Information Extraction from a Skewed Form Document in the Presence of Crossing Characters

Jin-Yong Yoo, Min-Ki Kim, Sang Yong Han and Young-Bin Kwon
Computer Vision Lab., Dept. of Computer Science and Engineering
Chungang Univ., 221 Heuksuk-Dong, Dongjak-ku
Seoul 156-756, Korea
e-mail: {yong,mkkim,ybkwon}@ripe.chungang.ac.kr, han@peace.cse.cau.ac.kr

Abstract. In this paper, we propose an information extraction method which can restore the handwritten character information from prescribed and skewed form documents. The proposed method includes the following aspects: boundary and successive internal line detection, accurate skew angle measurement, line removal and broken character restoration using morphological analysis model of crossing shape. Using the proposed method, more than 95% of the horizontal and vertical crossing lines are correctly restored.

1 Introduction

Form document analysis has an important role in the area of office automation. Form documents of prescribed shapes are used daily at public institutes, government and public offices. In a form document, characters or numerals are usually surrounded by the prescribed lines which are composed of vertical and horizontal lines. Therefore, it is necessary to cope with the problem of correctly extracting the filled-in data. In spite of existence of some forms which are printed with dropout ink, majority of forms are printed without dropout ink for the sake of cost. When characters or numerals are touching, crossing or overlapping with prescribed lines, the line removal and character restoration from the tables and the form documents are frequently required. This procedure contains two different steps: (i) separating characters from the form frames and (ii) restoration of the characters broken during separation. Some approaches are reported to solve this problem. Okada[1] proposes a procedure to extract the handwritten components. A robust method to extract the user entered components from forms that have gray-level picture backgrounds in addition to texture is proposed. User entered information on a filled-in form is extracted by regis-
tration with a blank master form. A morphological inter-image subtraction in gray-level space (GLS) is used for extraction of those components. Handwritten components are robust and of good quality for the recognition step, but it takes a quite long time to process. A form dropout system [2] which can automatically capture form structure from a blank form and extract the filled-in data from filled-in forms is also proposed. Forms are characterized by the presence of horizontal and vertical frames that delineate the usable space. The proposed system focuses on the horizontal frames. A binary image data structure, called block adjacency graph (BAG), is used to locate the horizontal frames even in cases where the form model is unknown and there is modest amount of skew introduced during scanning. This system contains a data deformation problem when they are touched with preprinted text.

Small amount of skew angle is often found on many form documents. It may be caused by the user or the auto-feeder of the scanner. In this case, line extraction may be difficult due to the skew angle. Moreover, when handwritten characters or numerals within a form document touch horizontal and/or vertical lines, the extraction of information becomes more difficult. In order to find out an accurate extraction of the handwritten part from a form, a skew detection and correction technique is used. After skew correction, a distortion on the tables or handwritten characters can occur, and it may become difficult to extract the information correctly.

In this paper, a method which extracts the characters crossing the lines in skewed form documents is proposed. The proposed technique consists of two important aspects: boundary line tracing and skew angle estimation based on the four detected boundary lines, and successive line removal and broken character restoration. After the introduction, overall procedures of the proposed method are reviewed. In section 3, the boundary detection and the skew angle estimation technique are described. In section 4, a procedure for line removal and the restoration of overwritten characters is explained. Following the experimental results, conclusions are presented.

2 Configuration of Information Extraction Procedures

Figure 1 shows the overall configuration of our information extraction procedures. An input image is a form document which contains a table. Input images often contain a skew angle caused by the user or the auto-feeder of the scanner. The handwritten information can be written at an angle due to writer's habit. Thus, a skew angle must first be detected in order to extract the prescribed line correctly.

To estimate a skew angle, boundary line tracing is done after binarization. Boundary line tracing is a procedure to detect the external lines of a given table. Lines of a form documents are not straight but irregular with various thickness. Thus, it is better to trace a medial axis of boundary line from starting point. A line tracing is accomplished by saving the consecutive medial axis points. Using four detected boundary line information, a skew angle is calculated by