# Automatic Text Skew Estimation in Document Images

Su Chen and Robert M. Haralick

Department of Electrical Engineering
University of Washington
Seattle, WA 98195

Ihsin T. Phillips

Department of Computer Science
Seattle University
Seattle, WA 98122

## Abstract

*This paper describes an algorithm to estimate the text skew angle in a document image. The algorithm utilizes the recursive morphological transforms and yields accurate estimates of text skew angles on a large document image data set.*

*The algorithm computes the optimal parameter settings on the fly without any human interaction. In this automatic mode, experimental results indicate that the algorithm generates estimated text skew angles within* $0.5^\circ$ *of the true text skew angles with a probability of* 99%.

*To process a 300dpi document image, the algorithm takes 10 seconds on SUN Sparc 10 machines.*

## 1 Introduction

Many document layout analysis systems demand very small text skew angles on document images. If the skew angles are as large as a couple of degrees, the document layout analysis systems usually require the images be deskewed. This requires procedures to detect the text skew angle of a document image and then deskew the image via image rotation. The deskewed image allows more compact representation of image objects, e.g. paragraphs, where they can be represented by the rectilinear bounding boxes. And more efficient algorithms could exist due to the simplified representation.

Earlier work on detection of text skew is basically a two stage process [1] [2]. In the first stage, feature positions for alignment are detected. In the second, various hypothesis tests are applied to select a "good" subset from the detected alignment feature positions. Depending on the choice of alignment features and the criteria for the hypothesis testing, the text skew estimation algorithms are variously divided. Most methods require many heuristics and there have been no systematic experimental protocols to evaluate their performance.

In this paper, we discuss an engineering approach for the automatic text skew estimation in document images. Statistical methods are used throughout the algorithm design process. Our text skew estimation algorithm is based upon the recursive morphological opening and closing transforms [3] [4]. The algorithm is fully automatic in that there is no need for users to set any algorithm parameters. The algorithm itself can estimate the optimal parameter settings on the fly. In addition, a very large document image database has been used for the training and evaluation of the proposed text skew estimation algorithm.

Section 2 formally states the document text skew estimation problem. Then we describe the text skew estimation algorithm in Section 3. In Section 4, an experimental protocol for evaluating the algorithm is presented. Finally, the experimental results are summarized.

## 2 Problem statement

We assume that the origin of the coordinate system is at the geometric center of the image $I$. The $X$-coordinate is the column direction and the $Y$-coordinate is the row direction. When the document image is rotated counter-clockwise, the text skew angle is positive; otherwise, it is negative.

**Definition 1** *The* **text skew angle** *of a document image is denoted by* $\varphi$ *and is defined as its dominant (most frequently occurring) text baseline direction, i.e. the counter-clockwise or clockwise orientation of text baseline with respect to* $X$*-coordinate.*

Let $I$ denote a bi-level document image. Let $\varphi$ and $\hat{\varphi}$ represent the true and the estimated document text skew angle. The problem of the document text skew estimation can be formulated as follows:

*Text skew estimation problem:*

*Given a document image* $I$ *with text lines at unkown skews of* $\varphi_1, \varphi_2, \cdots, \varphi_n$. *Find* $\hat{\varphi}$, *an estimate of the true document text skew angle* $\varphi$, *to maximize the probability* $P(\varphi \mid I)$.

## 3 Algorithm description

Our text skew estimation algorithm [9] is a three stage process. In the first stage, text line directions are detected. The detection is based on the recursive closing transform (RCT) and recursive opening transform (ROT) described in [3] [4]. In the second stage, a subset of the text line directions along the dominant direction are selected. Lastly, a Bayesian estimate of the document text skew angle is calculated based on the selected text baseline directions. The algorithm can be summarized as follows:



0-8186-7128-9/95 $4.00 © 1995 IEEE

Proceedings of the Third International Conference on Document Analysis and Recognition (ICDAR '95)
0-8186-7128-9/95 $10.00 © 1995 IEEE

### Step 1: sub-sampling

A sampling algorithm is used to reduce the input image resolution to 100dpi. Although this may sacrifice a little bit on the estimation accuracy of the document text skew angle, it is a trade-off between the performance and the computational efficiency.

