A Robust And Fast Skew Detection Algorithm for Generic Documents

Bin Yu[1] *and Anil K. Jain*
Department of Computer Science, Michigan State University
East Lansing, MI 48824-1027
binyu@cps.msu.edu, jain@cps.msu.edu

Abstract

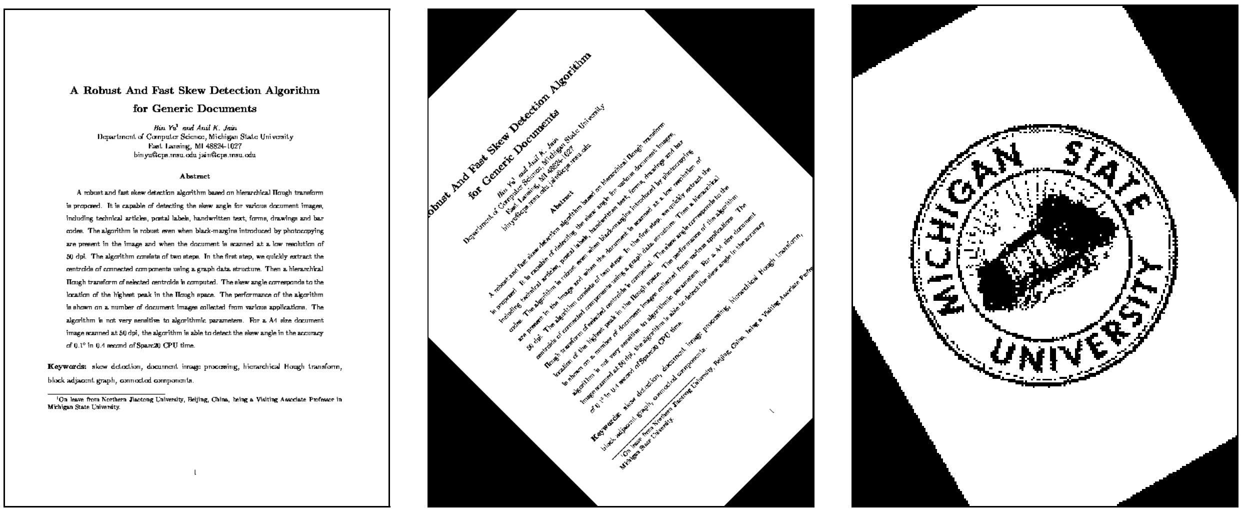
A robust and fast skew detection algorithm based on angular hierarchical Hough transform is proposed. It is capable of detecting the skew angle for various document images, including technical articles, postal labels, handwritten text, forms, drawings and bar codes. The algorithm is robust even when black-margins introduced by photocopying are present in the image and when the document is scanned at a low resolution of 50 dpi. The algorithm consists of two steps. In the first step, we quickly extract the centroids of connected components using a graph data structure. Then, an angular hierarchical Hough transform of selected centroids is computed. The skew angle corresponds to the location of the highest peak in the Hough space. The performance of the algorithm is shown on a number of document images collected from various applications. The algorithm is not very sensitive to algorithmic parameters. For an A4 size document image scanned at 50 dpi (typically 413×575 pixels), our algorithm is able to detect the skew angle with an accuracy of 0.1° in 0.4 seconds of CPU time on a Sun Sparc 20 workstation.

Keywords: skew detection, document image processing, angular hierarchical Hough transform, block adjacent graph, connected components.

[1]On leave from Northern Jiaotong University, Beijing, China.

1 Introduction

A text line in a document is defined as a group of characters, symbols, and words that are adjacent, relatively close to each other, and through which a straight line can be drawn. The dominant orientation shared by major text lines in a document determines the skew angle of that document. A document with a zero skew angle should have horizontally or vertically printed text lines which are parallel to the respective edges of the paper. Figures1(a) and (b) show synthetic document images with zero skew angle and a skew angle of 45.0°, respectively. Nonzero skew may be introduced when a document is scanned or photocopied. Many of the important document analysis algorithms, including optical character recognition (OCR) and page layout analysis, are sensitive to the orientation (or skew) of the input document image. With the rapid growth of the document entry and interpretation systems, it is important to develop algorithms to perform skew detection and correction automatically. In this paper we are concerned with situations where the skew in the document image is due to a rigid rotation of the document (Fig.1(b)). The more general problem of skew detection for document in which the text is not aligned along straight lines (Fig.1(c)) is beyond the scope of this paper.



