MVA'90 IAPR Workshop on Machine Vision Applications Nov. 28-30,1990, Tokyo

Model based Understanding of Document Images

Koichi KISE[†] *Ken-ichi MOMOTA*[‡] *Masaki YAMAOKA*[‡] *Jun-ichi SUGIYAMA*[‡]

Noboru BABAGUCHI[‡] *Yoshikazu TEZUKA*[‡]

† *Department of Electrical Engineering,*
University of Osaka Prefecture

‡ *Department of Communication Engineering,*
Osaka University

ABSTRACT

Document image understanding is a task to generate the structured description about contents of a document. In this paper, we propose a new method of document image understanding which employs the domain specific knowledge base called document model. Document model is structural representation of constraints on the layout structure as well as the logical structure of a target document. Since the variation of the structure can be described in document model, intermediate results of understanding generally include multiple candidates. In order to generate plausible description from such candidates, we introduce the strategy of hypothesis generation and testing. From the experiments for 100 visiting cards, we demonstrate the effectiveness of our method.

1. INTRODUCTION

In recent years, we have been facing the problem of how to deal with large amount of the existing documents which are not in a *processable form* for computer systems. Document image understanding, which is a technology to extract structured description about contents from a document image, has been gaining in importance aiming at paper-free office.

In order to realize document image understanding, many efforts have been made for some sorts of documents, e.g., postal addresses[1,2], office letters[3], magazine index pages[4]. Several works try to achieve higher performance than conventional methods with the help of AI techniques. However, there seems to still remain some problems to be explored: 1) how to describe the domain specific knowledge for understanding, 2)how to deal with uncertainty of intermediate processing results.

Although the rule based approach[1] is attempted to solve the first problem, it would be insufficient to represent the hierarchy inherent in document structure. In addition, most of the existing methods take little account of the variation of the structure. On the other hand, the second problem is concerned with the control of inference. An expedient way is numeric driven inference[2,3]. The uncertainty is mapped to numerical values such as confidence values and certainty factor. The values will bring a great effect on final results, nevertheless, they may be given, in many cases, ad hoc. We consider that the inference should be controlled logically rather than numerically.

In this paper, we describe a new approach to cope with the problems. For the first problem, we propose a knowledge base called *document model* which is based on *frame* representation to describe the hierarchy and the variation straightforwardly. In document model, the knowledge about layout structure are described as well as the logical constraints about contents of a document. For the second problem, we introduce the strategy of hypothesis generation and testing. We will generate all possible results, which is obtained based on the layout structure, as hypothesis, and test them by logical constraints about contents.

2. DOCUMENT MODEL

In general, a document consists of many components which are hierarchically structured. For instance, a visiting card, which we concern here, includes components of 6 levels: document, group, subgroup, item_group, item and character. Document model is a domain specific knowledge base which represents the constraints on the layout structure of a document as well as the contents of a document. In the rest of this paper, the knowledge about the layout structure is called layout knowledge, and the knowledge about the constraints on contents is called logical constraints.

In a document image, a component can be regarded as a rectangle with an attribute, e.g., name, address and telephone for a visiting card. The hierarchy is represented as nested regions of components in the image. For the purpose of knowledge description, we restrict the hierarchy as follows: a component at a level includes lower level components which are arranged either horizontally or vertically.

We utilize frame representation to specify the layout structurally. As shown in Fig.1, each component corresponds to a distinct frame which is linked to each other with two kinds of slots: *part_of* slot to represent the hierarchy of components, and *similarity* slot to represent the difference between components and relative positions.

In order to obtain high describability and readability of document model, we employ *layout predicates* to fill the facet of slots, such as "horizontal_centering", "upper_end", "horizontal_alignment", "right_indented". Currently we use 24 predicates for a visiting card. Most of these predicates are defined as the conjunction of a pair of characteristic features that indicate width, area of a rectangle and its value. Note that the value is not a numerical one, but an *interval* to represent tolerance of the characteristic feature.

