# Entropy Based Skew Correction of Document Images

K.R. Arvind, Jayant Kumar, and A.G. Ramakrishnan

MILE Lab, Electrical Engineering,
Indian Institute of Science, Bangalore, India 560012

**Abstract.** The document images that are fed into an Optical Character Recognition system, might be skewed. This could be due to improper feeding of the document into the scanner or may be due to a faulty scanner. In this paper, we propose a skew detection and correction method for document images. We make use of the inherent randomness in the Horizontal Projection profiles of a text block image, as the skew of the image varies. The proposed algorithm has proved to be very robust and time efficient. The entire process takes less than a second on a 2.4 GHz Pentium IV PC.

## 1 Introduction

An optical character recognition system takes in the document image generated by a scanner, segments it into words and then to characters, then finally recognizes using some feature based classifier. Although, this procedure seems simple enough, improper feeding of the document into the scanner could produce a skew. This skew could hamper the word/character level segmentation of the document and hence drastically affect the performance the recognition system. Hence we see that, before the processing of the document image, a robust skew detection and correction mechanism is extremely important.

Xiaoyan Zhu and Xiaoxin Yin[1] have proposed a method to correct skew of document images containing both text and non-text. They first divide the image into blocks. Then they classify the blocks into Text/Non-Text using Fourier transform the projection profiles of the blocks as features and Support Vector Machine as the classifier. They determine the skew only for the text blocks by taking the standard deviation of the projection profiles for various angles. They conclude that the angle at which the standard deviation is maximum, is the skew angle of the document image.

Le et al.[2] select a square region dominated by text from the document image and calculate the skew angle by this area. Avanindra and Subhasis Chaudhuri[3] divide the document into blocks and use the median of the cross-correlations of all blocks to determine the skew angle.

Bruno and Rafael[4] have used nearest neighbor clustering approach for skew detection in document images. It works with complex documents and has a range angle detection of 0 to 360 with precision of 0.1. The algorithm starts by boxing each block of black pixels by using component labeling. Applying the least square method to the middle top point and middle bottom point separately of all blocks of the text-line, it forms two lines located at strategic points. The bottom line is located at the baseline of the text-line and crosses the descending characters. The top line is located above the text-line and crosses the ascending characters. This happens because there are more non-salient characters (e.g. a, c, e, o and u) in the Latin alphabet than ascending and descending ones. If the number of descending strokes is greater than the ascending one then the document is upside-down.

Marisa E. Morita et al.[5] have proposed a morphology-based method to detect and correct handwritten word skew in the treatment of dates written on bank checks. Their aim is to limit the number of parameters and heuristic features necessary for a good skew correction. Their approach is based on the morphological pseudo-convex hull.

In this paper, we propose an entropy based skew correction method. In section 2, we describe our method in detail. In section 3, we discuss the experimental results. Our method can handle document images having multiple skewed blocks. It is time efficient, consuming less than half a second for both skew detection and correction on a block image.

## 2 System Description

The system is divided into three steps. In the initial step, the document image is cleaned of all noise elements such as spurious dots and lines. Next, it is segmented into its constituent blocks. This is carried out similar to the block segmentation method given by Arvind *et. al.*[6]. Then, the blocks are skew corrected based on their horizontal projection profiles and entropy. The procedure is summarized below:

- Clean the document
- Extract the blocks
- Skew correct the blocks

### 2.1 Noise Removal

We apply Connected Component Analysis (CCA) and obtain the number of ON pixels and aspect ratios for every component. Then, we find the minimum and maximum values of them. Let them be *minp*, *maxp* and *mina*, *maxa* respectively.

$$TP = \frac{np - minp}{maxp - minp} \tag{1}$$

$$TA = \frac{na - mina}{maxa - mina} \tag{2}$$

-where *np* is number of on pixels in the component and *na* is aspect ratio of the component

If TP or TA is less than 0.002, which is computed empirically, then we remove it. We need to remove extraneous dots and especially vertical lines as they could hamper the horizontal projection profiles(HPP) and thereby reduce the effectiveness of the skew correction process.

