Edge-based Connected Component Approach for Skew Correction of Complex Document Images

No Institute Given

Abstract. Skew correction of complex document images is a difficult task. We propose an edge-based connected component approach for robust skew correction in documents with complex layout and content. The algorithm essentially consists of two steps - an 'initialization' step to determine the image orientation from the centroids of the connected components and a 'search' step to find the actual skew of the image. During initialization, we choose two different sets of points regularly spaced across the the image, one from the left to right and the other from top to bottom. The image orientation is determined from the slope between the two succesive nearest neighbors of each of the points in the chosen set. The search step finds succesive nearest neighbors that satisfy the parameters obtained in the initialization step. The final skew is determined from the slopes obtained in the 'search' step. Unlike other connected component based methods, the proposed method does not require any binarization step that precedes connected component analysis. The method works well for scanned documents with complex layout of any skew with a precision of 0.5 degrees.

1 Introduction

During digitization of documents, it often happens that the document page is not aligned correctly. The relative inclination angle of the page being acquired must be detected and accounted for as it can cause serious performance deterioration of any text processing system. Thus, skew estimation is one of the first tasks to be solved after document scanning. Although skew detection and correction problems are more than two-decades old, faster, more accurate and robust solution remain a matter of interest today.

Many researchers have used the horizontal projection profile [9, 8, 10] method for skew detection. The deviation of the projection histogram is an evaluation function for the skew angle. The skew angle corresponds to the rotation angle where the mean square deviation of the projection histogram is maximized. Algorithms based on projection profile assumes that documents have text arranged along parallel straight lines and the text represents most of the document image. In the presence of pictures or images in the document, the performance of these algorithms is rarely satisfactory. Projection profile methods are, in general, limited to estimate skew angle within ± 10 degrees. Srihari and Govindaraju [1] used Hough transform method to detect skew in document images. Yu and Jain [2] also proposed a skew correction method employing Hough transform and

connected components. However, Hough transform based methods are computationally expensive. In the case of documents with complex layout, it is also difficult to choose a proper threshold on the accumulator matrix to pick the lines. Hashizume et al. [3] proposed nearest neighbor (NN) clustering for skew detection. For each component, the direction of its nearest neighbor is computed. The peak of histogram of the direction angle indicates the document skew angle. In [4], O'Gorman introduced the 'docstrum' analysis as a generalization of the approach. Pal and Chaudhuri [5] also used a connected component approach, selecting only those components whose height are less than the average height. This leads to fairly accurate identification of the pixels lined up along the text line. The mean line and base line of the text are then obtained using Hough transform on these components. Lu and Tan [6] uses nearest-neighbor chains (NNCs) to develop a skew estimation method with a high accuracy and with language-independent capability. The NNCs are extracted from the adjacent NN pairs, in which the slopes of the NNC with a largest possible number of components are computed to give the skew angle of document image. The most desirable feature of the nearest neighbor approach is that it does not assume any specific layout and hence it can correct skew of any angle irrespective of the document page layout. However, all connected-component based methods involve a binarization step which is not a trivial task for complex documents. Improper binarization could introduce noise and broken characters. These undesired extra components can affect the accuracy of nearest neighbor approaches.

2 Edge-Box based approach for skew detection

The proposed approach uses the edge map to find the connected components thereby bypassing the binarization step. Performing connected component analysis on the edge map yields a better result as there is no binarization noise involved. It also does away with the problem of choosing a proper threshold for binarization, which is a difficult task for complex images. The choice of window size in local binarization methods is a critical parameter that can drastically affect the binarization result. Hence, local binarization methods do not work well if there are characters of varying sizes. In this work, we have used the Canny edge detector [7] for obtaining the edge map. An 8-connected component labeling follows the edge detection step. We call each connected component obtained as above an edge-box (EB). Since the edge captures both the outer as well as the inner boundaries of the character, there could be two or more EBs arising out of a single character. For example, the letter 'e' has 1 inner EB (EB_{int}) and 1 outer EB (EB_{out}). We need to eliminate the undesired EBs as they could interfere in the subsequent estimation of the text baseline. Thus, we filter out unwanted components as follows:

```
if ((N_int >0) AND (N_int < 3))
   Remove EB_int
else
   Remove EB_out
end
```

