Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

e-PCP: A robust skew detection method for scanned document images

Prasenjit Dey, S. Noushath*

Hewlett-Packard Laboratories, #24 Salarpuria Arena, Koramangala, Bangalore 560030, India

ARTICLE INFO

Article history:
Received 20 April 2009
Accepted 24 June 2009

Keywords:
Skew detection
Piece-wise coverings (PCP)
Confidence measure
Text-flow orientation detection
Enhanced-PCP
Skew correction

ABSTRACT

We present here an enhanced algorithm (e-PCP) for skew detection in scanned documents, based on the work on Piecewise Covering by Parallelogram (PCP) for robust determination of skew angles [C.-H. Chou, S.-Y. Chu, F. Chang, Estimation of skew angles for scanned documents based on piecewise covering by parallelograms, Pattern Recognition 40 (2007) 443–455]. Our algorithm achieves even better robustness for detection of skew angle than the original PCP algorithm. We have shown accurate determination of skew angles in document images where the original PCP algorithm fails. Further, the increased robustness of performance is achieved with reduced number of computation compared to the originally proposed PCP algorithm. The e-PCP algorithm also outputs a confidence measure which is important in automated systems to filter cases where the estimated skew angle may not be very accurate and thus can be handled by manual intervention. The proposed algorithm was tested extensively on all categories of real time documents and comparisons with PCP method is also provided. Useful details regarding faster execution of the proposed algorithm is provided in Appendix.

1. Introduction

Skew estimation of document refers to the process of finding the angle of inclination made by the document with respect to horizontal axis, which is often introduced during document scanning. For any ensuing document image processing tasks (such as page layout analysis, OCR, document retrieval etc.) to yield accurate results, the skew angle must be detected and corrected beforehand. Though a large number of skew estimation methods have been proposed, development of a solitary skew estimation algorithm which is relatively fast and yet handles wide range of documents is still an elusive goal. It is because of this very reason, document skew estimation research is still active although it has been studied for several decades now.

The algorithms for skew estimation can mainly be classified as the ones based on (i) projection profile (PP) [6,3], nearest neighbor (NN) [1,10], (iii) Hough transform (HT) [9,4,7,12] and (iv) cross-correlation (CC) [8,5].

Among these methods, PP based methods are most commonly used. These methods compute projection profiles of the document at various angles and compute the skew angle of the document based on some maximization criteria. However, these methods are computationally intensive and they need to carry out expensive rotation operation at every angle. Moreover, it is sensitive to the layout of the image [6]. The NN techniques calculate skew angle between each component and its nearest neighbor and the histogram of the angles are formed thereby. The peak in the histogram corresponds to the skew angle of the document. Another class of skew estimation methods is based on the HT. The idea is that collinear pixels in Cartesian space constitutes cluster of (ρ, θ) bins in Hough space. The peak in the Hough space corresponds to skew angle of the document. However, there are two main demerits of both NN and HT based methods:

(1) one has to extract text regions from the document which is again a non-trivial task for documents whose layouts are complex and
(2) they are computationally intensive.

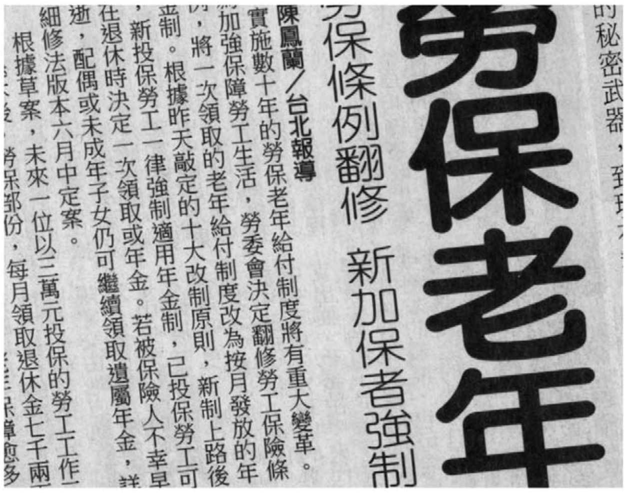
On the other hand, there are also methods that compute skew angle of the document based on maximum variance of transition counts (TC) [11] and based on cross-correlations (CC).

Recently a robust skew estimation algorithm based on piecewise coverings of objects by parallelograms (PCP) was proposed [2]. In this approach, the document image is divided into several non-overlapping slabs and the objects within each slab is covered by parallelograms at various angles. The angle at which objects are *best covered* corresponds to skew angle of the document. The PCP algorithm has been demonstrated to achieve faster and robust results than PP, HT and NN based methods in [2].

However, there exists one main drawback with this approach. When vertically flowing text (VFT) in a document (which is common in Chinese and Japanese documents) touches the borders of

* Corresponding author. Tel.: +91 80 25042256.
E-mail addresses: pdey@hp.com (P. Dey), nawali_naushad@yahoo.co.in, noushath.s@hp.com (S. Noushath).