(a) (b) (c)

Figure 1: A synthetic document image with (a) zero skew angle, (b) a skew angle of 45.0°, (c) a document image where text is not aligned along a straight line.

A generic document of interest in this paper can be a machine printed or handwritten text on paper or envelope, a page of technical article composed of text lines of different fonts, bar codes, drawings, figures and tables, a form and a photocopied document with black margins and noise.

Several attempts to detect the skew angle in a document image have been reported. Most of the popular methods use projection profile [1] and its variants [2, 3, 4, 5]. These methods usually work well when the skew angle is less than $\pm 10^\circ$ in the case of text only. Another approach uses k-NN (nearest neighbor) clustering [6, 7] of the connected components. This approach has a relatively high accuracy and a large computational cost [8], typically $\mathcal{O}(N^2)$, where N is the number of connected components. Hough transform has also been used for skew detection [9, 10, 11]. Despite its high accuracy, Hough transform approach is typically slower than the non-iterative projection profile methods and its computational cost greatly depends on the desired accuracy and detection range of skew angle.

The major computational cost in skew detection methods primarily involves (i) locating the text lines and (ii) skew angle estimation using projection, k-NN clustering or Hough transform. Approaches based on morphological transforms [12], text baseline extraction [13], connected component extraction [2, 11] and burst image creation from black run length [10] have been proposed for locating text lines.

Most previous skew detection methods work well for documents containing printed text-only and fail in situations where the text lines are not in majority, e.g. in a generic document described above. Without recognizing character and/or understanding the document, humans usually determine document orientation by collecting some regularly aligned symbols, mostly text. A symbol line is defined as a group of regularly aligned symbols that are adjacent, relatively close to each other, and through which a straight line can be drawn. The symbols can be individual machine printed or handwritten characters with various fonts and sizes, small groups of attached characters (which can be due to low scanning resolution), stripes in a bar code, layout token, etc. In our algorithm, each symbol is a connected component.

The skew detection algorithm presented in this paper consists of two steps. The first step quickly extracts the centroids of connected components based on a graph data structure, called block adjacency graph (BAG) [14] and the selection of components based on the size of their bounding boxes. The second step is the skew detection by an angular hierarchical Hough transform (AHHT); the skew angle corresponds to the highest peak in the Hough space. A block diagram of the algorithm is shown in the Fig.2.

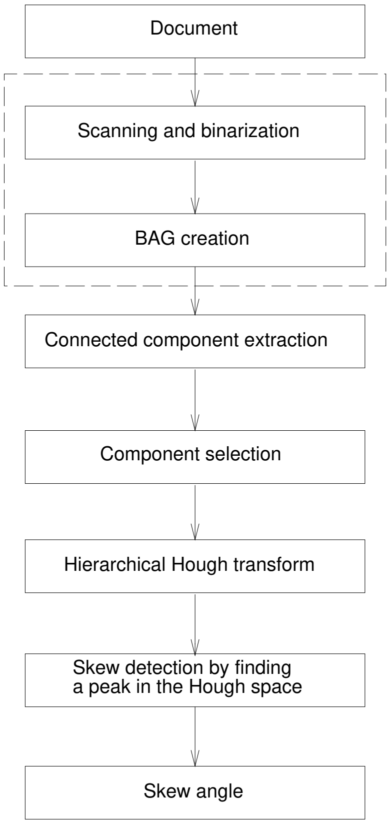


Figure 2: Block diagram of the algorithm (the steps bounded by the dashed box can be implemented concurrently).

The proposed algorithm has been evaluated on a large number of document images, including technical articles from journals, text on address labels on envelopes or postal cards, handwritten text, bar codes and forms. The algorithm is relatively fast and robust for detecting the skew in a generic document.

We will discuss the extraction and selection of connected components in Section 2 and explain the AHHT in Section 3. In Section 4 a number of test images and the detected skew angles are given to show the performance of the new algorithm. Section 5 contains our conclusions and suggestions for improving the algorithm.