### 2.2 Block Segmentation

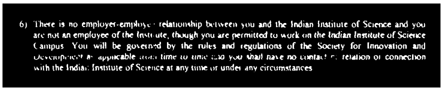
Arvind *et. al.*[6] have segmented blocks by run-length smoothening the image with the parameter selected such that the intra and inter character gaps, up to a paragraph level, are filled. Then, they apply morphological erosion operator to remove thin joints between blocks. Finally CCA is used to segregate the blocks.

### 2.3 Skew Detection and Correction

**Horizontal Projection Profile.** HPP of an image is a vector where each vector element contains the sum of the pixel values in the corresponding row.

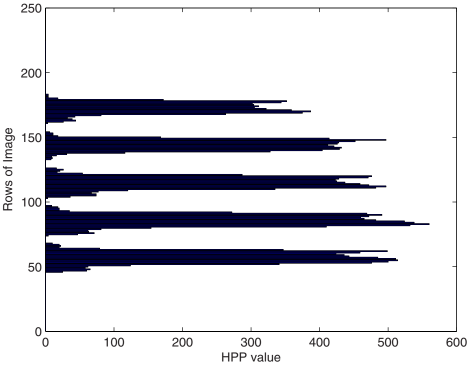
$$HPP(i) = \sum_{j=1}^{No.of columns} I(i,j) \tag{3}$$

where $I(i,j)$ is the image.



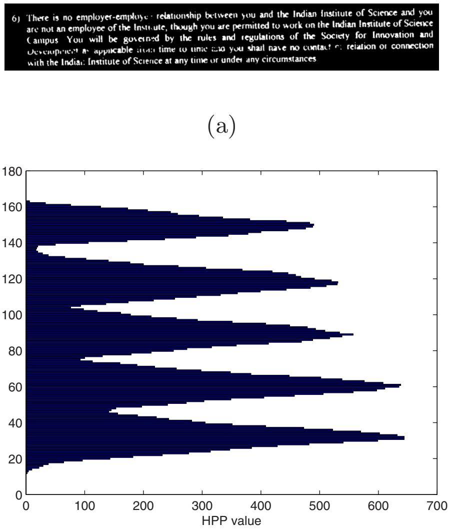
6) There is no employer-employee relationship between you and the Indian Institute of Science and you are not an employee of the Institute, though you are permitted to work on the Indian Institute of Science Campus You will be governed by the rules and regulations of the Society for Innovation and Development as applicable from time to time and you shall have no contact or relation or connection with the Indian Institute of Science at any time or under any circumstances

(a)



(b)

**Fig. 1.** (a) Example of a typical correctly oriented block. (b) HPP of the correctly oriented block.



6) There is no employer-employee relationship between you and the Indian Institute of Science and you are not an employee of the Institute, though you are permitted to work on the Indian Institute of Science Campus You will be governed by the rules and regulations of the Society for Innovation and Development as applicable from time to time and you shall have no contact or relation or connection with the Indian Institute of Science at any time or under any circumstances

(a)

(b)

**Fig. 2.** (a) Example of a typically skewed block (b) HPP of the skewed block

Figure 1(a) shows a typical block segmented from a document image and Figure 1(b) its horizontal projection profile(HPP). Figures 2(a) and 2(b) depict the skewed block and its HPP respectively. We observe that the HPPs of both the properly oriented block and the skewed block seem to have a similar pattern of increase and decrease in the HPP. The HPP values actually repeat in case of a skew corrected block. There exists a repetitive pattern as compared to the HPP of a skewed block. Hence we see that the randomness associated with the projection profiles increases as the skew of the block varies. We use this property to detect the skew of a particular block.