0031-3203/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2009.06.019



的秘密武器，致

勞保老年

勞保條例翻修　新加保者強制

陳鳳蘭／台北報導

實施數十年的勞保老年給付制度將有重大變革。加強保障勞工生活，勞委會決定翻修勞工保險條例，將一次領取的老年給付制度改為按月發放的年金制。根據昨天敲定的十大改制原則，新制上路後，新投保勞工一律強制適用年金制，已投保勞工可在退休時決定一次領取或年金。若被保險人不幸早逝，配偶或未成年子女仍可繼續領取遺屬年金，詳細修法版本六月中定案。

根據草案，未來一位以三萬元投保的勞工工作

勞保部份，每月領取退休金七千兩

Fig. 1. A document with vertical flowing text touching the borders.

the document as shown in Fig. 1, this method fails to yield desired skew angle. This is because the maximization criterion (this will be discussed in Section 2) of PCP approach [2] will not *strongly* favor a particular angle to arrive at the actual estimate of the skew. Consequently, the method may lead to a wrong estimate of the skew angle for such kinds of documents.

Moreover, one can encounter documents of type shown in Fig. 1 very frequently, especially while scanning large documents (such as newspapers, posters etc.) whose content often goes beyond the scanner bed. This necessitates the need of a mechanism in the original PCP approach, which automatically determines the flow of text and then determines the slab orientation accordingly (i.e. either horizontal or vertical). Hence, our proposed enhanced-PCP (e-PCP) algorithm enhances the conventional PCP in this aspect. In contrast to the PCP method [2], the overall enhancements achieved in the proposed method are as follows:

(1) Improved robustness across any kind of documents, especially for VFT documents.
(2) Reduction in number of computations and yet retaining the accuracy of the algorithm.
(3) Insensitive to size of the slab widths by automatically determining the slab orientation.
(4) A robust confidence measure module for reliable skew estimation, which is useful in automated document processes.

Rest of the paper is organized as follows: Review of the PCP algorithm and its shortcomings are given in Section 2. The proposed e-PCP algorithm and its computational load are described in Section 3 and Section 4 respectively. Experimental results are presented in Section 5, and we finally draw some conclusions in Section 6.

2. Review of the PCP algorithm

In this section we briefly review the PCP algorithm [2]. This algorithm is based on the concept that document contains many rectangular objects (text lines, text regions, forms, rectangular pictures etc.) and when there is no skew in the document, these objects can be *best covered* by rectangles. On the other hand, when there is a skew in the document, these objects can only be *best covered* by parallelograms.

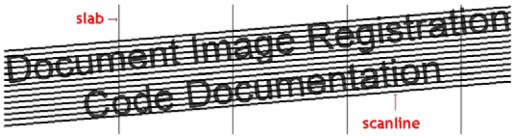


Fig. 2. Parallel scan lines drawn at an angle of text line skew.

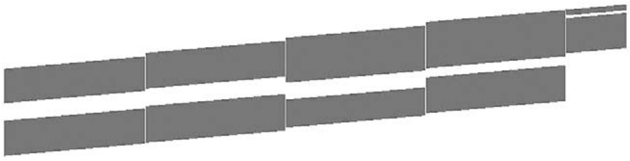


Fig. 3. Parallelograms constructed for the text present in Fig. 2.

In the process, the document is first divided into a number of non-overlapping vertical slabs, and scan-lines are drawn at all angles within the skew angle range (e.g. -15° to $+15^\circ$). Each scan line is further divided into as many sections as the number of slabs, where a section refers to a part of the scan line within a slab. Fig. 2 demonstrate the process of dividing images into slabs and drawing scan lines.[1] In this example, since width of the image is not a multiple of slab width, the last slab is small compared to others. Each section of the scan line is examined for occurrence of any black pixel. If a section contains at least one black pixel, all pixels along that section will be changed to gray, else it will be counted as a white section. Fig. 3 shows parallelograms constructed for the objects shown in Fig. 2 by changing those sections to gray which contain at least one black pixel.

This process of scan-line drawing is repeated for all angles within the skew angle range and number of white sections at each angle is computed. The intuitive idea is that when scan lines are drawn at angle corresponding to skew angle of the document, there will be more number of white sections than the gray ones, which is quite evident in Fig. 4. This is true even in case of complex cases such as when document has large scale figures, forms or tables, multilingual documents, etc. [2]. Thus the process of estimating skew angle reduces to maximizing the following criteria:

$$\theta^* = \underset{\theta}{\operatorname{argmax}}\, WS(\theta) \tag{1}$$

where $WS(\theta)$ is the number of white sections when scan lines are drawn at angle θ.

Unlike HT, PP or NN based algorithms, this method produces robust results for many real time documents [2]. Nevertheless, it suffers from following two major drawbacks which deserves further study:

(1) *Subjectiveness of slab width*: As mentioned earlier, when text lines are aligned vertically and if their content touches borders of the document, the success of the algorithm depends on the appropriate size of the slab width. For example, the document shown in Fig. 5 has a skew angle of 0°. The estimated skew angle for the document for different slab widths is shown in Table 1. We see that the estimated skew angle is highly dependent on the slab width, which is fixed to a particualr value a priori in PCP method [2]. Hence, for such kind of documents, determining appropriate slab width is highly subjective in nature.
(2) *Number of computations*: The whole process of computing the number of white section has to be repeated for each skew angle in the given skew range. In future, if the skew range has to

[1] For illustration purpose, scan lines are drawn at an angle corresponding to the skew angle of the text, and with some gap between them.