where EB_{int} denotes the EBs that lie completely inside the current EB under

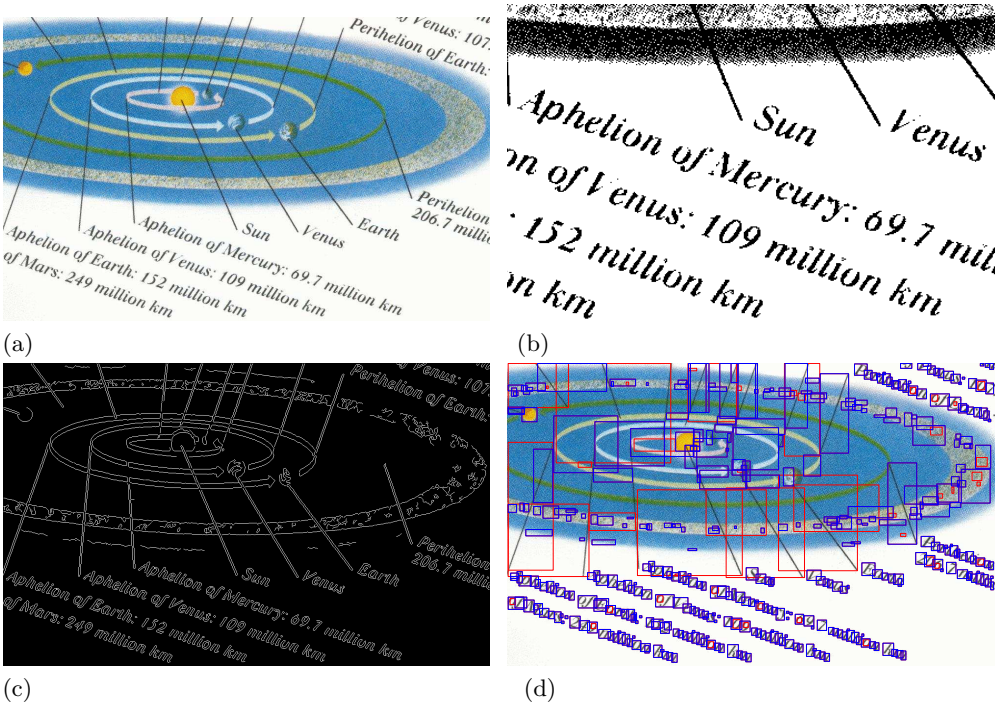

Fig. 1. (a) Example of a complex image (b) A portion of the binarized image obtained using Otsu's method showing broken characters (c) The output of Canny edge detection and (d) The edge boxes obtained after connected component analysis on the edge map. The EBs shown in red are filtered out before EB clustering for obtaining the text orientation. The use of EBs effectively bypassed binarization which is a difficult task in complex images.

consideration and N_{int} is the number of EB_{int}. We treat those EBs that have number of ON pixel less than 8 as noise and are eliminated. The aspect ratio is constrained to lie between 0.2 and 5 to eliminate highly elongated regions. After this filtering step, we are left with only those that could be a text character. We compute the centroid of each of these EBs and use these set of points to determine the skew angle.

2.1 The Initialization Step

Since the characters are regularly spaced and have a consistent orientation, the slopes between each successive EB can be used to estimate the text orientation. Though we do not assume any particular paragraph format, all the text lines in the image are assumed to be parallel to each other. The bounding box information of the EBs is used to determine whether the image is vertically or horizontally aligned. In particular, we sort the centroids of all the bounding boxes based on their y-coordinates. We select a number of initial starting seeds uniformly spaced apart from these ordered set of points. For each seed point, we search for its two successive NNs that satisfy the following constraints:

$$(|\phi_1| < 45^\circ) \,\text{AND}\, (|\phi_2| < 45^\circ) \tag{1}$$

$$[max(\phi_1, \phi_2) - min(\phi_1, \phi_2)] < \mathrm{T}_\phi \tag{2}$$

$$[max(\delta_1, \delta_2) - min(\delta_1, \delta_2)] < 3 \times max(W_{EB}, H_{EB}) \tag{3}$$

where $\phi_1$and ϕ_2 denote the slope between the seed point with the its two successive NNs as illustrated in the Fig. 2. T_ϕ is a threshold on the difference between the value of the two slopes and is set to 8 degrees. W_{EB} and H_{EB} are the width and height of the current seed respectively. The additional constraint on the distance δ_1 and δ_2 between the two neighbors restrict the NNs coming from two different text lines which satisfy the slope constraints. Since characters in

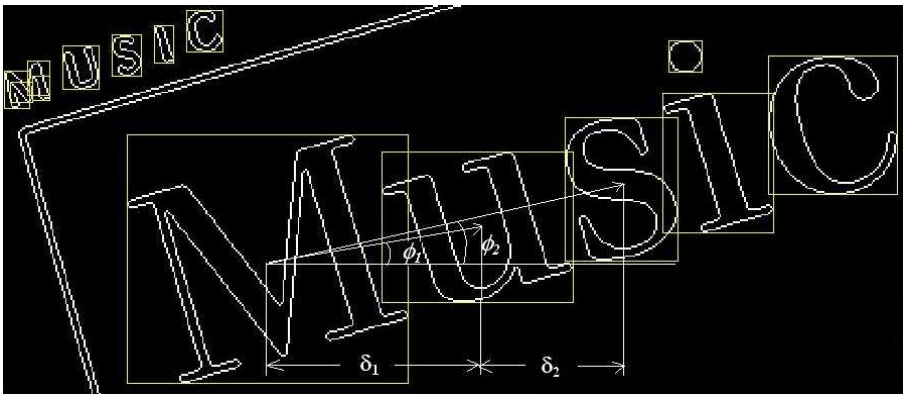

Fig. 2. The slope and distance parameters that guide the selection of initial seed points.

a word/sentence lie close to each other, we limit the search to only a subset of points lying in the neighborhood of each seed point to reduce the computation time. In the event where none of these points satisfy Eqns. 1 to 3, the seed point under consideration is rejected. Eqn. 1 will be generally satisfied by these points if the text were oriented horizontally.

