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

Identification of Item Fields in Table-form Documents with/without Line Segments

Tsuneo SOBUE and Toyohide WATANABE *
Department of Information Engineering,
Graduate School of Engineering,
Nagoya University

Abstract

Many methods to recognize the layout structures of table-form documents have been proposed until today. Most of them interpret table-form document images using the knowledge which is adaptable to the specification of layout structures of individual table-form documents. H.Naruse et al. proposed a successful method, based on neighboring/connective relationships among item fields in table-form documents, to recognize the layout structures of table-form document images. Our method is an advanced version of their method. In comparison with their method, our method further recognizes the layout structures of table-form document images whose item fields are not surrounded with real line segments. The main idea in our approach is to transform the original table-form document images so as to be adaptable to their method.

1 introduction

Table-form documents are generally characterized as a collection of item fields which are surrounded with horizontal and vertical line segments. On the basis of this characteristic, H.Naruse et al. proposed the successful method to recognize the layout structures of table-form documents(we call this method Naruse method, hereafter)[1]. However, we can observe various kinds of table-form documents in the real world: some are not always surrounded with horizontal and vertical real line segments. H.Kojima et al. classified table-form documents into two types: open-type and closed-type[2]. In the closed-type of table-form documents, every item field is always surrounded with some line segments. While, in the open-type of table-form documents any item fields are not surrounded with line segments. Naruse method is applicable to only some of closed-types of table-form documents. Also, Naruse method does not validate the recognized results. Validating the recognized results is very important[3].

In this paper, we address an experimental method which recognizes such table-form documents successfully as an enhanced version of Naruse method. The characteristics in our method are to apply the preprocessing to the original table-form document images so as to be adaptable to Naruse method, to expand the verification process in both the preprocessing and post-processing with a view to validating the processing results, and to improve the structure description tree in order to represent the knowledge of layout structure and description rules for applying to various forms, description, organizations, etc. Fundamentally, our method makes use of an approach which supplements newly assumed line segments to separate item fields independently or redraws real line segments in place of other kinds of line segments so that the document images should be applied well by Naruse method. With a view to verifying whether such supplemented line segments are consistent to the original relationships among line segments or whether such supplemented line segments should be interpreted well by the knowledge of modified structure description tree, the recognized results are validated.

2 Representation of knowledge

Table-form documents are composed of the meaningful item fields which are surrounded with horizontal and vertical line segments. These item fields compose complicated structures under their neighboring and constructive relationships. Naruse method adapts a binary tree, so-called the structure description tree, to represent the layout structures of table-form documents logically. The structure description tree is composed of the global structure tree and local structure trees. The global structure tree describes the neighboring relationships among blocks which are meaningful sets of item fields. The nodes in the global structure tree correspond to

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

these blocks. The left and right edges link the blocks which are located to the lower and right sides of current block respectively as shown in Figure 1. The local structure tree describes the internal structures of blocks. The nodes in the local structure tree correspond to the item fields and are divided into the vertical node 'v', the horizontal node 'h' and the terminal node 't' as shown in Figure 2. However, it is impossible to apply Naruse method to table-form documents whose item fields are not surrounded with horizontal and vertical real line segments. This is because Naruse method is based on the assumption that every item field is surrounded with horizontal and vertical real line segments.

We improve the structure description tree so as to represent various types of layout knowledge. In order to be adaptable to table-form documents whose item fields are not surrounded with horizontal and vertical real line segments, it is necessary to consider the kind of line segments in addition to the approach in Naruse method. Therefore, we add the information, which indicates the kinds of line segments, such as real line segment, broken line segment and so on, to vertical and horizontal nodes in the local structure tree. It is easy to know the kinds of line segments which are frames of individual item fields on the basis of this information. Further, Naruse method does not validate the recognized results. However, it is important to validate the recognized results, and the knowledge for validating the recognized results is needed. We adapt the subordinate relationships as the knowledge for validating the recognized results. The subordinate relationships are the relationships among meaningful sets of item fields[4]. For example, in case that the data field is subordinate to the name field the relationship is illustrated in Figure 3.