### Step 2: filling the inter-character gap

The RCT with respect to a structuring element $K_c$ is computed on the sub-sampled document image $I_s$. A threshold $T_c$ is estimated given the histogram of the RCT of $I_s$. Then the RCT is thresholded using the threshold value $T_c$: pixels that have values greater than zero but not greater than $T_c$ are set to binary one. This effects a morphological closing of the image $I_s$ by a structuring element given by $(\oplus_{T_c-1}K_c)$, $K_c$ dilated with itself $T_c - 1$ times,

$$I_c = I_s \bullet (\oplus_{T_c-1}K_c).$$

where $I_c$ denotes the thresholded RCT image of $I_s$. Ideally the output thresholded image closes only the intercharacter gaps and the interline gaps remain untouched. The structuring element $K_c$ could be chosen as either a $2 \times 2$ box or a $2 \times 3$ box.

### Step 3: removing ascenders and descenders

The ROT with respect to a structuring element $K_o$ is computed on the bi-level image $I_c$. A threshold $T_o$ is calculated based on the histogram of the ROT. The ROT is then thresholded using the threshold value $T_o$: pixels that have values greater than or equal to $T_o$ are set to binary one. Ideally, the output thresholded image will eliminate all the ascenders, the descenders and the over-fills. Let $I_o$ denote the thresholded ROT image of $I_c$. The processing sequence is equivalent to the following morphological opening operation:

$$I_o = I_c \circ (\oplus_{T_o-1}K_o) = [I_s \bullet (\oplus_{T_c-1}K_c)] \circ (\oplus_{T_o-1}K_o)$$

where for simplicity, we would normally take $K_o = K_c = K$.

### Step 4: line fitting

The connected component labeling is performed on the bi-level image $I_o$. Each detected connected component is considered a set of points in $Z^2$. Let $(x_i, y_i)$, where $i = 1, 2, \cdots, n$, denote the set of points in a connected component. A least square line fitting procedure is applied to minimize the sum of squared distance from the points to the fitted line. The line direction $\varphi$ and its variance are functions of the second order spatial moments of the points (denoted by $\mu_{xx}$, $\mu_{xy}$, $\mu_{yy}$) [7]:

$$\hat{\varphi} = -\frac{1}{2}\arctan\left(\frac{2\mu_{xy}}{\mu_{xx}-\mu_{yy}}\right)$$

$$\hat{\sigma}^2_{\varphi} = \frac{1}{n-2}\cdot\frac{\mu_{xx}+\mu_{yy}+2\mu_{xy}\sin 2\hat{\varphi}-(\mu_{xx}-\mu_{yy})\cos 2\hat{\varphi}}{\mu_{xx}+\mu_{yy}-2\mu_{xy}\sin 2\hat{\varphi}+(\mu_{xx}-\mu_{yy})\cos 2\hat{\varphi}}$$

### Step 5: robust skew estimation

The directions of the fitted lines may vary. Such variations arise because of the presence of multiple text baseline directions on a document image. In addition, some extraneous text baseline directions may also be introduced through non-textual objects or portions of skewed text lines from other pages. Therefore, we compute the dominant line orientation $\varphi_d$ and select the subset of line orientations (denoted by $(\hat{\varphi}_i, \hat{\sigma}^2_{\varphi_i})$, $i = 1, 2, \cdots, L$) that are close to $\varphi_d$.

Assume that the text skew angle $\varphi$ has a normal prior probability distribution $N(0, \hat{\sigma}^2_{\varphi_0})$. Then we calculate the Bayesian estimate of the skew angle by

$$\hat{\varphi} = \frac{\sum_{i=0}^{L}\omega_i\hat{\varphi}_i}{\sum_{i=0}^{L}\omega_i}$$

where $\omega_i = 1/\hat{\sigma}^2_{\varphi_i}$, $i = 0, 1, 2, \cdots, L$ [9].

## 4 Experimental protocol

### 4.1 Training the algorithm

#### Image population

Our image data set is based on the "UW-I English Document Image Database" [5] [6]. The database was developed at the University of Washington in 1993. It is intended for researchers in the areas of optical character recognition and document image understanding. It provides a substantial sized database for the algorithm developments and evaluations.