**Entropy.** The entropy is a statistical measure of randomness. It is given by equation 4.

$$E(i) = \sum_{i} -HPP(i) * log(HPP(i)) \tag{4}$$

The entropy associated with HPP of objects that are repetitive in nature would be lesser compared to the HPP of random objects such as graphic images or skewed text blocks. i.e. as the randomness increases, the entropy also increases. Assuming that the maximum skew would not be greater than 10 degrees, we rotate the image upto ±10 degrees and obtain the HPPs along with their corresponding entropy values. The resolution of rotation is 1 degree. The angle for which the entropy is minimum, is the rough skew angle of the segmented block. We have taken block image displayed in Figure 4(a) and run the proposed algorithm to correct its skew. Table 1 displays the rotated angles and its corresponding entropy values. It can be observed from the table, that the angle for which entropy value is minimum i.e.1 degree is the coarse skew angle.

**Table 1.** The angles through which the blocks are rotated and their corresponding entropy values

| Angle | Entropy | Angle | Entropy |
|---|---|---|---|
| -10 | 5.64 | 10 | 5.44 |
| -9 | 5.59 | 9 | 5.37 |
| -8 | 5.53 | 8 | 5.30 |
| -7 | 5.46 | 7 | 5.23 |
| -6 | 5.40 | 6 | 5.17 |
| -5 | 5.33 | 5 | 5.11 |
| -4 | 5.25 | 4 | 5.05 |
| -3 | 5.20 | 3 | 4.98 |
| -2 | 5.11 | 2 | 4.87 |
| -1 | 5.03 | 1 | 4.35 |
| 0 | 4.78 | 0 | 4.78 |

After obtaining the coarse skew angle, we do a finer skew angle detection by obtaining the entropy values of the HPP of the block image rotated through ±1 degree with a resolution of 0.1 degrees. Table 2 depicts entropy values obtained during the finer skew angle detection process.

**Table 2.** The entropy values obtained for fine skew angle detection

| Angle | Entropy | Angle | Entropy |
|---|---|---|---|
| 0 | 4.74 | 1.1 | 4.402 |
| 0.1 | 4.73 | 1.2 | 4.45 |
| 0.2 | 4.66 | 1.3 | 4.50 |
| 0.3 | 4.61 | 1.4 | 4.56 |
| 0.4 | 4.55 | 1.5 | 4.62 |
| 0.5 | 4.49 | 1.6 | 4.68 |
| 0.6 | 4.44 | 1.7 | 4.74 |
| 0.7 | 4.38 | 1.8 | 4.79 |
| 0.8 | 4.34 | 1.9 | 4.84 |
| 0.9 | 4.33 | 2 | 4.88 |
| 1 | 4.36 | | |

The minimum entropy value is for 0.9 degrees. After obtaining the exact skew angle, the image is rotated by that angle in the opposite direction using bilinear interpolation. Thus the block is skew corrected. Although the skew detection and correction could be done on the entire document image, it is done at block level since every block with its logos, typing defects etc. could have a different skew. After skew correction, further processing such as classification and OCR could be conducted on the block.

## 3 Experimental Results

### 3.1 Data Description

Our data consists of 100 document images each of English and Kannada, scanned at 200 dpi and stored in 1-bit depth monochrome format. These documents contain signatures, logos and other such things along with free-flowing text paragraphs.

### 3.2 Implementation

The algorithm has been implemented in ANSI-C language and compiled using GCC compiler. It has been executed on a Pentium IV 2.4 GHz, 512 MB RAM PC.

### 3.3 Timing Analysis

Table 3 depicts the timing for various stages in the skew correction process for 100 block images consisting of both English and Kannada images.