To describe the variation of the layout structure, *class-instance* relation are introduced in document model (See Fig.1). The layout predicates which are common to the instances are stored in a *class* frame, and the layout predicates which are peculiar to each instance are stored in an *instance* frame. An instance frame is connected to a class subframe with *is_a* link to inherit the features of the layout predicates in a class

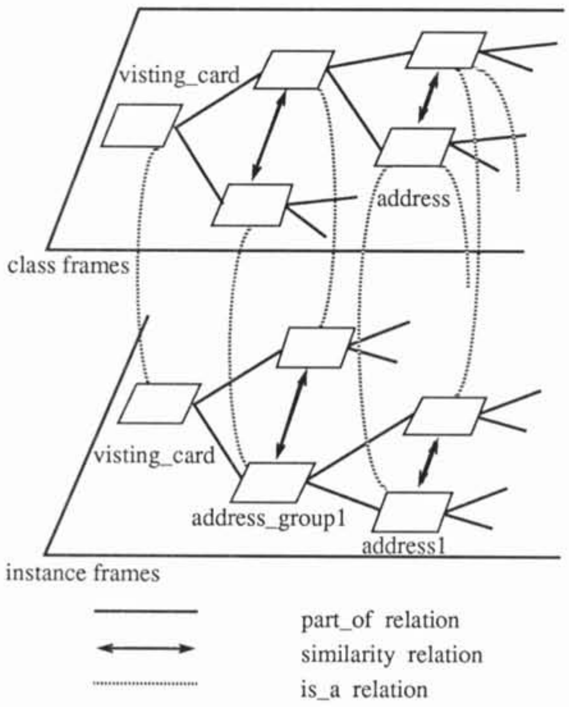


Fig. 1 Document model

frame.

As the logical constraints, a relation between words in two items are described as consistent or inconsistent. For example, it is inconsistent that "社長" (president) and "研究員" (researcher) exist in two title items. In case the consistency relation is described, the relation between words which has no description is viewed as inconsistent. This knowledge is also described in a similarity slot.

3. HYPOTHESIS GENERATION

At the hypothesis generation, our system tries to generate candidates of components described in document model using the layout knowledge. To avoid rejecting correctly extracted components, multiple candidates of components will be accepted as hypotheses. These hypotheses are generated by layout structure analysis, character segmentation and recognition. In the following, we present the method of layout structure analysis precisely. For the details of character segmentation and recognition, see [5].

The input to layout structure analysis is the basic rectangles generated by recursive projection of a document image. Since these rectangles are smaller than any other regions of components, layout structure analysis can be viewed as both merging basic rectangles to generate a component region and assigning an appropriate attribute to it.

Layout structure analysis is the recursive processing guided by the part_of relation between components. After all possible components are extracted at a certain level, next target components at their lower level will be extracted based on the extracted component. The extracted component is called *base* for the components at a lower level.

The processing at each level begins with the collection of basic rectangles included in a target base. Since the components included in the base are arranged vertically(horizontally), the basic rectangles can be merged horizontally(vertically). For convenience, we assume that the rectangles should be arranged vertically.

Focusing on the rectangle x which is located on the top of the base region, a set of components $\{f_1, f_2, ...\}$ whose element may include x can be constructed by referring the descriptions of frames at the target level. In case there is no variation in the layout structure, only one component may be selected.

Candidates of a component can be generated in the top-down manner as follows: 1) assume one component f_i in the set, 2)generate the regions by merging rectangles downward from x, and assign the attribute of f_i to the region. Note that the generated candidates should satisfy all of the layout predicates in the frame for f_i. Since the layout predicates includes the interval values for the features, multiple candidates are generally obtained.

In order to generate the rest of components included in the base, we select one of the candidates as an assumption, and then regard the rectangle which is adjecent to the region of the assumption as x. Based on the rectangle x, the next candidates of a component can be generated in the same manner mentioned above. This processing is continued until no rectangle remains in the base region.

If all of the rectangles are included in assumptions, the set of the assumptions are called hypothesis. After the hypothesis is obtained, backtracking is entailed to generate other hypotheses. In case no candidate can be obtained at any stage of the processing, backtracking is also entailed to select the alternative assumption.

In general, multiple hypotheses may be obtained from a base. It can be represented as follows:

$$b \Rightarrow \{h_1, h_2, ...\} \qquad (1)$$

$$h_i = c_{i1} \ and \ c_{i2} \ and \ c_{i3} ... \qquad (2)$$

where b, h_i and c_{ij} denote a base, a hypothesis and a candidate of a component, respectively. The Eqs. (1), (2) indicate that: (1) a set of mutually inconsistent hypotheses is generated from a base, (2) a hypothesis is represented as the conjunction of candidates.

4. HYPOTHESIS TESTING

At the hypothesis testing, our system tests the hypothesis through generation of feasible contents of the components. A description about contents of an item are generated by word sequence recognition to test item candidates individually. Subsequently these descriptions are grouped to generate descriptions about contents for the upper level components. The newly generated descriptions are checked by logical constraint satisfaction to test the candidates at an upper level.