2 Extracting Centroids of Connected Components

The information needed to determine the document skew is one or more symbol lines, each consisting of several separately aligned symbols. If a group of symbols is closely aligned along a line then their centroids are also roughly aligned along a line with the same direction. The centroid, therefore, is regarded as the feature for skew detection.

Yu et al. [14] described the block adjacency graph (BAG) which was used to extract the contours of the connected components in a binary image. This data structure can be created concurrently when the document is being scanned row by row. Processing steps such as contour extraction, vectorization and feature extraction are efficiently implemented based on this data structure [15].

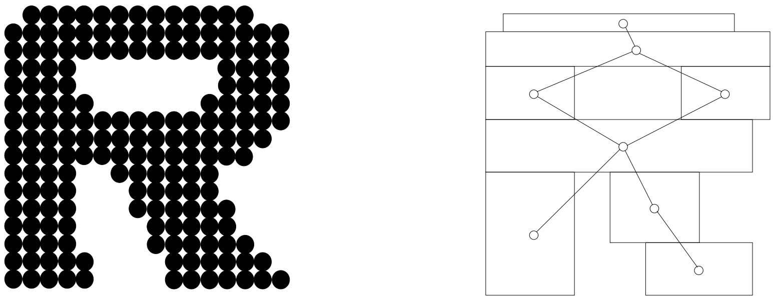


Figure 3: A connected component and its BAG.

Figure3 shows a connected component and its BAG which consists of eight connected blocks. Note that each block is an approximate bounding box of a partial area in the component and each area involves either one or a series of runs connected and closely justified in margin, given a tolerance. It is easy to obtain the centroid coordinates of the connected component by searching the graph and averaging the centroids of the blocks connected in the BAG. Since Hough transform is relatively time consuming, instead of pixels or runs, we apply it on the set of centroids of connected components, which greatly reduces the computational cost. Since symbol boundaries are actually thick lines of sparse density, using centroids of connected components instead of using pixels or runs as features will increase the accuracy of skew detection. Furthermore, some connected components are filtered out with respect to the size of their bounding boxes, which are introduced due to noise, black margins, lines in tables or forms, figures, etc. Therefore, only useful data are fed to the

Hough transform. Figure4 shows the intermediate results of our algorithm. Here, Fig.4(a) shows an input document with skew, 4(b) is the corresponding binary image, 4(c) is the BAG data structure, 4(d) contains the connected components extracted based on the BAG, 4(e) shows selected connected components and their centroids which are fed to the Hough transform, and 4(f) shows the deskewed image.

3 Skew Detection by An Angular Hierarchical Hough Transform

Hough transform [16] is a well known technique for detecting lines and curves in an image. For line detection, a set of points in Cartesian space $(x - y)$ are mapped to sinusoidal curves in Hough space $(\rho - \theta)$ via the following transform:

$$\rho = x \cos\theta + y \sin\theta. \tag{1}$$

Each time a sinusoidal curve intersects another curve at a particular value of ρ and θ, the likelihood that a line corresponding to that (ρ, θ) coordinate value is present in the original image is increased by using an accumulator array $\mathcal{H}(\rho, \theta)$. The computational steps in the basic Hough transform (BHT) are as follows:

```
For (all points (x, y)) {
    For (θ1 ≤ θ < θ2) {
        calculate ρ = x cos θ + y sin θ
        increment H(ρ, θ)
    }
}
Find the maximum in the array H(ρ, θ).
```

The computational cost of the BHT is $\mathcal{O}(N\Psi)$, where N is the number of points in the Cartesian space and Ψ is the number of different values of θ used in constructing the accumulator array. Note that $\Psi = \Delta\theta/\delta\theta$, where $\Delta\theta = \theta_2 - \theta_1$, $0° \leq \theta_1, \theta_2 \leq 180°$ and