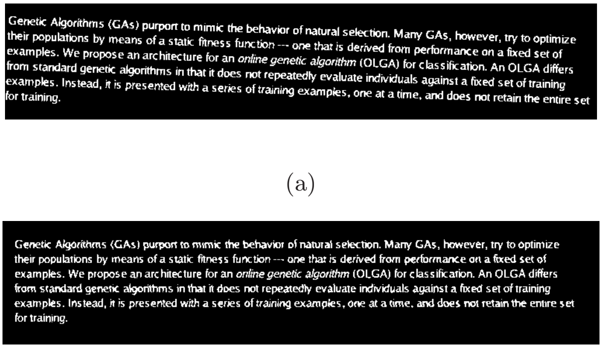
**Table 3.** The Detection and Correction time for 100 block images

| | |
|---:|:---:|
| Total Detection Time | 10.57s |
| Total Correction Time | 33.05s |
| Total Time | 43.62s |
| Average Time | 0.44s |

### 3.4 Time Complexity

During skew detection the entropy calculation is done for a constant number of times (20 in our case). We find the minimum of all the entropy values calculated. Since finding the minimum value among all entropy values, involves a constant number of comparisons, the time complexity of our algorithm mainly depends on the time complexity of entropy calculation. Entropy calculation for each angle involves iteratively checking each pixel in the image. Also, we use bilinear interpolation for rotating the image for constant number of times. Hence the overall, time complexity is O(height*width), which is linear with respect to total number of pixels in the image.

Figure 3 and 4 depict some of the results we have obtained for English and Kannada block images.



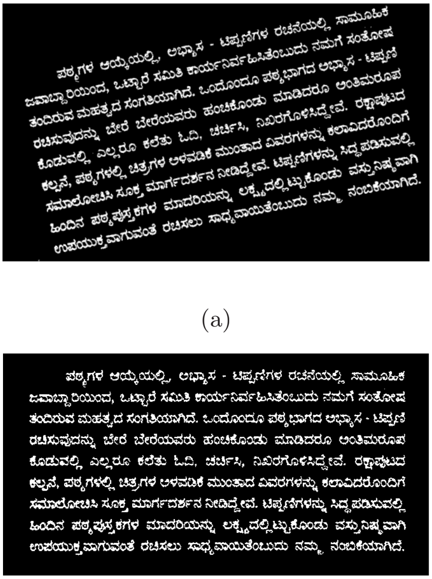
Genetic Algorithms (GAs) purport to mimic the behavior of natural selection. Many GAs, however, try to optimize
their populations by means of a static fitness function --- one that is derived from performance on a fixed set of
examples. We propose an architecture for an *online genetic algorithm* (OLGA) for classification. An OLGA differs
from standard genetic algorithms in that it does not repeatedly evaluate individuals against a fixed set of training
examples. Instead, it is presented with a series of training examples, one at a time, and does not retain the entire set
for training.

(a)

Genetic Algorithms (GAs) purport to mimic the behavior of natural selection. Many GAs, however, try to optimize
their populations by means of a static fitness function --- one that is derived from performance on a fixed set of
examples. We propose an architecture for an *online genetic algorithm* (OLGA) for classification. An OLGA differs
from standard genetic algorithms in that it does not repeatedly evaluate individuals against a fixed set of training
examples. Instead, it is presented with a series of training examples, one at a time, and does not retain the entire set
for training.

(b)

