AN AUTOMATIC ALGORITHM FOR TEXT SKEW ESTIMATION IN DOCUMENT IMAGES USING RECURSIVE MORPHOLOGICAL TRANSFORMS

Su Chen and Robert M. Haralick

Department of Electrical Engineering
University of Washington
Seattle, Washington 98195
{chen, haralick}@george.ee.washington.edu

ABSTRACT

The text skew estimation algorithm utilizes recursive morphological transforms. With hand tuned parameters the algorithm produces estimated text skew angles which are within 0.1° of the true text skew angles 99% of the time.

We also developed methodology to allow the algorithm to determine the optimal algorithm parameter settings on the fly without any human interaction. Under this automatic mode, our experimental results indicate that the algorithm generates estimated text skew angles which are within 0.5° of the true text skew angles 99% of the time.

To process a 3300 × 2550 document image, the algorithm takes about 10 seconds on SUN Sparc 10 machines if discounting the document image file reading time.

1. INTRODUCTION

Document layout analysis systems usually rely upon images of unskewed texts, i.e. images of texts positioned in their up-right orientation. However, during the process of document image production (e.g. printing, photocopying, faxing and scan-digitization), the document images are often rotated by some angle. Texts on the document images are inevitably skewed. The presence of text skews may result in failure of the document layout analysis algorithms. Therefore, techniques have been developed to estimate the text skew angle given a document image [1] [2]. The estimated text skew angle can then be utilized to de-skew the text images.

Earlier work on detection of text skew is basically a two stage process [1] [2]. In the first stage, one detects the feature positions for the alignment. 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 them.

The text skew estimation algorithm in this paper is based upon our newly developed recursive 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, there is strong experimental evidence that demonstrates that the performance of our text skew estimation algorithm is at least as good as the results reported in [1] and [2], although our document image population is much larger.

2. PROBLEM STATEMENT

2.1. Image Coordinate System

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. The image rotation is specified with respect to this coordinate system. When the document image is rotated counter-clockwise, the text skew angle is positive.

2.2. Document Text Skew Angle

In real document images, there may exist multiple text baseline directions. The following gives our definition of the text skew angle:

Definition 1 *The* **text skew angle** *of a document image is denoted by φ 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.*

We constrain φ to be in the range of: $-\pi/2 < \varphi \leq \pi/2$.

2.3. Problem Statement of Skew Estimation

Let I denote a bi-level document image. Let φ and $\hat{\varphi}$ represent the true and the estimated document text skew angle.

Document Text Skew Estimation Problem:

Given an observed document image I. Find $\hat{\varphi}$ to optimally estimate the true document text skew angle φ.

3. A NEW ALGORITHM FOR TEXT SKEW ESTIMATION

Our new text skew estimation algorithm is a three stage process. In the first stage, directions are detected for each

0-8186-6950-0/94 $4.00 © 1994 IEEE

text line. In the second stage, the dominant detected directions are selected. These are assumed to correspond to the text baseline direction. Lastly, a Bayesian estimate of the document text skew angle is calculated based on the selected text baseline directions.

3.1. Document Image Model

Let $I \in Z^2$ represent a bi-level document image. Reasonable assumptions on the document image I are as follows:

- I is scan-digitized at a resolution of 300dpi.
- Binary one pixels are feature pixels (text, graphics, halftone); whereas the binary zero pixels are non-feature pixels (white space).
- I is a mixed mode document image: It may contain both textual and non-textual objects.
- Characters that constitute a text word share approximately a common baseline and a x-height
- There exists a dominant text baseline direction
- The intercharacter gaps within a word are smaller than the interline gaps.