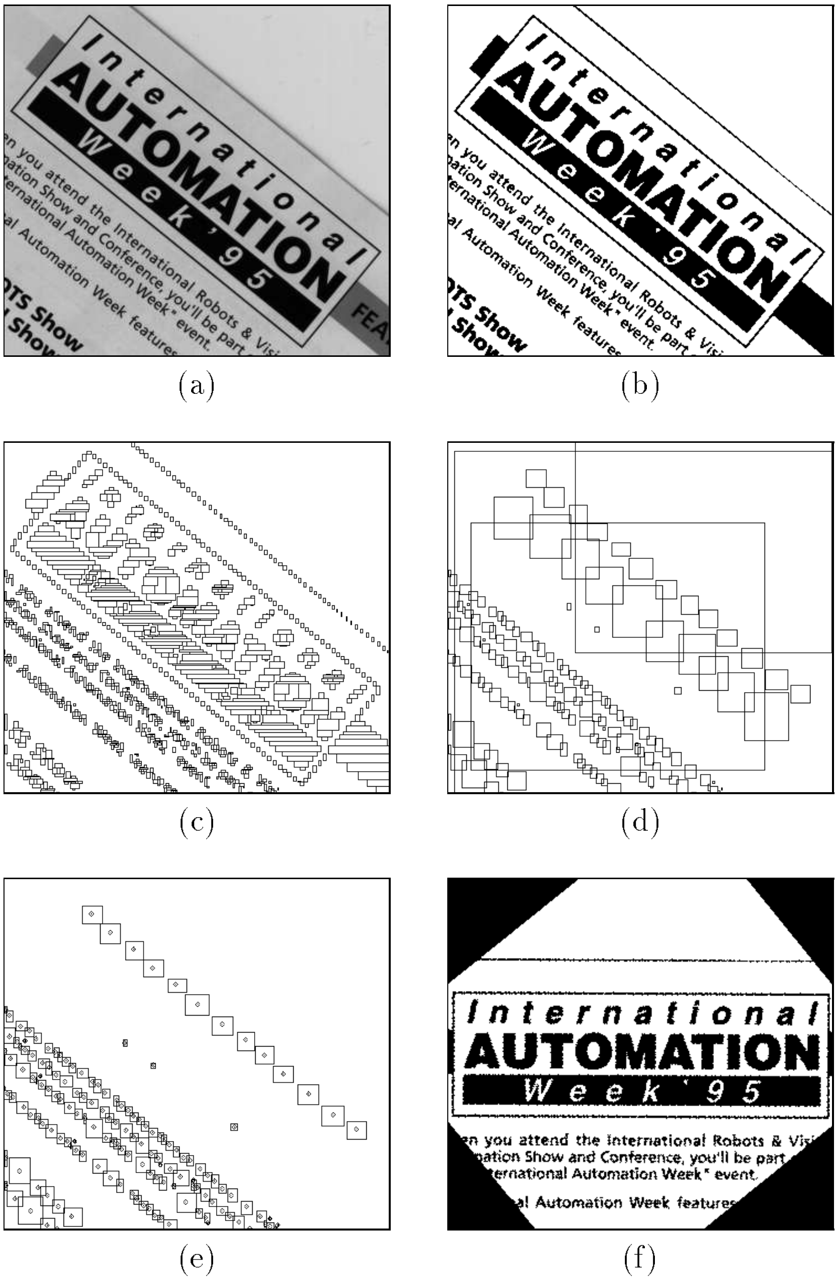


Figure 4: Intermediate results in the skew detection algorithm: (a) a grey scale document image with skew; (b) binary image of (a); (c) BAG; (d) connected components; (e) selected components and their centroids which are fed to the Hough transform; (f) the skew-corrected image after skew detection.

$\theta_1 < \theta_2$. The parameter $\delta\theta$ is the angular resolution which controls the accuracy of the skew detection (orientation of the line). For computational reasons, most skew detection algorithms using the Hough transform limit themselves to only a subset of $(0°, 180°)$ [8, 10].