**Fig. 3.** (a) Example of a skewed English block image (b)The skew corrected block image



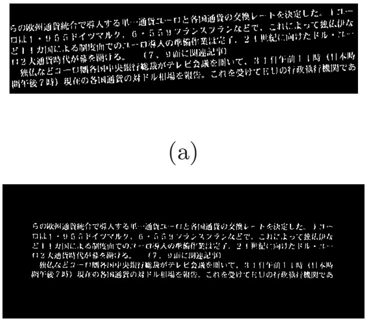
ಪಠ್ಯಗಳ ಆಯ್ಕೆಯಲ್ಲಿ, ಅಭ್ಯಾಸ - ಟಿಪ್ಪಣಿಗಳ ರಚನೆಯಲ್ಲಿ ಸಾಮೂಹಿಕ
ಜವಾಬ್ದಾರಿಯಿಂದ, ಒಟ್ಟಾರೆ ಸಮಿತಿ ಕಾರ್ಯನಿರ್ವಹಿಸಿತೆಂಬುದು ನಮಗೆ ಸಂತೋಷ
ತಂದಿರುವ ಮಹತ್ವದ ಸಂಗತಿಯಾಗಿದೆ. ಒಂದೊಂದೂ ಪಠ್ಯಭಾಗದ ಅಭ್ಯಾಸ - ಟಿಪ್ಪಣಿ
ರಚಿಸುವುದನ್ನು ಬೇರೆ ಬೇರೆಯವರು ಹಂಚಿಕೊಂಡು ಮಾಡಿದರೂ ಅಂತಿಮರೂಪ
ಕೊಡುವಲ್ಲಿ ಎಲ್ಲರೂ ಕಲೆತು ಓದಿ, ಚರ್ಚಿಸಿ, ನಿಖರಗೊಳಿಸಿದ್ದೇವೆ. ರಕ್ಷಾಪುಟದ
ಕಲ್ಪನೆ, ಪಠ್ಯಗಳಲ್ಲಿ ಚಿತ್ರಗಳ ಅಳವಡಿಕೆ ಮುಂತಾದ ವಿವರಗಳನ್ನು ಕಲಾವಿದರೊಂದಿಗೆ
ಸಮಾಲೋಚಿಸಿ ಸೂಕ್ತ ಮಾರ್ಗದರ್ಶನ ನೀಡಿದ್ದೇವೆ. ಟಿಪ್ಪಣಿಗಳನ್ನು ಸಿದ್ಧಪಡಿಸುವಲ್ಲಿ
ಹಿಂದಿನ ಪಠ್ಯಪುಸ್ತಕಗಳ ಮಾದರಿಯನ್ನು ಲಕ್ಷ್ಯದಲ್ಲಿಟ್ಟುಕೊಂಡು ವಸ್ತುನಿಷ್ಠವಾಗಿ
ಉಪಯುಕ್ತವಾಗುವಂತೆ ರಚಿಸಲು ಸಾಧ್ಯವಾಯಿತೆಂಬುದು ನಮ್ಮ ನಂಬಿಕೆಯಾಗಿದೆ.

(a)

ಪಠ್ಯಗಳ ಆಯ್ಕೆಯಲ್ಲಿ, ಅಭ್ಯಾಸ - ಟಿಪ್ಪಣಿಗಳ ರಚನೆಯಲ್ಲಿ ಸಾಮೂಹಿಕ
ಜವಾಬ್ದಾರಿಯಿಂದ, ಒಟ್ಟಾರೆ ಸಮಿತಿ ಕಾರ್ಯನಿರ್ವಹಿಸಿತೆಂಬುದು ನಮಗೆ ಸಂತೋಷ
ತಂದಿರುವ ಮಹತ್ವದ ಸಂಗತಿಯಾಗಿದೆ. ಒಂದೊಂದೂ ಪಠ್ಯಭಾಗದ ಅಭ್ಯಾಸ - ಟಿಪ್ಪಣಿ
ರಚಿಸುವುದನ್ನು ಬೇರೆ ಬೇರೆಯವರು ಹಂಚಿಕೊಂಡು ಮಾಡಿದರೂ ಅಂತಿಮರೂಪ
ಕೊಡುವಲ್ಲಿ ಎಲ್ಲರೂ ಕಲೆತು ಓದಿ, ಚರ್ಚಿಸಿ, ನಿಖರಗೊಳಿಸಿದ್ದೇವೆ. ರಕ್ಷಾಪುಟದ
ಕಲ್ಪನೆ, ಪಠ್ಯಗಳಲ್ಲಿ ಚಿತ್ರಗಳ ಅಳವಡಿಕೆ ಮುಂತಾದ ವಿವರಗಳನ್ನು ಕಲಾವಿದರೊಂದಿಗೆ
ಸಮಾಲೋಚಿಸಿ ಸೂಕ್ತ ಮಾರ್ಗದರ್ಶನ ನೀಡಿದ್ದೇವೆ. ಟಿಪ್ಪಣಿಗಳನ್ನು ಸಿದ್ಧಪಡಿಸುವಲ್ಲಿ
ಹಿಂದಿನ ಪಠ್ಯಪುಸ್ತಕಗಳ ಮಾದರಿಯನ್ನು ಲಕ್ಷ್ಯದಲ್ಲಿಟ್ಟುಕೊಂಡು ವಸ್ತುನಿಷ್ಠವಾಗಿ
ಉಪಯುಕ್ತವಾಗುವಂತೆ ರಚಿಸಲು ಸಾಧ್ಯವಾಯಿತೆಂಬುದು ನಮ್ಮ ನಂಬಿಕೆಯಾಗಿದೆ.

