MVA '96 IAPR Workshop on Machine Vision Applications, November. 12-14, 1996, Tokyo, Japan

An Efficient Recognition and Data Extraction Method for Table-Form Documents

Lin Yu Tseng and Rung-Ching Chen[+]
Department of Applied Mathematics
National Chung Hsing University
Taichung, Taiwan, R.O.C.

Abstract

In Asia, many documents processed in offices are table-form documents. Hence the automatic processing of table-form documents is an important issue of the office automation research. In this paper, we propose an efficient representation method for table-form documents. The representation method is based on three types of line segments. The line segments are normalized and sorted, hence the representation method provides an unified and efficient way for learning and recognition of table-forms. Based on this representation, a table-form documents may be learned interactively with the user. After a table-form document was learned, it can then be recognized and the data in data fields be extracted.

1 Introduction

Table-form documents are very popular in many Asian countries. One example is given in Figure 1. Some researches had been focus on the automatic processing of this kind of documents, for example, [1]-[6].In [1], in order to represent the documents that are logically the same but are different slightly in geometric shape as in the same class, a complex representation had been used. But in general, two table-form documents that are different in geometric shape are in fact different. Hence, we propose a much simpler representation for this kind of documents. This representation makes learning and recognition of the documents more efficient. For brevity, we use documents to mean table-form documents thereafter.

The remaining parts of the paper are organized as follows. In section 2, we describe the representation method for the table-form document. In section 3, the learning process is described. In section 4, the recognition of table-forms and data extraction are described. Finally, concluding remarks are given in section 5.

2 The Representation of a Table-Form Document

Previously, we proposed a stroke extraction method to extract the strokes in a Chinese character[7]. This method extract four kinds of primitive strokes, namely, horizontal strokes, vertical strokes, left-slanting strokes and right-slanting strokes. Use this method, we can extract all horizontal line segments, vertical line segments and slanting line segments in a document. Each line segment is represented by two end points $P_1(x_1, y_1)$ and

[+] Rung-Ching Chen is with the Department of Electronics Engineering, Cheng Kuo Technology Collage and is now working the Ph.D. degree at the Department of Applied Mathematics, National Chung Hsing University.

$P_2(x_2, y_2)$. For the horizontal line segment, P_1 represents the left end point, for the vertical line segment, P_1 represents the upper end point, for the slanting line segments, P_1 represents the left end point. When the line segments were extracted, we check if the table-form is skew and fix the problem if it is skew. After all line segments were extracted, the scale of the document is normalized to L×H. We choose L and H to be 400 and 500 in our experiments. Then all coordinates of the end points of line segments are normalized accordingly.

All horizontal line segments are sorted by their P_1's from top to bottom, and for those with the same y_1's, form left to right then. All vertical line segments are sorted by their P_1's from left to right, and for those with the same x_1's, from top to bottom then. All slanting line segments are also sorted by their P_1's in the same way as horizontal line segments are. A document is then represented by the number of horizontal line segments, the number of vertical line segments, the number of slanting line segments and the three sorted sequence mentioned above. The representation for the table-form shown in Figure 1 is also given in Figure 1.

3 Learning

The system must first learn a document, then it can recognize this document afterward and extract the data from the data fields of this document. Based on the representation method stated in section 2, the learning process is described as follows. A blank table-form document is scanned and deskewed, if it is needed, three types of line segments are extracted and a representation for this document is obtained. In a document, there are three kinds of fields, namely, the name field, the data field, and the mixed field. A mixed field is a combination of name fields and data fields. Using horizontal line segments, vertical line segments and slanting line segments, the boundary points of all fields can be determined. With a scanning of the interior of a field, this field can be determined to be a data field or a name/mixed field. We provide an interactive interface for users to specify those data fields in mixed fields. User are also asked to provide some attributes for each data field through this interactive interface. This information will help the OCR recognize more easily the text block extracted form the data field.

4 Recognition and Data Extraction

After a blank table-form was learned, its representation is stored in the form library. A filled-in table-form can now be recognized. It is first scanned and deskewed, if needed. Then the representation consists of the numbers of three types of line segments and the three sorted lists of line segments are derived. A fuzzy matching between this representation and those representations in the form library is made. There must be a table-form in the form library which matches this filled-in table-form best. If the match grade of this table-form is greater than the predefined threshold, this table-form is taken as the form to be recognized. Using the information about the locations of data fields, text blocks can be extracted from the data fields in the filled-in table-form document. These text blocks are then passed to the OCR.

5 Concluding Remarks

In general, table-form documents that have different geometric shapes represent different documents. In this paper, we propose an efficient representation method for the table-form document. Based on the representation method efficient learning method and recognition method can be derived. Because of the normalization, we solve the problem of magnification and contraction. Also, fuzzy matching makes the recognition more robust.

References:

[1] T. Watanabe, Q. Luo, N. Sugie, "Layout recognition of multi-kinds of table-form documents," IEEE transitions on pattern analysis and machine intelligent 1995 Vol. 17, No. 4, pp. 432-445.

[2] R. Casey, D. Ferguson, K. Mobiuddin, E.Walach, "Intelligent forms processing system," Machine Vision and Applications 5, pp. 143-155, 1992.