The computational cost of the Hough transform can be decreased by reducing the number of points (x, y) and $\Delta\theta$ or increasing the interval $\delta\theta$. Initially, we can select a relatively large value of $\delta\theta$, say, $\delta\theta_1$ to approximately determine the skew angle $\hat{\theta}$ in the desired angular range $\Delta\theta$. Then, we can implement the BHT again with the desired angular resolution $\delta\theta_0$ within the reduced range $(\hat{\theta} - \delta\theta_1, \hat{\theta} + \delta\theta_1), \delta\theta_0 < \delta\theta_1$. We call this method the angular hierarchical Hough transform (AHHT). The computational complexity of AHHT is $\mathcal{O}(N\Psi_1) + \mathcal{O}(N\Psi_2) = \mathcal{O}(N\Psi')$, where

$$\Psi' = \Psi_1 + \Psi_2 = \frac{2\delta\theta_1}{\delta\theta_0} + \frac{\Delta\theta}{\delta\theta_1}. \tag{2}$$

To minimize the value of Ψ', we take the derivative of Eq.(2),

$$\frac{d\Psi'}{d\delta\theta_1} = 0.$$

Solving for $\delta\theta_1$, we get

$$\delta\theta_1 = \sqrt{\frac{\Delta\theta\delta\theta_0}{2}} \tag{3}$$

and

$$\Psi' = 2\sqrt{\frac{2\Delta\theta}{\delta\theta_0}}.$$

Initially,

$$\Psi = \frac{\Delta\theta}{\delta\theta_0},$$

therefore,

$$\frac{\Psi'}{\Psi} = 2\sqrt{\frac{2\delta\theta_0}{\Delta\theta}}.$$

Equation (3) helps us determine the optimum value of $\delta\theta_1$ for given values of $\Delta\theta$ and $\delta\theta_0$. Typically, $\Delta\theta = 180°$ and $0.05° \leq \delta\theta_0 \leq 0.5°$. If $\delta\theta_0 = 0.1°$, then from Eq.(3) we have $\delta\theta_1 = 3°$ and $\Psi'/\Psi = 0.07$ which reduces the complexity of the Hough transform from $\mathcal{O}(1800N)$ to $\mathcal{O}(120N)$. At the same time, the storage requirement which is proportional to Ψ is reduced from $\mathcal{S}(\Psi)$ to $\mathcal{S}(0.5\Psi')$.

The major computation involved in the Hough transform is in the evaluation of sinusoidal function in Eq.(1). For an efficient implementation, we define an $n \times 2$ matrix of constants

$$A = \begin{bmatrix} \cos(\theta_1) & \sin(\theta_1) \\ \cos(\theta_1 + \delta\theta) & \sin(\theta_1 + \delta\theta) \\ \vdots & \vdots \\ \cos(\theta_1 + k\delta\theta) & \sin(\theta_1 + k\delta\theta) \\ \vdots & \vdots \\ \cos[\theta_1 + (n-1)\delta\theta] & \sin[\theta_1 + (n-1)\delta\theta] \end{bmatrix}.$$

The value of n (number of rows in the matrix A) is determined by the following constraints

$$\theta_1 + (n-2)\delta\theta < \theta_2 \leq \theta_1 + (n-1)\delta\theta.$$

Let

$$P^{(i)} = (p_1^{(i)}, \ldots, p_n^{(i)})^t$$

be the Hough transform of the ith centroid point $z_i = (x_i, y_i)^t$, where

$$p_k^{(i)} = \mathcal{Q}[\rho^{(i)}(\theta_1 + k\delta\theta)],$$

$\mathcal{Q}$ is the nonlinear transform for quantizing z_i's Hough transform at $(\rho^{(i)}, \theta_1 + k\delta\theta)$ and

$$\rho^{(i)}(\theta_1 + k\delta\theta) = x_i \cos(\theta_1 + k\delta\theta) + y_i \sin(\theta_1 + k\delta\theta).$$

Then

$$P^{(i)} = \mathcal{Q}(Az_i).$$

An improved Hough transform can now be described as follows:

For $(z_i; i = 1, \ldots, N)$ {

  calculate $P^{(i)} = \mathcal{Q}(Az_i) = (p_1^{(i)}, \ldots, p_n^{(i)})^t$

  For $(k; k = 1, \ldots, n)$ {

    increment $\mathcal{H}(p_k^{(i)}, \theta_1 + k\delta\theta)$

  }

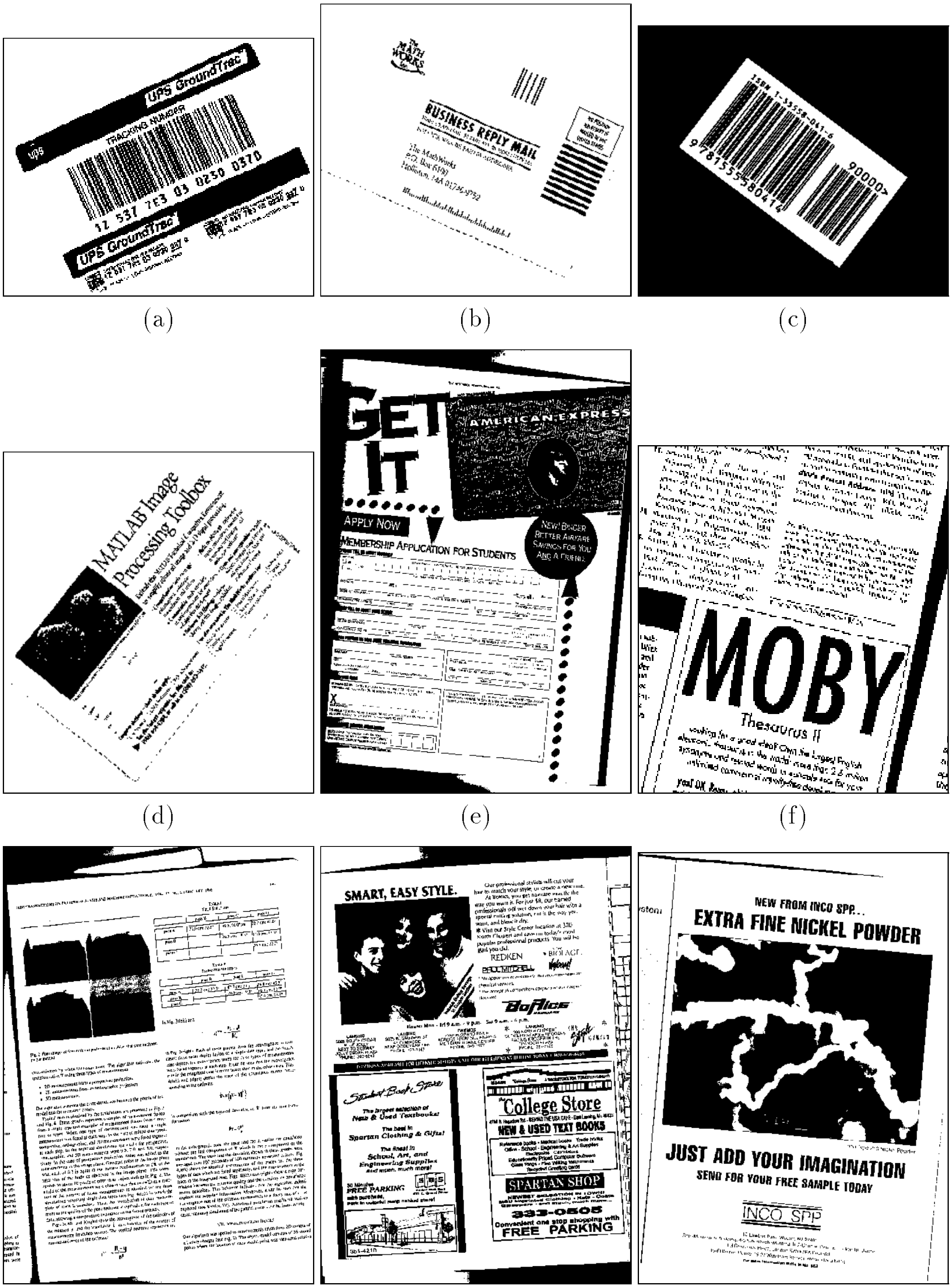
}

Find the maximum in the array $\mathcal{H}(\rho, \theta)$

4 Experimental Results

The skew detection algorithm proposed in this paper has been applied to a large number of document images to test its capability for different applications. These images were obtained by scanning envelopes, journals, magazines, postal labels, yellow pages and coupons using an HP ScanJet IIcx 24-bit color scanner. Some of the scanned images are shown in Fig.5. The image shown in Fig.5(a) is scanned from a postal label, in 5(b) from a post card, in 5(c) from the back cover of a handbook, in 5(d) from a coupon, in 5(e) from a yellow page, in 5(f) from a magazine, in 5(g) from the IEEE Transactions on Pattern Analysis and Machine Intelligence, in 5(h) and 5(i) from an advertisement brochure. Note that thick black margins appear in the images shown in Figs.5(c), (e), (f), (g) and (h). The images in Fig.5 contain a variety of fonts, logos, tables and bar codes. Table 1 gives information about the input image sizes, scanning resolution and estimated skew angle. The corresponding processing times on a SPARC20 CPU by our algorithm and the traditional BHT algorithm are also shown in Table 1. In our experiments, we use $\delta\theta_0 = 0.1°$, $\Delta\theta = 180°$ and, therefore, $\delta\theta_1 = 3°$. The algorithm works well on all of these document images which have different characteristics and resolutions. The correct orientations of the images based on the estimated skew angle are shown in Fig.6. However, our algorithm failed to detect the correct skew for some of these images when the scanning resolution was reduced further (for example, Fig.5(c) at 50 dpi and Fig.5(f) at 20 dpi).

The skew detection accuracies reported in the literature are difficult to compare because



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure 5: Input document images for skew detection: (a) a postal label; (b) a post card; (c) a bar code; (d) a coupon; (e) an application form in yellow pages; (f) photocopied document; (g) a page in the IEEE TPAMI; (h) and (i) advertisements in magazines.

Figure 6: Skew corrected results for the images shown in Fig.5.

Table 1: Performance of skew detection.

Image (Fig.5)	Size (pixels)	Resol. (dpi)	Time (s)		Estimated skew angle (degrees)
			AHHT	BHT	
(a)	268×225	50	0.1	0.9	25.1
(b)	319×300	50	0.1	3.6	-23.8
(c)	326×286	100	0.1	0.5	-36.9