(b)

**Fig. 4.** (a) Example of a skewed Kannada block image (b) The skew corrected block image



らの欧州通貨統合で導入する単一通貨ユーロと各国通貨の交換レートを決定した。１ユー
ロは１・９５５ドイツマルク、６・５５９フランスフランなどで、これによって独仏伊な
ど１１カ国による制度面でのユーロ導入の準備作業は完了。２１世紀に向けたドル・ユー
ロ２大通貨時代が幕を開ける。　（７、９面に関連記事）
　独仏などユーロ圏各国中央銀行総裁がテレビ会議を開いて、３１日午前１１時（日本時
間午後７時）現在の各国通貨の対ドル相場を報告。これを受けてＥＵの行政執行機関であ

(a)

らの欧州通貨統合で導入する単一通貨ユーロと各国通貨の交換レートを決定した。１ユー
ロは１・９５５ドイツマルク、６・５５９フランスフランなどで、これによって独仏伊な
ど１１カ国による制度面でのユーロ導入の準備作業は完了。２１世紀に向けたドル・ユー
ロ２大通貨時代が幕を開ける。　（７、９面に関連記事）
　独仏などユーロ圏各国中央銀行総裁がテレビ会議を開いて、３１日午前１１時（日本時
間午後７時）現在の各国通貨の対ドル相場を報告。これを受けてＥＵの行政執行機関であ

(b)

**Fig. 5.** (a) Skewed image of block with Chinese script (b) The skew corrected image

# 4 Conclusion

Thus we propose a robust and efficient skew detection and correction algorithm based on Horizontal Projection Profiles of a block image and their entropy. In this paper we have shown a range angle detection of ±10 degrees with a precision of 0.1 degrees. Extension to greater angles is possible and is a simple procedure, the drawback being the increase in time consumption. We have tested our algorithm on 100 document images (English and Kannada) i.e. 4421 block images. We have also tested our algorithm on few Chinese documents as shown in Figure 5 and we have observed that the algorithm performs well, independent of script.

# References

1. Zhu, X., Yin, X.: A New Textual/Non- textual Classifier for Document Skew Correction. In: Proceedings of the 16th International Conference on Pattern Recognition (ICPR 2002) (2002)
2. Le, D.S., Thoma, G.R., Wechsler, H.: Automatic page orientation and skew angle detection or binary document images. Pattern Recognition 27
3. Avanindra, Chaudhuri, S.: Robust Detection of Skew in Document Images. IEEE Trans on Image Processing 6, 344–349 (1997)
4. Avila, B.T., Lins, R.D.: A Fast Orientation and Skew Detection Algorithm for Monochromatic Document Images. In: Proceedings of the 2005 ACM symposium on Document engineering (2005)
5. Morita, M.E., Bortolozzi, F., Facon, J., Sabourin, R.: Morphological approach of handwritten word skew correction. In: Proceedings of International Symposium on Computer Graphics, Image Processing and Vision (SIBGRAPI 1998)
6. Arvind, K.R., Pati, P.B., Ramakrishnan, A.G.: Automatic text block seperation in document images. In: Proceedings of 4th International Conference on Intelligent Sensing and Information Processing (ICSIP 2006) (2006)