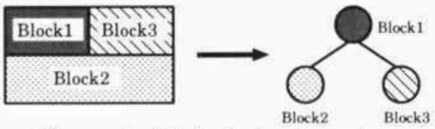


Figure 1: Global structure tree

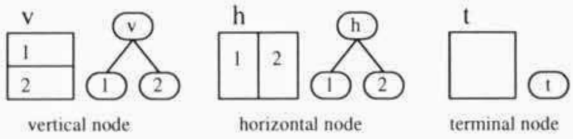


Figure 2: Node types in local structure tree

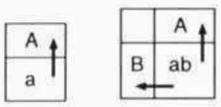


Figure 3: Subordinate relationship

3 Recognition of layout structure

Our layout recognition system of table-form documents is divided into two modules: structure analysis and structure recognition. The structure analysis module extracts the upper-left corners of individual item fields, and the upper-left corners are interpreted with the structure description tree. Then, the recognized result is validated.

3.1 Structure analysis

3.1.1 Preprocessing

Naruse method detects the upper-left corners as the characteristic points of item fields. However, it is impossible to detect the upper-left corners of item fields, which are not surrounded with horizontal and vertical real line segments, in Naruse method. Therefore, we apply the preprocessing to the original table-form document images so as to be adaptable to Naruse method. The preprocessing arranges the layout structures of the original table-form documents so that individual item data can be surrounded with horizontal and vertical real line segments completely. We apply the preprocessing for the following types of table-form documents.

- table-form documents whose corners are not orthogonal

  Naruse method extracts the upper-left corners which are generated as cross-points between horizontal and vertical real line segments. Therefore, Naruse method does not extract the upper-left corners in case that the corners are not orthogonal, because there are not the cross-points. We modify the corners, which are not orthogonal, to the corners which are orthogonal in the preprocessing as shown in Figure 4.

- table-form documents whose item fields are surrounded with broken line segments

  In this case, Naruse method does not extract the upper-left corners, because the detection mechanism of horizontal and vertical line segments in Naruse method is not adaptable to broken line segments. Then, we redraw real line segments in place of broken line segments in the preprocessing as shown in Figure 5.

- table-form documents in which there are not completely separators among item fields

  In the same as the case of table-form documents whose item fields are surrounded with broken line segments, Naruse method for this case does not extract the upper-left corners. Then, we supplement newly assumed line segments by the separation among item fields in the preprocessing as shown in Figure 6.

Of course, in the supplement of line segments it is not sure that the assumed line segments should always be correct. So, it is needed to verify whether the supplemented line segments can be consistent to original relationships among line segments.

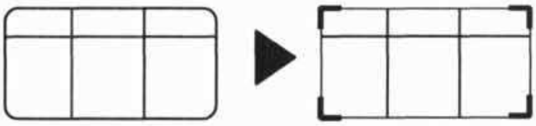


Figure 4: Round corners

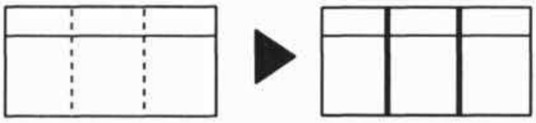


Figure 5: Broken line segments

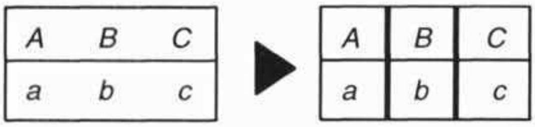


Figure 6: Non-separation among item fields

3.1.2 Detection of upper-left corners

We detect the upper-left corners of individual item fields, which are generated as cross-points between horizontal and vertical line segments, in the bottom-up means. It is very smart in Naruse method to look upon the upper-left corners as individual item fields which are rectangular blocks. This is because the relationships among the upper-left corners are identical to the representation of various kinds of layout structures abstractly, which are constructed from neighboring relationships among rectangular blocks.

3.2 Structure recognition

In the structure recognition module, the upper-left corners detected in the structure analysis module are interpreted with the structure description tree from the most uppest-left point to the most lowest-right point along the edges in the structure description tree. However, it is not sure that the detected upper-left corners are correct. Thus, we verify whether the upper-left corners are correct with the information about the kinds of line segments as frames of individual item fields, in the structure description tree. The verification process is as follows:

1. Verify whether the rectangular block, which the current upper-left corner represents, can satisfy the information about the kinds of line segments with the corresponding node in the local structure tree.
2. If the rectangular block is not satisfy, the rectangular block, which the current upper-left corner indicates, is magnified until the rectangular block satisfies the information as shown in Figure 7.
3. If the rectangular block satisfies the information, the current upper-left corner is corresponded to the node in the local structure tree. Thus, the next upper-left corner is verified.