(d)	297×327	50	0.1	4.7	51.2
(e)	415×596	50	0.4	13.5	-5.1
(f)	238×268	50	0.1	14.7	-13.5
(g)	422×598	50	0.5	34.8	4.5
(h)	414×587	50	0.4	15.3	3.1
(i)	412×576	50	0.2	6.3	-4.8

they depend on many factors such as the accuracy of the scanner and scanning resolution. Our algorithm was tested on synthetic document images of 16 pages of this manuscript at 75 dpi (the first page is shown in Fig.1). By synthetic we mean that these images were not scanned and the rotation was introduced manually. Our results indicate that we can obtain an accuracy of 0.1 degree, except for the pages containing Figs.4, 5 and 6. The AHHT method shows its power especially when $\delta\theta_0 < 0.1$ and it is not suitable in the case of $\delta\theta_0 > 1$ (when the value of Ψ is not very large), because of the large cell size in the Hough space.

5 Conclusions

The Hough transform can be used to accurately detect the document skew. However, the presence of noise, black margins, forms, and figures in the input image limits its effectiveness. Furthermore, its computational complexity is a limiting factor, especially when a high detection accuracy is required. The existing methods using the Hough transform are usually able to detect the skew angle of text-only documents in a limited angular range. Compared with the traditional skew detection methods using the Hough transform, our algorithm has the following three significant features: (i) instead of pixels and runs, we use centroids which are quickly extracted based on the BAG data structure; (ii) the Hough transform is implemented hierarchically with different angular resolutions and within different angular ranges; and (iii) a constant matrix containing coefficients of Hough transform is introduced to decrease the computation of sinusoidal function in Eq.(1). These features allow our algorithm to detect the skew angle of relatively more complex document images scanned at a low resolution with a high accuracy and within the whole angular range $(0°, 180°)$. The algorithm has been shown to be able to handle document images scanned at a resolution of 50 dpi and images which contain noise, black margins, figures, and forms.

We currently use a commercial scanning software (PixelFx) and, therefore, the BAG creation is performed only after the entire scanning of the document is completed. More than half of the total processing time needed by our algorithm is used to create the BAG. If the BAG creation module can be inserted inside the scanning and binarization module (see Fig.2), then this time-consuming procedure can be completed during the scanning itself.

