Figure 4(b).



(a) (b)

Figure 4: (a) Adjacent regions often contain parts of the same feature. (b) A pattern of local region selection for `detect-skew`.

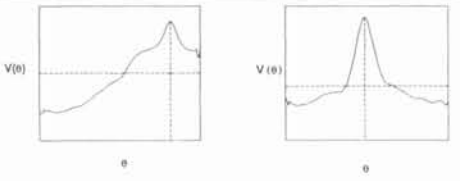
3.2 Complexity Variance Computation

During the complexity variance computation stage, two profiles of the local region's complexity variance are made. One, $V_h(\theta)$, is for scan lines oriented with respect to the horizontal axis of the image, and the other, $V_v(\theta)$, is for scan lines oriented with respect to the vertical axis. This makes it possible to detect skew in images which contain horizontal and/or vertical text, such as Japanese documents.

Complexity variance in `detect-skew` is a relatively straightforward computation, but it can be extremely time-consuming if not done carefully. `detect-skew` takes four measures to maintain a high level of efficiency.

1. Instead of counting all transitions from white to black pixels found on each scan line, `detect-skew` counts the number of black pixels found at sample points on each scan line. The distance between sample points is fixed.
2. `detect-skew` uses circular regions. The symmetry makes it possible to compute both the horizontal and vertical complexity variance at the same time, because all that is needed to go from one to the other is an exchange of coordinates. Circles also makes it simpler to identify where the scan lines intersect with the edges of the region.
3. Complexity variance is computed in exactly the same way for each local region. Before starting complexity analysis, a table is created which contains all information necessary for computing complexity variance. This saves `detect-skew` from having to duplicate calculations.
4. By fixing the region radius, the distance between scan lines, and the distance between sample points on the scan lines, this table can be computed at compile time, thus improving initialization time.

To decide which profile will give a better skew estimate, `detect-skew` makes a rough decision about which profile has a sharper peak. Peak sharpness is estimated by comparing the value of the profile maximum to the average value of the entire profile. If the peak is not sharp, then the ratio of these two values will be low (Figure 5(a)). On the other hand, if the profile is sharp, then this ratio will be high (Figure 5(b)). More precisely, let max_h and avg_h be the maximum and average value of $V_h(\theta)$, respectively, and max_v and avg_v be the maximum and average value of $V_v(\theta)$, respectively. Then the profile to use, $V(\theta)$ is chosen by the following rule: If $\frac{max_h}{avg_h} > \frac{max_v}{avg_v}$ then $V(\theta) = V_h(\theta)$; otherwise $V(\theta) = V_v(\theta)$.



(a) (b)

Figure 5: (a) Ratio of maximum to average profile value is low without sharp peak. (b) Ratio of maximum to average profile value is high with sharp peak.

Once the profile has been chosen, the skew angle θ_s for the region is defined as the angle which maximizes $V(\theta)$. This means that a skew estimation will be made for every region passed to the complexity variance routine. These skew estimates are collected and a final skew estimation for the overall document is made through a consensus vote, as described in the following section.

3.3 Consensus Voting of Skew Angle

`detect-skew` differs most dramatically from Ishitani's algorithm in the way the skew angle is deduced from profiles. Ishitani's method checks whether a profile has a sharp peak. If it does, then the angle which maximizes the profile is said to be the skew angle, and the skew detection is finished. With this method, the skew measurement depends entirely on the profile of one suitable local region. In `detect-skew`, since suitable local regions cannot be reliably isolated, a consensus must be reached among several profiles before a commitment to a final skew estimation will be made.

Typically, the skew estimates of several unsuitable regions will be randomly distributed. Measurements of suitable regions, on the other hand, should correspond with the overall document image skew. Therefore, if enough suitable regions are taken from the image, eventually a "consensus" will be reached, that is, most of the skew estimates will lie within a small neighborhood of some angle. In `detect-skew`, this angle is defined to be the mode of the skew estimates. After the nth skew estimate has been made, the following consensus criterion is applied. Let m be the mode of the n estimates, and c be the number of estimates θ_i such that $|m - \theta_i| < \epsilon$. Then if $c/n > 2/3$, a consensus has been reached.

Clearly this approach requires that the majority of the regions selected be suitable regions. But since region suit-

ability is not considered when regions are selected, this implies that the majority of the image must be text. This would appear to restrict the domain of input images. However, the consensus criterion was added as an afterthought to improve the overall efficiency of **detect-skew**. It does not make the final skew estimation. After a local region is analyzed, the consensus criterion permits **detect-skew** to decide whether it is necessary to process more regions. If the criterion is satisfied, confidence can be placed in **detect-skew** to compute an overall document image skew estimate without additional processing.

