Skew Detection of Scanned Document Images

Sepideh Barekat Rezaei, Abdolhossein Sarrafzadeh, and Jamshid Shanbehzadeh

***Abstract*—Skewing of the scanned image is an inevitable process and its detection is an important issue for document recognition systems. The skew of the scanned document image specifies the deviation of the text lines from the horizontal or vertical axis. This paper surveys methods to detect this skew in two steps, dimension reduction and skew estimation. These methods include projection profile analysis, Hough Transform, nearest neighbor clustering, cross-correlation, piece-wise painting algorithm, piece-wise covering by parallelogram, transition counts, morphology.**

***Index Terms*—Document recognition systems, preprocessing, skew, dimension reduction, skew estimation**

I. INTRODUCTION

Skew detection of scanned document images is one of the most important stages of its recognition preprocessing. The skew of the scanned document image specifies the deviation of its text lines from the horizontal or vertical axis. The skew of the document image can be a global (all document's blocks have the same orientation), multiple (document's blocks have a different orientation) or non-uniform (multiple orientation in a text line) [1]. Generally, dimension reduction and skew estimation are two steps of the skew detection of scanned document images. We describe these two steps later.

The first step of skew detection is dimension reduction. Each image is a point in the image space. Each dimension of the image space is related to one of its pixels. The first set of features that can be considered for an image is its pixel values. In other words, the value of each image pixel is one of its features. The dimension of images is high and the employment of all image pixels as its features creates complexity and high computational cost on the irrelevant features. Dimension reduction is the process of reducing the size of features or image pixels and finding another feature with much lower dimensions. We extract or select relevant features of an image with low dimensions from image pixels. To solve a particular problem, we choose several features to achieve the final goal. Features of scanned documents can be divided into three groups:

1. Irrelevant features: those features which are not necessary at all to achieve the desired goal.
2. Weakly relevant features: features that are not always necessary to achieve the desired goal, but may become necessary in certain conditions. These features can be divided into two categories: redundant and non-redundant features.
3. Strongly relevant features: those features which are always necessary to achieve the desired goal.

Dimension reduction methods can be divided into:

1. Feature transformation: in this method, initial set of features transforms to the other set, in respect to retaining the information as much as possible. These methods can be placed in two categories:
   a. Feature extraction: in this method, a new set of features is created by the initial feature set .
   b. Feature generation: in this method, the missing information detected and added to the feature set.
2. Feature selection: this method select an optimal subset of features based on an objective function. The optimal subset includes all of the strongly relevant and weakly relevant but non-redundant features.

Different dimension reduction methods can be used in skew detection of scanned document images. One or more methods can be used consecutively. At the end of this phase, a criterion function for skew detection is obtained.

The second step of skew detection is skew estimation. In this step, using the function defined in the previous step, the skew is estimated. The angle corresponding to the maximum or the minimum value of the function is usually considered as the skew. So, in this step, the maximum or the minimum of the criterion function is achieved.

Until now, many methods for skew detection of scanned document images have been proposed. These methods include projection profile analysis, Hough transform, nearest neighbor clustering, cross-correlation, piece-wise painting algorithm, piece-wise covering by parallelogram, transition counts, morphology. In the following section, we describe these methods and explain the features are used, reason for that feature's suitability for skew detection, the steps of the methods and the history of innovation and change.

II. SKEW DETECTION OF SCANNED DOCUMENT IMAGES

A. Projection Profile Analysis

In this method, the horizontal or vertical projection profile is used as a suitable feature for skew detection. Horizontal (or vertical) projection profile is the histogram of a one-dimensional array with a number of entries equal to the number of rows (or columns). The number of black pixels in a row (or column) is stored in the corresponding entry.

When the skew of the document image is zero degrees, the projection profile peak times will be longer. To understand the reason, consider the scan lines are drawn on

Manuscript received Jan 10, 2013; revised Jan 29, 2013.

Sepideh Barekat Rezaei is an M.Sc. student with the Department of Computer Engineering, Faculty of Engineering, Kharazmi University, Tehran, I.R. Iran (e-mail: s_barekat_r@yahoo.com).

Abdolhossein Sarrafzadeh is an Associate Professor and Head of Department of Computing, Unitec Institute of Technology, New Zealand (email: hsarrafzadeh@unitec.ac.nz).

Jamshid Shanbehzadeh is an Associate Professor with the Department of Computer Engineering, Faculty of Engineering, Kharazmi University (Tarbiat Moallem University of Tehran), Tehran, I.R. Iran (phone: +98 26 34550002; fax: +98 26 34569555; e-mail: jamshid@tmu.ac.ir).