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

Identification of Document Types from Various Kinds of Document Images Based on Physical and Layout Features

Hiroyuki MASAI *
Department of Information Engineering
Graduate School of Engineering
Nagoya University

Toyohide WATANABE †
Department of Information Engineering
Graduate School of Engineering
Nagoya University

Abstract

When we develop a general purpose document image understanding system, it is important, as the first step, to distinguish individual documents. We propose an approach which first classifies document images into some distinct types and then interprets them exactly by using an appropriate document model. In this paper, we define groups of documents and describe the classification method based on the verification mechanism by using physical and layout features of documents. Also, we show the experimental result in our method.

1 Introduction

The subject about automatic extraction of useful/meaningful information from document images is called document image understanding, and is one of interesting topics currently. Various methods and approaches for different kinds of documents have been proposed until today[1][2][3]. Most of them analyze document images interpretatively, using document models which are knowledge resources about the composition rule, description rule, layout structure, item position, item format, data format and so on. These conventionally developed systems were dependent on the particular applications because the document models are organized from application-specific domains[4]. On the other hand, there were some attempts to develop a general purpose document image understanding system[5]. However, most of them focused only on very limited documents, but did not propose methods/approaches to deal with different types of documents.

In this paper, we propose an experimental approach to classify various documents automatically with a view to applying appropriate document models to individually classified documents. This is the first step when we implement a general purpose document image understanding system from a practical point of view. The characteristic in our approach is to recognize document types on the basis of classification and verification mechanism. Such a paradigm makes not only the processing simple, but also the classification successful.

*Address: Furo-cho, Chikusa-ku, Nagoya 464-01, JAPAN
E-mail: masai@watanabe.nuie.nagoya-u.ac.jp

†Address: Furo-cho, Chikusa-ku, Nagoya 464-01, JAPAN
E-mail: watanabe@watanabe.nuie.nagoya-u.ac.jp

2 Groups of Documents

In the real world, there are various kinds of documents such as report-form documents, newspapers, library cataloging cards, name cards, table-form documents, checks and so on. Generally, these documents can be considered from three points of view about item areas: the first is how item areas are determined; the second is what item areas exist; and the third is what positional relationships item areas have. We define document groups, document types and document classes, on the basis of features about the above viewpoints. Also, since these features are concerned with item areas which are basic units in the layout structure, we call them layout features.

In this way, we can uniformly interpret documents by three viewpoints. Also, defining groups of documents according to them makes the classification of document images effective.

2.1 Document Group

Document groups are defined according to the first viewpoint: how item areas are determined. For example, geometrical features, such as one-character-up/down, centering and so on, can be found in name cards as shown in Figure 1(b) and items are specified by them. That is to say, those features can determine item areas. On the other hand, item areas of table-form documents in Figure 1(c) are determined by vertical/horizontal line segments. For report-form documents in Figure 1(a), item areas are determined by both columns and geometrical

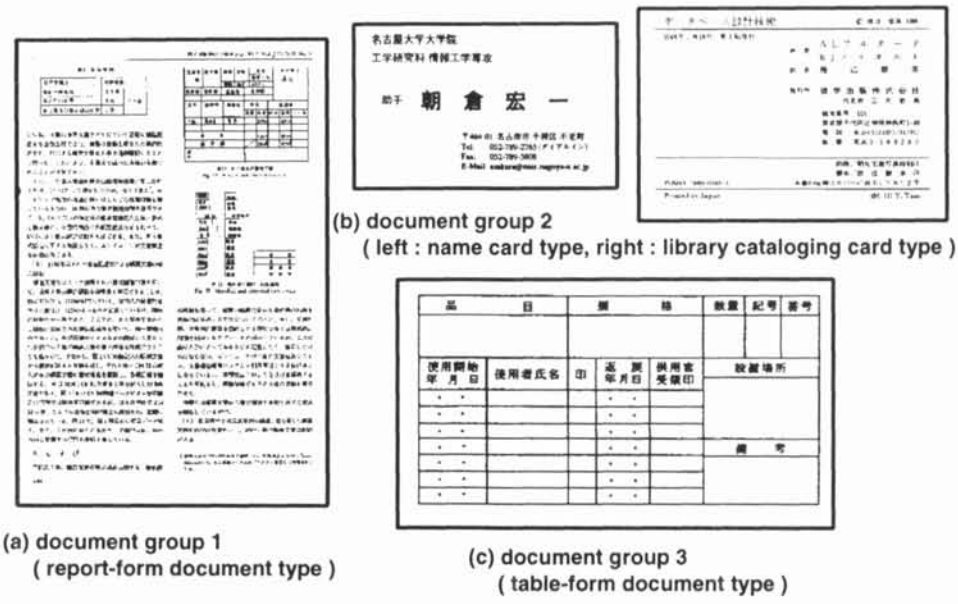

Figure 1: Examples of documents

features. In this way, item areas are determined by various factors and we can define document groups on the basis of the difference among them. Table 1 shows layout features of document groups.

