A Novel Fast Hierarchical Projection Algorithm for Skew Detection in Multimedia Big Data Environment

Li Cheng, School of Power and Mechanical Engineering, Wuhan University, Wuhan, China & School of Computer Science, South-Central University for Nationalities, Wuhan, China
Gongping Wu, School of Power and Mechanical Engineering, Wuhan University, Wuhan, China

ABSTRACT

Optical character recognition is an effective way for information input of paper media and skew detection of document images is a key stage of it. An algorithm for skew detection employing hierarchical projection is proposed in this paper. Projection histograms at various directions in a given range are acquired according to an initial angle step length. Then variances of it and absolute difference of the variances are calculated respectively and the angle corresponding to the maximum difference is served as rough skew estimate. The similar work above is implemented in which the projection angle range is two times the initial step length and symmetric about the estimate. Finally, the maximum value of the variances is found and the angle corresponding to it is served as skew angle. Experimental results show the algorithm has such advantages as fast processing speed, high detection accuracy, insensitivity to noise and suitable for complex layout.

KEYWORDS

Hierarchical Projection, Hough Transform Multimedia Big Data, Optical Character Recognition, Skew Correction, Skew Detection

INTRODUCTION

Nowadays, the amount of multimedia data is increasing dramatically, which makes people’s information source extends to multiple modalities of media objects, such as image, video, audio, flash and so on. As a rich, intuitive performance, suitable vision multimedia information, the application of image has been included as national defense, industrial manufacturing, media, health care etc. Optical Character Recognition (OCR) is an important means of information input for paper media image, especially in this big data era of information explosion; it has been applied in many fields, such as digitization of large number of ancient books, license plate recognition in the intelligent traffic management system and video caption recognition and so on. To recognize texts in the images by computer, skew corrections of the images are usually needed to make the text lines horizontal or vertical, and then individual characters are segmented, so it is required to first detect skew angle of the document images. Over the years, offline character recognition has been achieving great success due to demands in many fields. The major advantage of these off-line recognizers is to allow the previously written and printed texts to be processed and recognized.

Skew in an acquired image generally occurs due to the inaccuracies in the process of image acquisition. Skew detection and correction is one of the most important pre-processing steps involved

DOI: 10.4018/IJMCMC.2017070104

Copyright © 2017, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.
in the process of OCR. The paper will be dealing with the skew detection and correction of document images i.e. images containing textual information.

Many research findings have been obtained about skew detection of the document images. So far, there are five algorithms for skew detection (Felhi et al., 2011; Li et al., 2007): (i) Nearest Neighbor clustering algorithm (NN); (ii) The traditional Projection Profile (PP) based approach; (iii) Hough Transform algorithm (HT); (iv) Mathematics Morphology algorithm(MM); (v) Cross Correlation algorithm (CC).

In (Hashizume et al., 1986), NN was proposed, and many improved algorithms were developed (Lu et al., 2003; Michael et al., 2010). The algorithm is fast and not limited to images with large text areas or certain range of skew angles, but it generally shows low detection accuracy. The nearest neighbor-clustering algorithm (SUN et al., 2006) was applied to detect skew of Chinese document images, and the least square method was employed to estimate the angle, but the processable text complexity and detectable skew range are unclear.

PP for skew detection was first proposed in (Postl et al., 1986), and some improvements were made to this algorithm (Bloomberg et al., 1995), in which the processing pixels are reduced by down-sampling to save the computing time. However, both the original and improved algorithms are so sensitive to layout of the document images that they often fail to generate correct results when there are multiple fonts or many non-text areas (e.g. pictures, forms or graphs) in the images.

The HT based algorithm was first applied in (Srihari et al., 1989) and many improvements were made to it (Hinds et al., 1990; Le et al., 1994; Pal et al., 1996). The algorithm is robust and simple, but both the time and space complexity are very high. Yu and Jain proposed a multi-step strategy, which makes the algorithm faster and more accurate (Yu et al., 1996). The algorithm was dissected (Singh et al., 2008) from both the time and space complexity, and proposed many methods for improvement, so that performance of the algorithm was greatly improved. But it also has some problems. The improved hierarchical algorithm based on block adjacency graph (BAG) can significantly increase the speed, but it fails to give satisfactory results in the case where straight lines exist in the text lines as the major part. In (GAO et al., 2013), some improvements were also made to the algorithm so that the speed and accuracy were improved, but the processing effect of images with unclear boundary direction was not desirable.

Before skew detection using the mathematics morphology algorithm, opening and closing operations should be implemented for the document images. The CC algorithm can detect skew angle of the images accurately, but the operation takes too much time (Kumar et al., 2012). The combination of the mathematical morphology and Hough transform was applied (ZHANG et al., 2008), but experimental verification was not enough, so the effect is unclear.

According to analysis of the existing algorithms for skew detection of the documents images in the above literatures, all the algorithms have their own advantages, but also have some disadvantages, mainly including: (1) high computing cost and memory consumption; (2) limited detectable skew angle range; (3)sensitive to text layout and failure to adapt to document images with various characters and fonts and complex layout; (4)requiring the control parameters to be set manually, so the adaptability is limited. In response to these disadvantages, an algorithm named hierarchical projection histogram variance (HPV) is proposed. The algorithm can overcome the various disadvantages above, and better solve these problems. It is able to detect skew of the document images fast and accurately, so as to accomplish skew correction and prepare for the subsequent text segmentation.

The rest of this paper is organized as follows: A brief introduction with related works about skew detection for document images in big data environments is presented in Section 1. Principle of projection algorithm for skew detection is proposed in Section 2. Acceleration method in the
The Impact of Web 2.0 on E-Commerce Adoption and Use by Tourism Businesses – Can SMMEs Play the Trick?: A Case of the Eastern Cape Province
www.igi-global.com/chapter/the-impact-of-web-20-on-e-commerce-adoption-and-use-by-tourism-businesses--can-smmes-play-the-trick/183328?camid=4v1a

Optimal Spectral Resolution in Speaker Authentication Application in Noisy Environment and Telephony
www.igi-global.com/article/optimal-spectral-resolution-speaker-authentication/4068?camid=4v1a