The incorrect upper-left corners generated irregularly from noises or characters are rejected effectively because the correspondence between upper-left corners and some nodes in the structure description tree is not consistent.

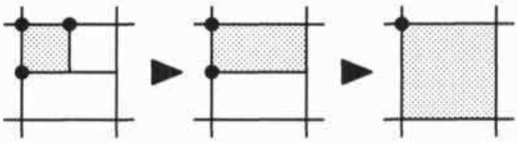


Figure 7: Magnification of rectangular block

4 Validation

It is very important to validate the recognized results. We validate the recognized results with the information about the subordinate relationships in the structure description tree. It is sure that item fields with subordinate relationships are rectangular blocks. Therefore, if item fields with subordinate relationships are not rectangular blocks in the recognized results, the recognized results are mistaken and it is necessary to modify the correspondence between the upper-left corners and the structure description tree, as shown in Figure 8. Thus, the upper-left corners are corresponded to the structure description tree so that item fields with subordinate relationships are rectangular blocks, and we re-recognize the layout structures of table-form documents. Re-recognizing the layout structures of table-form documents is repeated until the recognized results are correct completely.

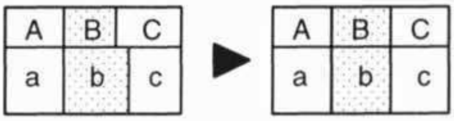
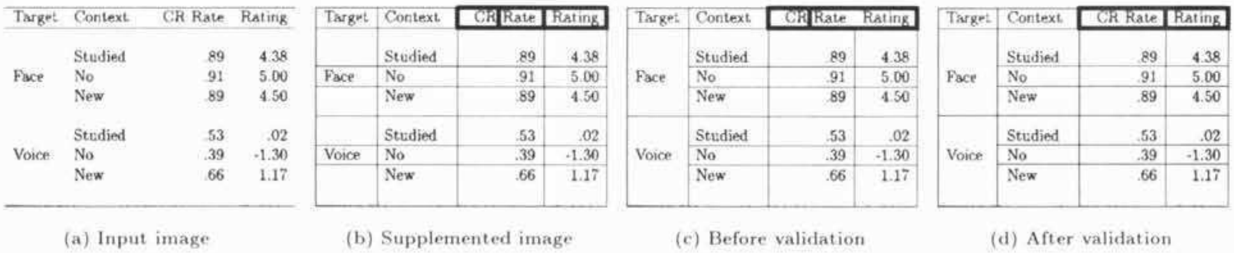


Figure 8: Validation



Target	Context	CR Rate	Rating
Face	Studied	.89	4.38
	No	.91	5.00
	New	.89	4.50
Voice	Studied	.53	.02
	No	.39	-1.30
	New	.66	1.17

(a) Input image

Target	Context	CR Rate	Rating
Face	Studied	.89	4.38
	No	.91	5.00
	New	.89	4.50
Voice	Studied	.53	.02
	No	.39	-1.30
	New	.66	1.17

(b) Supplemented image

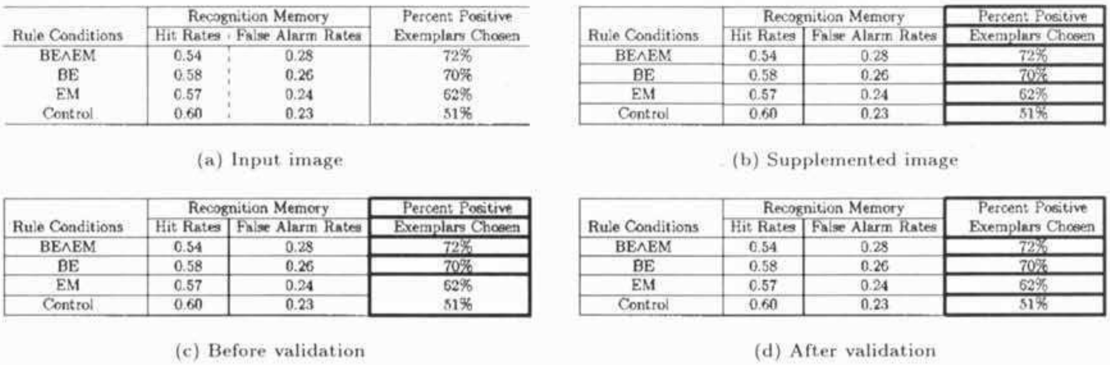
Target	Context	CR Rate	Rating
Face	Studied	.89	4.38
	No	.91	5.00
	New	.89	4.50
Voice	Studied	.53	.02
	No	.39	-1.30
	New	.66	1.17