Table 1: Layout features of document groups

document group	document type	layout features (How item areas are determined?)
1	report-form newspaper	·columns ·geometrical features
2	name card library	·geometrical features
3	table-form check	· vertical/horizontal line segments

report-form: report-form document
library: library catalogin card
table-form: table-form document

2.2 Document Type

Document types are defined according to the second viewpoint: what item areas exist. They individually belong to document groups and inherit layout features of document groups. Namely, document types in the same document group have the same layout features based on the first viewpoint, but ones about the second are different. For example, name card type and library cataloging card type as shown in Figure 1(b) belong to the document group 2 because their item areas are determined by geometrical features, but their logical structures are organized by items such as personal name, company's name and address in the former, and by items such as title, author and publisher in the latter. Namely, items which are allocated into item areas are different, and this indicates the number of item areas or factors which determine item areas as a result. Generally, logical structures are different in every application area. Table 2 shows layout features of document types.

Table 2: Layout features of document types

document type	layout features (What item areas exist?)
report-form	·the number of columns is about 2. ·item areas of chapter titles are determined by large white spaces.
newspaper	·the number of columns is about 10. ·item areas of sub-headlines are determined by both large white spaces and centering.
name card	·an item area of personal name is determined by the special geometrical feature.
library	·there is not an item areas of personal name.
table-form	·the number of item areas is over 2.
check	·the number of item areas is 1.

2.3 Document Class

Document Classes are defined according to the third viewpoint: what positional relationships item areas have. For example, documents which belong to the name card type have item areas which items, such as personal name, address and so on, are allocated into, and some documents have a positional relationship such that an item area of personal name is upside and one of address is downside; others do a reverse relationship. Namely, the difference of mutual positional relationships among item areas in document images defines individual document classes. Thus, documents which belong to the same document class have the same layout structures.

3 Classification Method

Our classification method is based on the verification mechanism. Namely, it is first to classify document images into groups of documents, and then

to verify classified documents by extracting layout features. Here, the first step classifies those groups by simple comparison processes. The reason that we adopt classification/verification mechanism is because we can accomplish the effective classification process. We identify document group, document type and document class for an input image on the basis of this mechanism as shown in Figure 2.

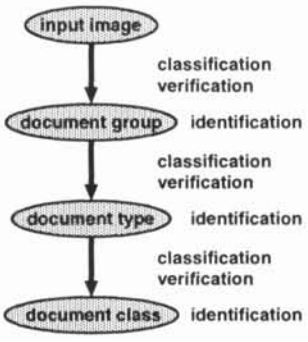

Figure 2: Classification method

4 Identification of Document Group

We identify document groups of input images as the first step of our classification method. The classification process is performed on the basis of physical features of documents, and the verification process is done by extracting layout features of individual document groups.

4.1 Classification of Document Group

We classify various documents into the corresponding document groups according to physical features. Physical features relate to the distribution of black pixels in document images. The extraction of physical features depends on the ratio of black/white pixels in document images by the simple methods, but these methods are not simple counting-up of black/white pixels, as it were. This extraction procedure deforms the original document images as a preprocessing. We analyze document images on the basis of two physical features, as shown in Table 3: the ratios of line elements and black pixels.

The ratio of black pixels in the document group 1 is large because this group generally consists of sentence-specific text data. Also, the ratio of line elements is small even if there exist line segments which organize columns, because the ratio of them in document images is small. For the document group 2, item areas are determined on the basis of geometrical features. Hence, the ratio of white pixels is large. Furthermore, since items are word-specific text data, the length of item data is short. Therefore, the ratio of black pixels is small. The ratio of line elements in the document group 3 is large because it is composed of item areas enclosed by horizontal and vertical line segments. Also, black pixels which construct line segments make the ratio of black pixels large.

Table 3: Physical features of document group

document group	ratio of line elements	ratio of black pixels
1	small	large
2	small	small
3	large	medium

4.2 Verification of Document Group

We verify individual document groups according to layout features shown in Table 1.