3.2. Text Baseline Direction Detection

In this section, an algorithm for the detection of the text baseline direction is described. The algorithm first sub-samples the input document image and then detects the block areas that correspond to the x-heights. The detection is based on the recursive opening transform (RCT) and the recursive closing transforms (ROT) described in [3] [4]. The RCT and ROT are able to compute in constant time per pixel the binary morphological opening and closing with all sized structuring element simultaneously. Each of the detected block areas is modeled as a straight ribbon, which can be modeled by sweeping a disk of a constant radius along a straight line. In the final step, the ribbon is fitted by a straight line and its orientation is extracted. The various components of the text baseline direction detection algorithm are described in next:

Sub-Sampling

In the document image model, we assumed that a document is scan-digitized at a spatial resolution of 300dpi. A sampling algorithm is implemented to reduce the 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. The experimental results in Section 4.2 justify our strategy.

Filling the Inter-Character Gap

The recursive closing transform (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 2×2 box or a 2×3 box. In the first case, little assumptions are made on the range of the text skew angle. It could lie anywhere in the interval $(-\pi/2, \pi/2]$. Whereas in the second case, it assumes that the text skew angle is relatively small, e.g. in the range of $(-\pi/4, \pi/4]$. We study the performance of the algorithm utilizing different structuring elements through our experiments which are discussed in Section 4.

Removing Ascenders and Descenders

The recursive opening transform (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 prescribed 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 $K_o = K_c = K$.

Line Fitting

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 [7] is applied to minimize the sum of squared distance from the points to the fitted line. The line direction φ and its variance are functions of the second order spatial moments of the points:

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

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

3.3. Selection of "Good" Lines

The fitted lines may differ significantly in their orientations. 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 (such as figures, streaking noise and etc.) or portions of skewed text lines from other pages (this happens frequently when scan-digitizing a document from a bound books). Therefore, a procedure is required to select a "good" subset from all the detected lines so that the text skew estimation can be more robust.

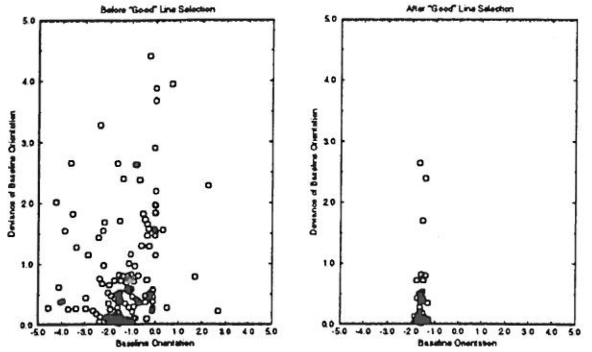

Figure 1: Illustrates the "good" line selection process. The figure plots the scatter plot of the detected text baseline orientation v.s. its standard deviation.

The task of "good" line selection can be translated into estimating the dominant text baseline direction and determining the subset of text baselines whose orientations are aligned with the estimated dominant text baseline direction.

Algorithm Description

In this section, we describe a robust iterative algorithm for "good" line selection. The algorithm first computes the median of the observed line orientations (denoted as μ_{med}) and the median of line orientation variances (denoted as σ^2_{med}). Then it forms a subset of lines whose orientations lie within an interval centered around μ_{med}. The size of the interval is proportional to the square root of σ^2_{med}. The selected lines are then re-input to the selection process. The selection process iterates and stops when both μ_{med} and σ^2_{med} converge.

Figure 1 shows the result of the "good" line selection process. The detected lines whose orientations are aligning with the dominant text baseline direction are selected.

3.4. Bayesian Text Skew Estimation:

After the "good" text baselines are selected, we can formulate the text skew estimation problem in a Bayesian framework. The text skew angle can be estimated based on the prior probability distribution of the text skew angle and the observed text baseline directions. Here is the problem statement:

Bayesian Text Skew Estimation Problem:

Given the prior probability distribution of the text skew angle and a set of selected "good" line orientations $l_i = (\hat{\varphi}_i, \hat{\sigma}^2_{\varphi_i})$, where $i = 1, 2, \cdots, L$. Find φ to maximize the joint probability distribution $P(\varphi, l_1, l_2, \cdots, l_L)$.