3.4 Estimating Overall Image Skew

After a consensus has been reached, or if there are no more local regions to process, the overall document image skew estimation θ_s is computed. In contrast with Ishitani's method, several local skew estimates are used to make the final skew estimation. The introduction of a statistical element to the estimation has greatly improved the reliability of the skew detector.

The overall skew estimation, θ_s is made from the following weighted average of the local skew estimates:

$$\theta_s = \frac{\sum_i \theta_i w_i}{\sum_i w_i}$$

for all i such that $|m - \theta_i| < \epsilon$ and where w_i is the maximum value of the complexity variance profile chosen for region i and m is the mode of all region skew angle estimates.

4. Experimental Results

detect-skew was implemented on a SPARCstation2 in the C programming language. Experiments have been carried out using 467 document images taken from books, facsimile transmissions, journals, magazines, and newspapers. These images vary widely with respect to text/non-text ratio, text orientation, fragmentation, size, noise, and density. Image text is in Japanese, English, or both. Examples are given in Figure 6. The following parameters were set in the experiments. (1) Local regions have a radius of 300 pixels. (2) Scan lines are 10 pixels apart. (3) Sample points along scan lines are 20 pixels apart. (4) No more than 15 local regions are analyzed for any image. (5) The range of the scan line orientation is ± 30 degrees. (6) The resolution of the scan line orientation is 0.50 degrees. The experimental procedure went as follows. Each image was first given to **detect-skew** unskewed, in order to determine is normal skew θ_n. The image was then rotated by a random angle $\theta_r (|\theta_r| < 30.0 \text{ deg})$ so that the image skew becomes $\theta_s = \theta_n + \theta_r$, and given to **detect-skew**, which returned an the estimated skew angle θ_e. The difference between the estimated and actual skew angles $\delta = |\theta_s - \theta_e|$ was used to judge the performance of **detect-skew**.

The results are given in Table 1. The average time per image was 3.16 CPU seconds.

5. Conclusion

A skew detection algorithm using directional profile analysis has been described. The skew detector can handle a large variety of documents with complex contents. The algorithm utilizes statistical information derived from small local regions and makes "consensus" decision based on the statistical estimation. Preliminary experiments had shown promising results in both speed and accuracy. Several investigations are planned to improve performance, including new region selection methods, and a reconsideration of profile analysis methods.

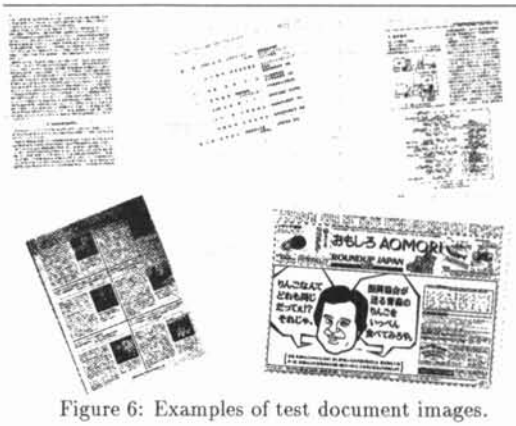


Figure 6: Examples of test document images.

δ	# Images	% Images
$0 \leq \delta \leq 0.50$	333	71.3
$0 \leq \delta \leq 1.00$	430	92.1
$0 \leq \delta \leq 2.00$	459	98.3
$\delta > 2.00$	8	1.7

Table 1: Experimental results.

References

[1] H.S. Baird, "The Skew Angle of Printed Documents", *SPIE's 40th Annual Conference and Symposium on Hybrid Imaging Systems*, Rochester, New York, May 1987.

[2] A. Dengel, "ANASTASIL: A System for Low-Level and High-Level Geometric Analysis of Printed Documents", *Structured Document Image Analysis*, H. Baird, H. Bunke and K. Yamamoto (eds), Springer-Verlag, 1992.

[3] R.O. Duda and P.E. Hart, "Use of Hough Transformation to Detect Lines and Curves in Pictures", *Communications of the ACM*, Vol. 15, No. 1, January 1992.

[4] Y. Ishitani, "Document Skew Detection Based on Local Region Complexity", *Second International Conference on Document Analysis and Recognition*, Tsukuba Science City, Japan, October 1993.

[5] A. Rastogi and S.N. Srihari, "Recognizing Textual Blocks Using the Hough Transform", TR 86-01, Dept. of Computer Science, State University of New York at Buffalo, 1986.