UW-I contains 1147 distinct document images. Among all the images, 979 of them are directly scan-digitized from technical journals published in the English language and the remaining 168 images are noise-free images synthetically generated from a set of LaTeX documents. UW-I provides with each real document image the ground-truth mean and standard deviation of the text skew angle and the number of observations in the measurement. The ground-truth text skew angle is not the true text skew angle, but rather an estimate. The synthetic image are without skew.

In our experiments, the data set was analyzed. Each of the document images in UW-I was rotated at various degrees of $0°, \pm 1°, \pm 2°, \pm 3°, \pm 4°, \pm 5°$, using a nearest neighbor interpolation algorithm. This yields a total population of $12617 = 11 \times 1147$ training images.

#### Experimental design

To gather the experimental data, we ran the algorithm under various tuning parameter settings and output the estimated text skew angle, its standard deviation and the number of observations. The experiments was carried out for each image in the population.

The following defines the configuration of the experiments. The selected parameter values are within their reasonable ranges: 1) Choose the structuring element $K$ from: $\mathcal{K} = \{2 \times 2 \text{ Box}, 2 \times 3 \text{ Box}\}$. 2) Choose



Proceedings of the Third International Conference on Document Analysis and Recognition (ICDAR '95)
0-8186-7128-9/95 $10.00 © 1995 IEEE

the RCT threshold $T_c$ from: $\mathcal{C} = \{3,4,5,6,7,8\}$. 3) Choose the ROT threshold $T_o$ from: $\mathcal{O} = \{T_o \mid \mid T_o - T_c \mid \leq 2, T_c \in \mathcal{C}\}$. 4) Choose $\sigma^2_{\varphi_0} = \infty$. The algorithm assumes uniform prior distribution on true text skew angle. Therefore, the cross product of the $\mathcal{K}, \mathcal{C}, \mathcal{O}$ constitutes all possible tuning parameter settings in the experiment. Let $(K, T_c, T_o) \in \mathcal{K} \times \mathcal{C} \times \mathcal{O}$ denote one tuning parameter setting.

### Experimental output evaluation

We consider the "goodness" measure to quantify the deviation of the estimated skew angle from the true text skew angle. For the synthetically generated document images, the true text skew angles are known. We use the squared difference between the true and the estimated skew angles as our goodness measure.

But for the real document images, the true text skew angles are unknown. Let $\hat{\mu}_{\varphi_D}$, $\hat{\sigma}^2_{\varphi_D}$, $N_D$ denote the ground truth text skew angle, its variance and the number of observations, respectively. Let $\hat{\mu}_{\varphi_A}$, $\hat{\sigma}^2_{\varphi_A}$, $N_A$ denote the algorithm estimated text skew angle, its variance and the number of observations, respectively. Also let $\mu_\varphi$ denote the true text skew angle. We define the goodness measure $\kappa$ in the following:

**Definition 2** *The* **goodness measure** *for the text skew estimation is denoted by $\kappa$ and is defined as $\kappa = (\mu_\varphi - \hat{\mu}_{\varphi_A})^2$ for synthetically generated document images; and $\kappa = \alpha\hat{\sigma}^2_{\varphi_D} + (1-\alpha)\hat{\sigma}^2_{\varphi_A} + \alpha(1-\alpha)(\hat{\mu}_{\varphi_D} - \hat{\mu}_{\varphi_A})^2$ for real document images, where $\alpha = \frac{N_D}{N_D+N_A}$.*

Based on this definition, we compute $\kappa$ for each of the tuning parameter configurations $(K, T_c, T_o) \in \mathcal{K} \times \mathcal{C} \times \mathcal{O}$ and for each of the document images in the population.