Bayesian Analysis

To simplify the analysis, we will make the following assumptions:

- The true text skew angle φ has a normal prior probability distribution: $P(\varphi) = \frac{1}{\sqrt{2\pi}\sigma_{\varphi_0}} e^{-\frac{(\varphi-\varphi_0)^2}{2\sigma^2_{\varphi_0}}}$.
- Each observed line orientation is normal given the true text skew angle φ: $P(\hat{\varphi}_i \mid \varphi) = \frac{1}{\sqrt{2\pi}\hat{\sigma}_{\varphi_i}} e^{-\frac{(\hat{\varphi}_i-\varphi)^2}{2\hat{\sigma}^2_{\varphi_i}}}$.

Under these assumptions, we can maximize the probability $P(\varphi, l_1, l_2, \cdots, l_L)$ and obtain the optimal estimate of the text skew angle $\hat{\varphi}$ and also the mean and variance of the estimator:

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

$$E[\hat{\varphi}] = \frac{\sum_{i=1}^{L} \omega_i}{\sum_{i=0}^{L} \omega_i}\varphi + \frac{\omega_0}{\sum_{i=0}^{L} \omega_i}\varphi_0$$

$$Var[\hat{\varphi}] = \frac{\sum_{i=1}^{L} \omega_i}{(\sum_{i=0}^{L} \omega_i)^2}$$

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

4. EXPERIMENTAL PROTOCOL

The prescribed skew estimation algorithm is not yet fully automatic because users need to set algorithm tuning parameters (such as the threshold values T_c and T_o). The selection of tuning parameters are dependent on the content of the input image and must be estimated on a per image basis.

Therefore, to equip the algorithm with the capability of determining the optimal tuning parameter setting on the fly and to evaluate its performance, we need an experimental protocol by which systematic experiments can be performed. Our experiments are conducted in two phases.

4.1. Phase I: Training the Algorithm

Image Population

Our image data set will be based on the "UW English Document Image Database (I)" [5] [6]. The database was developed at the University of Washington in 1993 and was 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.

The database 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 rest of the 168 images are noise-free images synthetically generated from a set of LaTeX documents.

Normally, real document images are skewed during the scanning process; while the synthetic image are without skew. The database provides with each real document image a ground-truth consisting of the mean and standard deviation of the text skew angle and the number of observations in the measurement. The ground-truth were provided by interactively specifying multiple sets of three points on a document image during the database preparation stage.

In our experiment, each of the document images in the database was additionally rotated at various degrees of $0°, \pm1°, \pm2°, \pm3°, \pm4°, \pm5°$, using a nearest neighbor interpolation algorithm. This becomes 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, i.e. the number of "good" text baseline directions. 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:

- Choose the structuring element K from: $\mathcal{K} = \{2 \times 2 \text{ Box}, 2 \times 3 \text{ Box}\}$.
- Choose the RCT threshold T_c from: $\mathcal{C} = \{3, 4, 5, 6, 7, 8\}$.
- Choose the ROT threshold T_o from: $\mathcal{O} = \{T_o \mid \mid T_o - T_c \mid \leq 2, T_c \in \mathcal{C}\}$.
- Choose $\sigma^2_{\varphi_0} = \infty$. The algorithm has no knowledge on the prior distribution of 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

In this section, we consider the criterion to measure the "goodness" of the output from the text skew estimation algorithm. The criterion quantifies the deviation of the estimated skew angle from the true text skew angle, denoted by . Therefore, it is used to determine the optimal algorithm tuning parameter settings.

For the synthetically generated document images, the true text skew angles are known. We use the squared distance 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. The text skew ground-truth provided by the database merely indicates a probability distribution of the true text skew angle.

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 μ_φ denote the true text skew angle. We define the goodness measure κ in the following:

Definition 2 *The* **goodness measure** *for the text skew estimation is denoted by κ 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 could compute κ 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 subsampled 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 or CART [8]. The constructed regression trees allow our text skew estimation algorithm to predict the optimal algorithm parameter settings on the fly.

4.2. Phase II: 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. Let $\hat{\mu}_{\varphi_A}$ denote the text skew angle predicted by the algorithm and let $\hat{\mu}_{\varphi_D}$ denote the ground-truth text skew angle provided by the database. 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 Parameter Setting 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 2.