Word sequence recognition

Contents of an item can be considered as a consistent word sequence. To represent it, we employ *connectivity* between words. A word P is *connectable* to a word Q if a concatenated sequence PQ is admissible in an item. This information is stored in a word dictionary which is made for each item.

The input to word sequence recognition is both item and character candidates. Since an item candidate has its attribute to select an appropriate word dictionary, a word sequence can be recognized in a top down manner. This enables us to reduce the number of dictionary words to be matched.

In order to deal with multiple candidates of character regions, we utilize directed acyclic graph structure whose nodes and arcs represent character regions with candidate categories, and the reading order of characters, respectively. To obtain an appropriate word sequence even if the region of an item candidate is incorrect, the starting point of a word sequence is assumed from the top node to the bottom. A word sequence is generated in such a way that the graph is traversed to match words in the depth-first manner. In order to reduce the search space, only the dictionary words which are connectable to the formerly matched word are used for matching. The generated word sequence can be regarded as the description about contents of the item.

Testing a hypothesis for the items is also achieved through word sequence recognition. In case that no word sequence is generated, item candidate can be determined to be inconsistent. Thus a hypothesis which includes such an item candidate is also inconsistent by Eq.(2). If the region of a word sequence is different from that of an item candidate, the region of a word sequence is regarded as an appropriate region. In most cases, multiple word sequences are obtained from an item candidate.

Logical constraint satisfaction

In order to test an upper level candidates, a description about contents of an upper level component is generated based on that of an item. The processing is guided by the history of hypothesis generation stored in the forms of Eq.(1),(2), and continued up to a document level in the reverse order of hypothesis generation.

A candidate c_{ij} in Eq. (2) generally has a set of mutual inconsistent descriptions $D_{ij}=\{d_{ij1}, d_{ij2}, ...\}$. At the beginning of logical constraint satisfaction, c_{ij} and d_{ijk} correspond to an item candidate and a word sequence, respectively. In order to generate the description about contents of the base b, the description d_{ijk} should be selected from the D_{ij} for all item candidates that belong to the hypothesis h_i.

The newly generated description, which represents the contents of the base, can be described as follows:

$d_{base}=d_{i1k}$ *and* d_{i2l} *and* ...

where d_{ijk} should be mutually consistent. To verify the consistency of the generated description, logical constraints in document model are checked for each pair of descriptions (d_{ijk} , d_{ilm}) in d_{base}. If d_{base} includes an inconsistent pair, it is determined to be inconsistent.

Testing a hypothesis is also achieved through generation of the description. In case the hypothesis h_i has no consistent d_{base}, it turns out to be inconsistent. Moreover, if the base b has no consistent description, it should be rejected. In most cases, the multiple consistent descriptions are generated for the base b.

In case multiple descriptions are obtained at the document level, the most plausible contents is selected based on the average of similarity values for characters.

5. EXPERIMENTAL RESULTS

To verify the performance of our method, experiments were conducted for 100 samples of visiting card images. The results are shown in Table 1 and 2.

Hypothesis generation

The performance of hypothesis generation is measured by both the average number of generated candidates and the reliability rate. The reliability rate is defined as follows:

$$1 - \frac{\text{[No. of components included in the candidates]}}{\text{[No. of components]}}$$

Note that a high reliability rate accompanied with a small average number of candidates suggests high performance. For the group level, the reliability rate of 100% was obtained, since the layout for groups includes no variation. For the subgroup and item_group level, some components in address group were not included in the candidates. This is due to irregular variation of the layout structure which is beyond the knowledge description.

For the item level, the candidates did not inlcude 12 items. Most of the errors are also caused by the irregular variation. Note that almost all the items in the erroneously extracted item_groups were correctly extracted, because the regions of the item_groups include the correct regions of items.

For character level, 98.8% of character regions and 94.2% of its attributes were included in the candidates. One of the major cause of failures is the distortion of small characters resulting from low resolution of scanning. Another cause is the existence of designed characters in organization items. This is because we could take no account of them in making the dictionary for character recognition. In order to show the ability of character recognition alone, we measured the correct recognition rate; only 78.7% of characters were correctly recognized.