### Training the algorithm

The purpose of training is to allow the skew estimation algorithm to decide its optimal algorithm tuning parameters on the fly. More specifically, we want the algorithm to predict the optimal RCT threshold ($T_c$) and the optimal ROT threshold ($T_o$) given their respective histograms. The optimalities of $T_c$ and $T_o$ are defined in the following way:

**Definition 3** *The* **optimal RCT threshold** *is denoted as $T_c^{opt}$ and is defined by*

$$T_c^{opt}(K) = arg \min_{T_c \in \mathcal{C}} [\max_{T_o \in \mathcal{O}} \kappa(K, T_c, T_o)]$$

**Definition 4** *The* **optimal ROT threshold** *given $T_c = T_c^{est}$ is denoted as $T_o^{opt}$ and is defined by*

$$T_o^{opt}(K, T_c^{est}) = arg \min_{T_o \in \mathcal{O}} \kappa(K, T_c^{est}, T_o)$$

The training process is divided into two sequential steps:

1. Build a regression function to predict the $T_c^{opt}$: The predictors are the histogram of the RCT of the sub-sampled image.
2. Build a regression function to predict the $T_o^{opt}$ given the previously built $T_c$ regression function: The predictors are the histogram of the ROT of the thresholded RCT image using the predicted $T_c^{opt}$ value.

Our regression model is based on the regression tree technique [8]. The constructed regression trees allow our text skew estimation algorithm to predict the optimal algorithm parameter settings on the fly.

## 4.2 Benchmarking the algorithm

In this section, we describe a series of experiments aimed to evaluate various aspects of the text skew estimation algorithm. Each experiment measures the probability distribution of the text skew angle estimation error, i.e. the difference between the ground truth and the predicted skew angles. The cumulative probability distribution of the absolute text skew estimation errors is computed: Prob $[ \mid \hat{\mu}_{\varphi_A} - \hat{\mu}_{\varphi_D} \mid \leq x ]$.

### Performance in optimal mode

The experiment studies the performance of the text skew estimation algorithm under the optimal settings of the $T_c$ and $T_o$ parameters. The $T_c$ and $T_o$ settings are optimal in the following sense:

$$[T_c^{opt}(K), T_o^{opt}(K)] = arg \min_{[T_c, T_o] \in \mathcal{C} \times \mathcal{O}} \kappa(K, T_c, T_o)$$

For each image in the total population of 12617 training images (Section 4.1), we determine the best parameter pairs $[T_c, T_o]$ and compute the estimated text skew angle. The probability distributions of the skew angle estimation error are plotted in Figure 1.

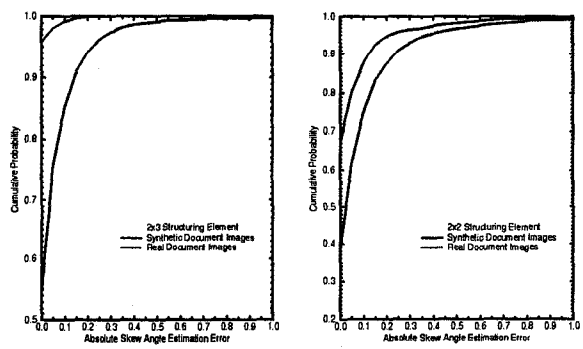


Figure 1: Illustrates the performance of the algorithm under the optimal parameter settings.

The algorithm exhibits worse performance on the real document images than the synthetic ones. There are two possible explanations for this. One is that our algorithm really performs worse on the real images. The other is the measurement uncertainty of



Proceedings of the Third International Conference on Document Analysis and Recognition (ICDAR '95)
0-8186-7128-9/95 $10.00 © 1995 IEEE

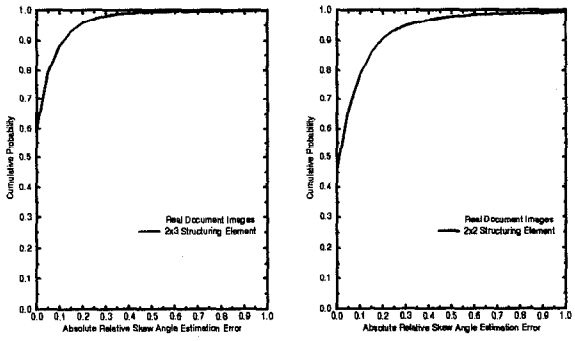


Figure 2: Illustrates estimation accuracy of the relative skew angles.