(c) Before validation

Target	Context	CR Rate	Rating
Face	Studied	.89	4.38
	No	.91	5.00
	New	.89	4.50
Voice	Studied	.53	.02
	No	.39	-1.30
	New	.66	1.17

(d) After validation

Figure 9: A successful example 1 of validation



Rule Conditions	Recognition Memory		Percent Positive
	Hit Rates	False Alarm Rates	Exemplars Chosen
BE∧EM	0.54	0.28	72%
BE	0.58	0.26	70%
EM	0.57	0.24	62%
Control	0.60	0.23	51%

(a) Input image

Rule Conditions	Recognition Memory		Percent Positive
	Hit Rates	False Alarm Rates	Exemplars Chosen
BE∧EM	0.54	0.28	72%
BE	0.58	0.26	70%
EM	0.57	0.24	62%
Control	0.60	0.23	51%

(b) Supplemented image

Rule Conditions	Recognition Memory		Percent Positive
	Hit Rates	False Alarm Rates	Exemplars Chosen
BE∧EM	0.54	0.28	72%
BE	0.58	0.26	70%
EM	0.57	0.24	62%
Control	0.60	0.23	51%

(c) Before validation

Rule Conditions	Recognition Memory		Percent Positive
	Hit Rates	False Alarm Rates	Exemplars Chosen
BE∧EM	0.54	0.28	72%
BE	0.58	0.26	70%
EM	0.57	0.24	62%
Control	0.60	0.23	51%

(d) After validation

Figure 10: A successful example 2 of validation

5 Experiments

Here, we show recognized results through some experiments in our prototype system. Input images are digitalized with 200 dpi and 256 gray levels. Figure 9(a) and Figure 10(a) show examples of input images. Figure 9(b) and Figure 10(b) show the images with supplemented/redrawed line segments. We observe that erroneous line segments are supplemented. Figure 9(c) and Figure 10(c) show the recognized result images before validation, and Figure 9(d) and Figure 10(d) show the recognized result images after validation. Erroneous line segments are rejected, and subordinate relationships among item fields are correct in Figure 9(d) and Figure 10(d). The experimental results make sure that our method is very successful to recognize various types of table-form documents.

6 Conclusion

In this paper, we addressed the method, which is an enhanced version of Naruse method, to recognize various types of table-form documents. As a result, our method was very adaptable to various kinds of table-form documents, whose item fields are surrounded with/without various kinds of line segments.

Acknowledgments

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 suggestions, discussions and cooperations.

References

[1] H.Naruse, T.Watanabe, Q.Luo and N.Sugie: "A Structure Recognition Method of Table-Form Documents on the Basis of the Information of Line Segments", *Trans.of IEICE I*, Vol. J75-D-II, No. 8, pp. 1372–1385 (1992) [in Japanese].

[2] H.Kojima, Y.Kiyosue and T.Akiyama: "A Study on Table Recognition with Complex Structure", *Proc.of 37th IPSJ*, Vol. 6W-8, pp. 1660–1661 (1988) [in Japanese].

[3] T.Watanabe and T.Fukumura: "A Framework for Validating Recognized Results in Understanding Table-form Document Images", *Proc.of the ICDAR'95*, Vol. II, pp. 536–539 (1995).

[4] Q.Luo, T.Watanabe and N.Sugie: "Automatic Acquisition of Layout Knowledge for the Structure Recognition of Table-Form Documents", *Trans.of IEICE I*, Vol. J76-D-II, No. 3, pp. 534–546 (1993) [in Japanese].