In the current algorithm we find a peak at a single point in the Hough space. In fact, one should examine a range of similar angles and all the polar distances in the accumulator array when estimating the skew direction. A two-dimensional window can be introduced for this purpose where all the accumulator cells inside the window with values larger than a threshold value will be summed. The skew angle is the weighted average within that window where the peak is found; the weight values are based on the entries in the accumulator array. Based on this process, the detected peak will be more robust in situations where the document image has several different candidate directions due to a non-rigid skew or text/symbol lines with different orientations.

It is worth mentioning that usually only a small part of a document is sufficient for skew detection. This will greatly reduce the processing time. However, how to quickly determine the significant area of the document for skew detection is an open problem.

Acknowledgment

We would like to thank Mr. Sharath Pankanti for providing the source code for the basic Hough transform and Miss Yu Zhong for her assistance in preparing this manuscript.

References

[1] W. Postl, "Detection of linear oblique structures and skew scan in digitized documents," in *Proc. of the 6th Int. Conf. on Pattern Recognition*, (Paris), pp. 687–689, 1986.

[2] H. S. Baird, "The skew angle of printed documents," in *Proc. of SPIE 40th Annual Conf. and Sym. on Hybrid Imaging Systems*, pp. 21–24, 1987.

[3] T. Akiyama and N. Hagita, "Automated entry system for printed documents," *Pattern Recognition*, vol. 23, pp. 1141–1154, 1990.

[4] T. Pavlidis and J. Zhou, "Page segmentation by white streams," in *Proc. of the 1st Int. Conf. on Document Analysis and Recognition*, (Paris), pp. 945–953, 1991.

[5] Y. Ishitani, "Document skew detection based on local region complexity," in *Proc. of the 2nd Int. Conf. on Pattern Document Analysis and Recognition*, (Tskuba Science City), pp. 49–52, 1993.

[6] A. Hashizume, P. S. Yeh, and A. Rosenfeld, "A method of detecting the orientation of aligned components," *Pattern Recognition Letters*, vol. 4, pp. 125–132, 1986.

[7] L. O'Gorman, "The document spectrum for structural page layout analysis," *IEEE Trans. Pattern Anal. and Machine Intell.*, vol. 15, pp. 1162–1173, 1993.

[8] L. O'Gorman and R. Kasturi, *Document Image Analysis*. IEEE Computer Society Press, 1995.

[9] S. N. Srihari and V. Govindaraju, "Analysis of textual images using the Hough transform," *Machine Vision and Applications*, vol. 2, pp. 141–153, 1989.

[10] S. Hinds, J. Fisher, and D. D'Amato, "A document skew detection method using run-length encoding and the Hough transform," in *Proc. of the 10th Int. Conf. on Pattern Recognition*, (Atlantic City), pp. 464–468, 1990.

[11] Y. Nakano, Y. Shima, H. Fujisawa, J. Higashino, and M. Fujinawaet, "An algorithm for the skew normalization of document image," in *Proc. of the 10th Int. Conf. on Pattern Recognition*, (Atlantic City), pp. 8–13, 1990.

[12] S. Chen and R. M. Haralick, "An automatic algorithm for text skew estimation in document images using recursive morphological transforms," in *Proc. of the First Int. Conf. on Image Processing*, (Austin), pp. 139–143, 1994.

[13] J. Liu, C. M. Lee, and R. B. Shu, "An efficient method for the skew normalization of a document image," in *Proc. of the 11th Int. Conf. on Pattern Recognition*, (The Hague), pp. 122–125, 1992.

[14] B. Yu, X. Lin, Y. Wu, and B. Yuan, "Isothetic polygon representation for contours," *CVGIP: Image Understanding*, vol. 56, pp. 264–268, 1992.

[15] B. Yu, "A method for automatic understanding of symbol connected diagrams," in *Proc. of the 3rd Int. Conf. on Document Analysis and Recognition*, (Montreal), 1995. Accepted for presentation.

[16] J. Illingworth and J. Kittler, "A survey of the Hough transform," *Computer Graphics and Image Processing*, vol. 44, pp. 87–116, 1988.