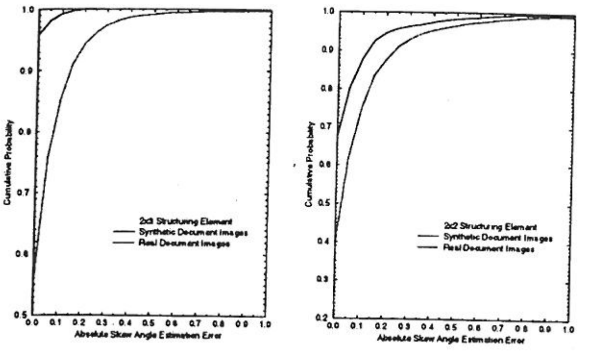

Figure 2: Illustrates the performance of the algorithm under the optimal parameter settings. It plots the probability distribution of the absolute skew estimation errors.

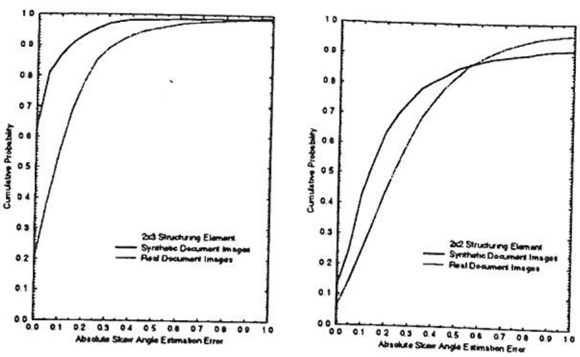

Figure 3: Illustrates the performance of the algorithm under the automatic mode. It plots the probability distribution of the absolute skew estimation errors.

The experimental results validated our approach to the skew estimation in document images. On the $1848 = 11 \times 168$ synthetic images, the algorithm detected text skew angles which were within 0.1° of the true text skew angles at a probability of about 99% when using the 2×3 box structuring element; while this probability drops to about 90% when using the 2×2 box structuring element. The reason behind the choice of either the 2×3 box structuring element or the 2×2 box structuring element was explained in Section 3.2. On the $10769 = 11 \times 979$ real images, the algorithm still exhibits very good performance. The probabilities that the estimated text skew angles lie within 0.1° of the ground-truth text skew angles are 86% and 75% while using the 2×3 and 2×2 box structuring elements respectively.

There are two possible explanations for this decrease in performance. One is that our algorithm performs worse on real images. The other is the measurement uncertainty of the ground-truth text skew angles on real images during the database preparation stage. The ground-truth text skew angle is no longer the true text skew angle, but rather an estimate. Our study indicates that the effect of the measurement uncertainty accounts for about 3% loss in the algorithm performance in the interval $[0^\circ, 0.2^\circ]$.

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 $U[-5^\circ, 5^\circ]$. 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 skew angle estimation errors were computed, as shown Figure 3.

On the $504 = 3 \times 168$ synthetic test images, the algorithm detected text skew angles which were within 0.5° of the true text skew angles at a probability of about 99% when using the 2×3 box structuring element; while this probability drops to about 86% when using the 2×2 box structuring element. On the $2937 = 3 \times 979$ real test images, the respective probabilities are 96% when using the 2×3 box structuring element and 84% when using the 2×2 box structuring element. Again, we could compensate the curves for the effect of the measurement uncertainty of the ground-truth text skew angles on real document images.

Timing Performance

The experiment studies the timing performance of the algorithm on SUN Sparc 10 machines. When the input document image size is 3300×2550 and if we discount the time needed for reading the image files from the hard-disk into the memory, the algorithm requires almost a constant 10 seconds in execution time under the automatic mode.

5. 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] S. Chen and R.M. Haralick, "Recursive Opening Transform", *CVPR, Champaign*, June 1992.

[4] S. Chen and R.M. Haralick, "Recursive Erosion, Dilation, Opening and Closing Transforms", to appear in *IEEE Trans on Image Processing*, Mar. 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.