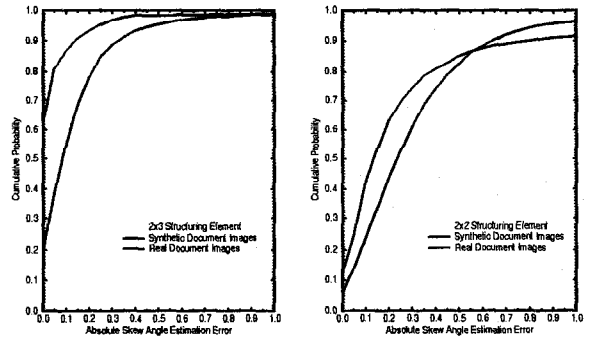


Figure 3: Illustrates the performance of the algorithm under the automatic mode.

the ground-truth text skew angles on the real images during the database preparation stage. To quantify the two possible effects, we adopt a methodology to measure the estimation accuracy of the relative skew angles.

The basic idea is the following: Given an input document image $I$, although it is impossible to know its true text skew angle, we can rotate the image by a known angle $\theta$, called the relative skew angle. In order for the text skew estimation algorithm to perform correctly, it is necessary that the estimated text skew angles on both the rotated image and the orignal image $I$ differ also by the same angle $\theta$. The difference between the true relative skew angle $\theta$ and the estimated relative skew angle $\hat{\theta}$ can be measured and characterized, which is independent of the true text skew angle of the input image $I$. Figure 2 summarizes the algorithm's estimation accuracy of the relative skew angles. By comparing Figure 2 with Figure 1, we discover that the main differences between the two sets of curves are in the interval $[0°, 0.2°]$, which is consistent with the fact that the measured standard deviations of the ground-truth text skew angle are typically in the range $[0.03°, 0.1°]$. From the differences, we see that the effect of the measurement uncertainty of the ground-truth text skew angles accounts for about 3% loss in the algorithm performance on real images in the specified interval.

### Performance in automatic mode

The experiment studies the performance of the text skew estimation algorithm with on-line optimal parameter setting prediction using the constructed regression trees.

To prepare the testing images, we rotated each image from the 1147 source image population by three random angles drawn from the uniform distribution $U[-5°, 5°]$. This constitutes a total of $3441 = 3 \times 1147$ testing images. Then for each image in the test image population, we obtained an estimate of the text skew angle through the automatic text skew estimation algorithm. Finally, the probability distributions of the estimation errors were computed, as shown Figure 3.

### Timing performance

To process an input document image of size $3300 \times 2550$, the algorithm takes almost a constant 10 seconds on SUN Sparc 10 machines.

## References


[1] H.S. Baird, "The Skew Angle of Printed Documents", *Proc. of SPIE Symposium on Hybrid Imaging Systems*, Rochester, N.Y. 1987, 21-24.

[2] A.L. Spitz, "Skew Determination in CCITT Group 4 Compressed Document Images", *Proc. of the first Annual Symposium on Document Analysis and Information Retrieval*, March 16-18, pp. 11-25, 1992.

[3] R.M. Haralick, S. Chen and T. Kanungo, "Recursive Opening Transform", *CVPR, Champaign*, June 1992.

[4] S. Chen and R.M. Haralick, "Recursive Erosion, Dilation, Opening and Closing Transforms", *IEEE Trans on Image Processing*, Vol.4, No. 3, March 1995.

[5] I.T. Phillips, S. Chen and R.M. Haralick, "English Document Database Standard", *ICDAR*, Japan, 1993.

[6] S. Chen, M.Y. Jaisimha, J. Ha, I.T. Phillips and R.M. Haralick, *Reference Manual*, 1993. UW English Document Image Database - (I) Manual.

[7] G. Vosselman, *Protocol on the performance of least squares line anc circle fitting* ISL Technical Report, University of Washington, May 1993.

[8] L. Breiman, J.H. Friedman, R.A. Olshen and C.J. Stone, *Classification and Regression Trees*, Wadsworth International Group, Belmont CA, 1984.

[9] S. Chen and R.M. Haralick, "An automatic algorithm for text skew estimation in document images using recursive morphological transforms", *ICIP*, Austin, November, 1994.




Proceedings of the Third International Conference on Document Analysis and Recognition (ICDAR '95)
0-8186-7128-9/95 $10.00 © 1995 IEEE