In the document group 1, separators such as line segments and white spaces organize columns and geometrical features such as one-character-up/down and centering are also found. And, item areas are determined by both of them. So, the verification of document group 1 is performed by checking up whether there are columns and geometrical features. In the document group 2, items are separated by white spaces and geometrical features. Namely, since items of document group 2 are composed of short text data, there is not the repetition of text lines but one-character-up/down and centering. So, the verification of document group 2 is performed by checking what text lines line up irregularly. In the document group 3, vertical/horizontal line segments organize item areas. So, the verification of document type 3 is performed by checking up whether item areas are enclosed by vertical and horizontal line segments.

5 Identification of Document Type

Here, we describe a part of identification method of document types according to layout features shown in Table 2.

The identification of report-form document type or newspaper type is accomplished on the basis of the number of columns and the existence of chapter titles or sub-headlines, which are recognized by means of geometrical features. Namely, if the number of columns is about 2 and chapter titles exist, document images are identified as report-form document type. If the former is about 10 and the latter is sub-headline, they are done as newspaper type. The identification of name card type or library cataloging card type is done by checking whether certain item (personal name in name card type), which is separated on the basis of special geometrical features, exists or not. Namely, if there are such special features, document images are identified as name

Table 4: An experimental result

document group	document type	input	identification of document group				identification of document type	
			classification		verification			
			correct	error	correct	error	correct	error
1	report-form	50	45	5	40	5	40	0
2	name card	50	50	0	50	0	43	7
	library	50	45	5	45	0	38	7
3	table-form	50	40	10	39	1	39	0
total		200	180	20	174	6	160	14
ratio [%]		—	90.0	10.0	96.7	3.3	91.9	8.1

card type. Otherwise, they are library cataloging card type. The identification of table-form document type or check type is to check up the number of item areas which are adjacent to each other. If many item areas can be found in document images, then we can identify these documens as table-form document type. Otherwise, they are check type.

6 Identification of Document Class

Here, we simply describe a part of identification method of document class based on layout features. For example, documents which belong to name card type have an item area of personal name. Document classes which belong to name card type are defined according to the mutual relationships among item areas including an item of personal name. Thus, if we can recognize what/how many item areas exist at upside or down side, we can identify document classes.

7 Experiments

Here, we show an experimental result about a part of our classification method. Sample document images are 50 sheets of report-form documents, name cards, library cataloging cards and table-form documents. Also, these images are digitalized by the image scanner with 200 dpi and 256 gray levels. Table 4 shows a result about classification of document groups according to physical features, and a result about verification of document groups and a result about identification of document types on the basis of layout features.

When we pay our attentions to the ratios of correct results at each step, we can consider that this experiment about our classification method is successful. Also, for 200 sheets of input images, 160 sheets of them are identified correctly. Thus, the ratio of the last correct identification is 80.0[%].

8 Conclusion

In this paper, we proposed an approach to classify various kinds of documents automatically as the first step for developing a general purpose document image understanding system. Also, we defined document groups, document types and document classes, and addressed a method to classify into them according to physical and layout features. Finally, we reported experimental results about our classification method.

One of our future work is to develop the reclassification mechanism for document images which are erroneous at the verification process so as to grow up the ratio of the last correct identification.

Acknowledgements

We are very grateful to Prof. T. Fukumura of Chukyo University, and Prof. Y. Inagaki and Prof. J. Toriwaki of Nagoya University for their perspective remarks, and also wish to thank Dr. Y. Sagawa, Mr. K. Asakura and our research members for their many discussions and cooperations.

References

[1] T. Watanabe, Q. Luo, Y. Yoshida, and Y. Inagaki: "A Stepwise Recognition Method of Library Cataloging Cards on the Basis of Various Kinds of Knowledge", *Proc. of 10th IPCCC*, pp.821-827(1991).

[2] T. Watanabe, H. Naruse, Q. Luo and N. Sugie: "Structure Analysis of Table-form Documents on the Basis of the Recognition of Vertical and Horizontal Line Segments", *Proc. of 1st ICDAR*, pp.638-646(1991).

[3] Q. Luo, T. Watanabe and N. Sugie: "A Structure Recognition Method for Japanese Newspapers", *Proc. of 1st SDAIR*, pp.217-234(1992).

[4] T. Watanabe, Q. Luo and N. Sugie: "Structure Recognition Methods for Various Types of Documents", *Int'l Journal of MVA*, vol.6,pp.163-176(1993).

[5] A.Yamashita and T. Amano, "A Model Based Layout Understanding Method for Document Images", *Trans. of IEICE*, **J75-D-II**,10,pp.1673-1681 (1992)(in Japanese).