Table 1 Results of hypothesis generation and testing

Level	No. of components	Generation			Testing
		Ave. no. of candidates		Reliability rate	Extraction rate of components
document	100	——		——	——
group	300	1.0		100%	99.0%
subgroup	667	2.1		98.2%	99.6%
item_group	170	3.9		95.3%	95.3%
item	947	10.2		99.8%	90.4%
character	8811	Region	1.3	98.8%	93.0%
		Category	8.3	94.2%	

Hypothesis testing

One of the role of hypothesis testing is to select an appropriate component from multiple candidates. To demonstrate the performance, the extraction rate of components is shown for each level in Table1. For the character level, the extraction rate corresponds to the recognition rate. High performance of our method was verified throughout all levels. In particular, the recognition rate of 78.7% was improved to 93.0%. It is also worthy of note that the extraction rate for subgroups was improved by hypotheses testing. It indicates that word sequence recognition is effective to correct the regions of items. Figure 2 illustrates the example of hypothesis generation and testing for items in address group. 13 hypotheses including 19 candidates are generated in (a), and correct items are selected by testing in (b). As shown in the figure, our method is flexible enough to extract components from the image with the complicated layout structure.

Another role of hypothesis testing is to generate descriptions about contents of the items. This role is of great importance in document image understanding. The results are shown in Table2. The understanding rate indicates the number of correctly generated descriptions per the number of items. The description is regarded to be correct if all characters in an item are correctly recognized. Although the description is determined as an error even if one character in an item is incorrectly recognized, good results were obtained except for organization, telephone, fax, telex, and postcode items. Errors in organization items are also caused by the designed characters. On the other hand, the major reason of errors for other items is the difficulty of word sequence recognition; since most of the words in these items are numerals, they are connectable to any other words. Except for these items, our method is robust enough to generate correct contents.

6. CONCLUSION

We have presented the model based approach of document image understanding. The knowledge about the layout structure as well as the logical constraints are described in document model. We realize high expressivity, describability and readability of document model with the aid of the frame representation and layout predicates. To compile plausible intermediate results of understanding, the strategy of hypothesis generation and testing is introduced. In our method, hypothesis generation plays the role to restrict the descriptions roughly by the attributes. Full descriptions of components are generated through the hypothesis testing based on connectivity and consistency between words. The experimental results demonstrate that our method is effective to the documents with the complicated structure, although there is still room for further refinement of character recognition and word sequence recognition.

REFERENCES

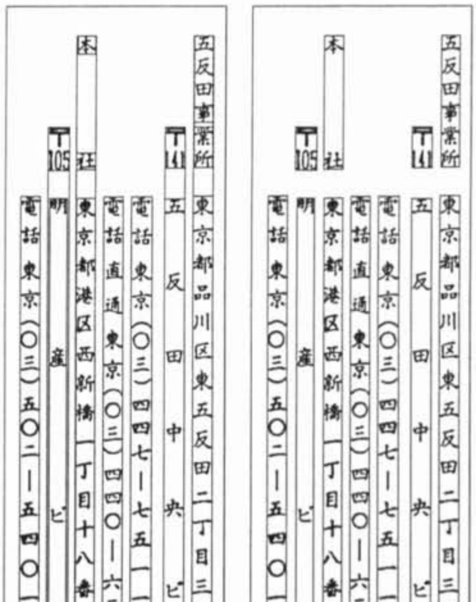
[1]D.Niyogi and S.N.Srihari:"A Rule-based System for Document Understanding", Proc.AAAI-86, pp.789-793, 1986.

[2]C.H.Wang, P.W.Palumbo and S.N.Srihari: "Performance of a System to Locate Address Blocks on Mail Pieces", Proc.AAAI-88, pp.837-841, 1988.

[3]A.Dengel and G.Barth:"ANASTASIL:A hybrid knowledge-based System for Document Layout Analysis", Proc.IJCAI-89, pp.1249-1254, 1989.

[4]F.Esposito, D.Malerba, G.Semeraro:"An Experimental Page Layout Recognition System for Office Document Automatic Classification: An Integrated Approach for Inductive Generalization", Proc. 10th ICPR, pp.557-562, 1990.

[5]K.Kise, K.Yamada, N.Tanaka, N.Babaguchi, Y.Tezuka: "Visiting Card Understanding System", Proc. 9th ICPR, pp.425-429, 1988.



五反田事業所
東京都品川区東五反田二丁目三番五号
〒141
五反田中央ビル
電話東京(〇三)四四七一七五一一(代表)
電話直通東京(〇三)四四〇一六二四一番
本社
東京都港区西新橋一丁目十八番十七号
〒105
明産ビル
電話東京(〇三)五〇二一五四〇一(代表)

(a) hypothesis generation (b) hypothesis testing

Fig. 2 Actual results

Table 2 Results of understanding

	organization	position	title	name	header	address	postcode	telephone	fax	telex	total
No. of components	100	136	130	100	41	119	119	167	10	25	947
Understanding rate	81.0%	94.9%	98.5%	89.0%	87.8%	93.3%	82.4%	74.3%	20.0%	84.0%	86.5%