A similar operation is performed on another set of seed points obtained from the centroids ordered with respect to its x-coordinates. This time Eqn. 1 is modified as follows to obtain seeds and their associated slopes in the vertical direction.

$$(|\phi_1| > 45^\circ) \text{ AND } (|\phi_2| > 45^\circ) \tag{4}$$

The number of seed points rejected in the two cases indicate the orientation of the text. The one which gives a lower value is chosen as the correct orientation. In this way, we obtain a number of triplets comprising of the seed point and its two successive NNs along with its associated slope which will guide us in identifying those EBs that lie in a single text line.

2.2 The Search Step

From each triplet of EBs thus obtained, we find all those EBs that lie in the same text line as the seed using the initial slope. Since we have the slope associated with each of the seed, we use only Eqn. 2 in the search step. Every time two succesive EBs satisfy this slope constraint, the mean of ϕ_1 and ϕ_2 is stored. We pick only those values of slopes that lie between one standard deviation away from the mean slope. The median of the selected values of slopes is taken as the final estimate of the skew angle and is used to correct the input image.

3 Experimental Results

The test images used in our work are acquired from a HP scanjet 2400 at a resolution of 300 dpi. We have considered images with complex layouts to illustrate the feasibility of the proposed method. The thresholds 0.1 and 0.25 are used for the hysteresis thresholding step of Canny edge detection. An 8-connected component labeling follows the edge detection step and the EBs are obtained. The search for the seed points is carried out wtihin a square region confined to 5 times the width of the EB under consideration. The slopes of all the clusters of EBs that lie between one standard deviation away from the mean slope are considered and the median value is used to correct the skew. Ideally, there should be at least one seed in each text line. The choice of the seeds could affect the estimate of the skew. The estimate is better if the seeds lie in as many different text lines as possible. Sorting the centroids with respect to their x or y-coordinates and then choosing the seeds at regular intervals allow us to obtain the seeds from different text lines. Fig. 3 shows the results of the algorithm applied on documents with few texts, complex layout and content, multi-column text and a Kannada (an Indian script) document. To evaluate the performance of our algorithm, we consider a test image and rotate it by various angles and the skew is computed for each angle of rotation. As shown in Table 1, the proposed method is applicable to documents of any arbitrary skew with a precision of less than 0.5^o.

Table 1. Accuracy of skew estimate obtained using our algorithm on an image rotated over several angles

Angle of Rotation	-85^o	-65^o	-45^o	-25^o	-5^o	5^o	25^o	45^o	65^o	85^o
Estimated Skew	-84.7	-65.2	-44.4	-25.5	-5.4	4.8	24.7	44.5	64.7	84.6

Fig. 3. The results of the proposed algorithm on some skewed document images with few texts, complex layout and content, multi-column text and Kannada script. The skew-corrected output images are ordered in the second row.

4 Conclusion & Future Work

A robust method for skew correction is presented in this paper. The method does not assume any specific page layout and is shown to work well on complex document images with any arbitrary skew. The advantage of using EBs are two-fold. Firstly, it does away with the problems associated with binarization which is not a trivial task complex images. Secondly, it enables the algorithm applicable to documents having text characters of different sizes. The initialization step of the algorithm automatically determines the seeds and the search direction. The slopes obtained from the search step gives an accurate estimate of the skew angle. The method can be extended to documents with multiple skew by locally correcting cluster of EBs corresponding to each seed.

References

1. S N Srihari and V Govindaraju, "Analysis of textual images using the Hough transform", Machine Vision and Applications, vol. 2(3), 1989.
2. B. Yu and A K Jain, "A robust and fast skew detection algorithm for generic documents", Pattern Recognition, vol. 29 (10), 1996.
3. 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.
4. L O'Gorman, "The document spectrum for page layout analysis", IEEE Trans. PAMI, vol. 15, pp. 1162-1173, 1993.
5. U Pal and B B Chaudhuri, "An improved document skew angle estimation technique", Pattern Recognition Letters, vol. 17, pp, 1996.
6. Y Lu and C L Tan, "A nearest-neighbor chain based approach to skew estimation in document images", Pattern Recognition Letters, vol. 24, pp. 23152323, 2003.
7. J Canny, "A Computational Approach To Edge Detection", IEEE Trans. PAMI, vol. 8, pp. 679-714, 1986.
8. A Bagdanov and J Kanai, "Projection Profile Based Skew Estimation Algorithm for JBIG Compressed Images", ICDAR, pp. 401-405, 1997.
9. G Ciardiello, G Scafuro, M T Degrandi, M R Spada and M P Roccotelli, "An experimental system for office document handling and text recognition", ICPR, vol. 2, pp. 739-743, 1988.
10. Y Ishitani, "Document Skew Detection Based on Local Region Complexity", ICDAR, pp. 49-52, 1993.