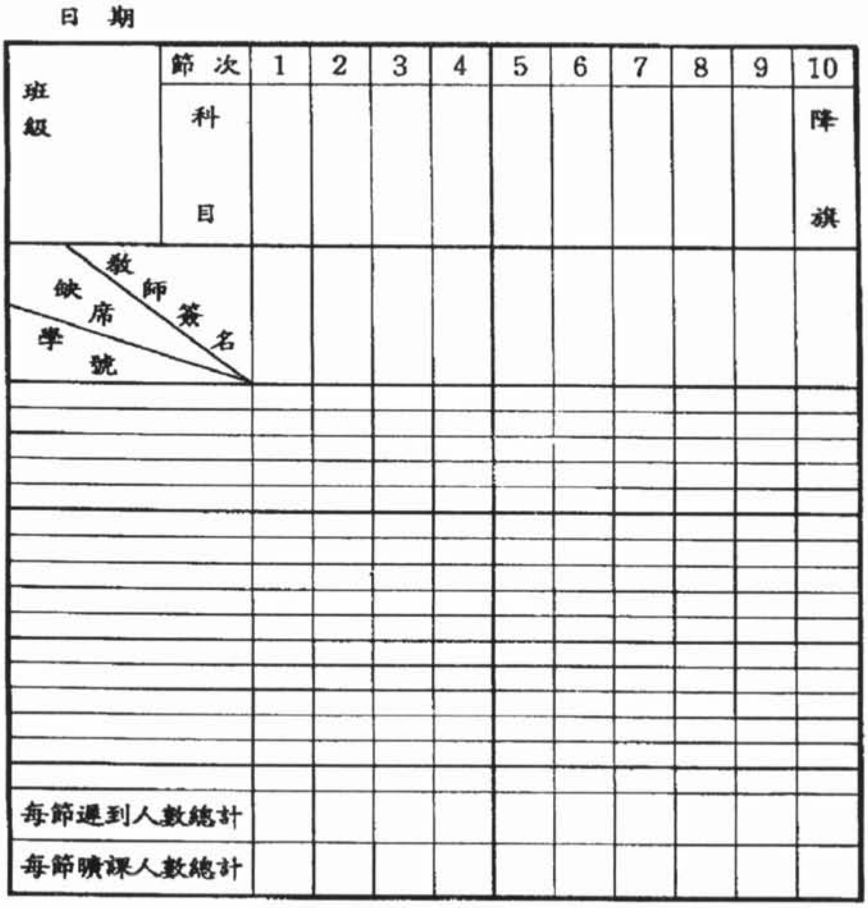
[3] K-C. Fan, J-M. Lu, J-Y. Wang, "A feature point clustering approach to the segmentation of form documents," Proc. Of the 3rd ICDAR 1995, Montreal, Canada, pp.623-626.

[4] J-Y. Lin, Z. Chen, "Identification of business forms using relationships between adjacent frames", Image and Recognition of the Chinese image processing and pattern recognition society volume 3, no.1, 1995. pp. 22-40.

[5] T. Watanabe, H. Naruse, Q. Luo, and N. Sugie, "Structure of table-form documents on the basis of the recognition of vertical and horizontal line segments," Proc. Of the 1st ICDAR 1991, Saint-Malo , France, pp. 638-646

[6] T. Watanabe, Q. Luo , N. Sugie, (1993) "Structure Recognition methods for various types of documents," Machine Vision and Applications 6, pp. 163-176

[7] L. Y. Tseng and C. T. Chuang, "An Efficient Knowledge-based stroke Extraction method for Multi-font Chinese character," Pattern Recognition, Vol. 25, No.12, pp. 1455-1458.



日　期

班級	節次	1	2	3	4	5	6	7	8	9	10
	科目										降旗
教師簽名 缺席學號											
每節遲到人數總計											
每節曠課人數總計											

備註：1.登記符號（曠課「○」，遲到「⊖」，註銷「⊗」）。
　　　2.如有註銷或更正請老師務必簽名。
　　　3.每節遲到、曠課人數總計請老師塡寫。
　　　4.如有上課違規學生，請到敎官室塡處罰單。

```
number of horizontal line segments: 22

number of vertical line segments: 13

number of slanting line segments: 2

h( 0):(  0,  0)-(400,  0) h( 1):( 72, 24)-(400, 24) h( 2):(  0,117)-(400,117)
h( 3):(  0,198)-(400,198) h( 4):(  0,213)-(400,213) h( 5):(  0,228)-(400,228)
h( 6):(  0,243)-(400,243) h( 7):(  0,257)-(400,257) h( 8):(  0,272)-(400,272)
h( 9):(  0,287)-(400,287) h(10):(  0,303)-(400,303) h(11):(  0,317)-(400,317)
h(12):(  0,333)-(400,333) h(13):(  0,348)-(400,348) h(14):(  0,362)-(400,362)
h(15):(  0,377)-(400,377) h(16):(  0,392)-(400,392) h(17):(  0,406)-(400,406)
h(18):(  0,421)-(400,421) h(19):(  0,436)-(400,436) h(20):(  0,466)-(400,466)
h(21):(  0,500)-(400,500)
v( 0):(  0,  0)-(  0,500) v( 1):( 72,  0)-( 72,117) v( 2):(114,  0)-(114,500)
v( 3):(142,  0)-(142,500) v( 4):(170,  0)-(170,500) v( 5):(198,  0)-(198,500)
v( 6):(226,  0)-(226,500) v( 7):(254,  0)-(254,500) v( 8):(282,  0)-(282,500)
v( 9):(310,  0)-(310,500) v(10):(338,  0)-(338,500) v(11):(366,  0)-(366,500)
v(12):(400,  0)-(400,500)
s( 0):( 29,117)-(243,198) s( 1):(  0,153)-(243,198)
```

Figure 1. A table-form and its representation.