INTERNATIONAL WORKSHOP ON INDUSTRIAL APPLICATIONS OF
MACHINE INTELLIGENCE AND VISION (MIV-89), Tokyo, April 10–12, 1989

# Automatic Extraction and Classification of Data Items from Library Cataloging Cards by a Knowledge-based Approach

Toyohide WATANABE, Qin LUO, Masahiro MIZOGAMI, Yuuji YOSHIDA and Yasuyoshi INAGAKI

Department of Information Engineering, Faculty of Engineering,
Nagoya University
Furo-cho, Chikusa-ku, Nagoya 464-01, JAPAN

***** abstract *****

It is important to store the library information in the machine-readable form into the computer systems( e.g. the library information systems ) effectually with a view to extending the library information managements and services. Today, such a task can be satisfied with many costs and heavy man-powers. Therefore, it is very desirable to develop some effective method.

In this paper, we report our experimental approach to extract and classify data items automatically from library cataloging cards. A basic strategy concept in our approach is to utilize various kinds of knowledge cooperatively, concerning cataloging cards: structure information of the card description, relationship information among data items, format information of data values and so on. In comparison with many traditional character recognition approaches, our approach is adaptable to even cataloging cards, composed of blurred and indistinct characters and/or described by various layout structures, without difficulty.

***** keywords *****

description of layout structure, description of relationships among data items, multi-level knowledge of objects, library cataloging cards, character recognition

## 1. INTRODUCTION

Library cataloging cards in each library have been necessarily used as the indicator to manage and find up individual books since the establishment. They include mini-max information about the books in spite of the narrow space. It is the most basic requirement to store them into computer files as machine-readable data with a view to making various kinds of library information services/managements powerful. However, it is tremendously wasteful and tiresome to compose machine-readable data directly from the cataloging cards. This task is a mechanical transformation process of a sort, in addition to the voluminous number of cataloging cards which have equiped until now. Therefore, a more effectual data composition method must be developed shortly.

In this paper, we propose an experimental approach to extract printed characters from the cataloging cards successfully. Our approach differs from many traditional approaches[1, 2] studied as one application of the subjects of the character recognition or the pattern recognition. In our approach, the artificial intelligence techniques[3] are effective on the basis of distinguishment multi-level features, attended inherently to the cataloging cards: the knowledge such as structure information of the card description, relationship information among the data items, characteristic information of the data components and format information about each data value is very useful to understand the properties of objects. The basic strategy concept in our approach is to make use of various kinds of knowledge cooperatively according to each processing phase though the traditional approaches have been concentrating on the adaptation of character recognition techniques.[4]

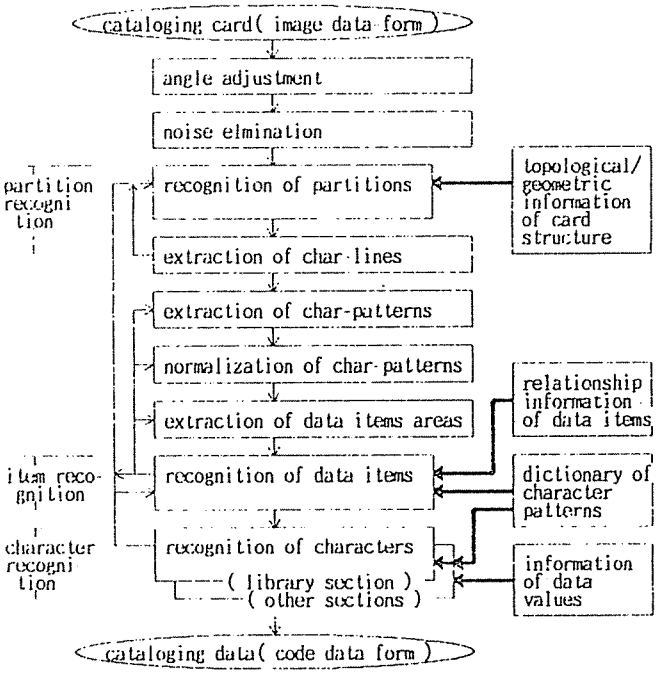



Fig.1 Processing flow in our approach

## 2. OUTLINE OF APPROACH

Although the cataloging cards are designed with different layout structures in each library, the numbers and kinds of data items, accommodated into the cataloging cards, are almost similar.[5] Moreover, the layout structures and description formats of the cataloging cards are specified under the same framework in each library, except the library-dependent management data. Namely, the cataloging cards are always composed on the basis of common constructive rules, standard description formats and topological/geometric relationships among individual data items. If such knowledge could be applied successfully with respect to extracting and classifying data items from the cataloging cards automatically, we can investigate a more effective method in point of the flexibility and adaptability of recognition strategies. In the traditional approaches, such an extraction and classification strategy has been adapting the character recognition techniques with the quantitative analysis of characteristic information about character patterns.[6] While, in our approach the strategy does not only concentrate on the character recognition, but divides the recognition process into several processing phases: the classification of layout segments; the recognition of mutual relationships among data items; the identification of data items; the character recognition and so on.

We show the general processing flow of our approach in Fig.1. In Fig.1, the particular object recognition procedures and usual

image processing routines are working cooperatively. The former controls the latter successively, and transfers the clearly interpreted objects to the latter with newly generated information. Of course, such control mechanisms are also applicable to mutual relationships among the object recognition procedures. The strategy principle in our approach is that each object recognition procedure always interprets the newly properties of partitioned objects with its own knowledge, in addition to the knowledge generated by the previously executed procedures. Therefore, the properties of recognized objects are stepwise refined according to the progress of processing phases. Namely, the successive object recognition procedures refine the insufficient results so as to be consistent with the well defined properties under cooperative mechanisms, even if the results in some phase were not correctly recognized or could not be sufficiently identified. 3 types of object recognition procedures are prepared in our approach: one is for the segmentation of data items' areas; one is for the identification of each data item; and one is for the image-code transformation of data values.

Our approach based on the knowledge-based techniques is more flexible and adaptable than the traditional approaches applied only by the character recognition techniques. This is because our approach applies the recognition tasks to the restrictive object domains interpreted continuously by various kinds of knowledge, concerning the layout structure of cataloging cards, the relationships among data items, and the data formats. Though the traditional approaches concentrate on matching with the characteristic information of character patterns. For example, our approach is useful even for the following cases:

(1) printed characters are indistinct and blurred;
(2) cataloging cards are dirty with noises;
(3) printed character lines are bumpy;
(4) data formats and layout structures are not always uniform.

## 3. DATA ITEMS OF LIBRARY CATALOGING CARDS

The cataloging cards accommodate various kinds of data items, in addition to the bibliography information such as title, author-statement, name-of-publisher, date-of-publication, pagination and so on. These data items are arranged on the basis of the NCR ( Nippon Cataloging Rules ) in case of Japanese cataloging cards. Moreover, these are allocated into the layout structure predefined by the standard description rule of each library. For example, in the University Library of Nagoya University, the normal data items are allocated geometrically as illustrated in Fig.2. We can divide these data items into 4 information sections as shown in Table 1. The data items in the foreign cataloging cards are also composed under similar frameworks.

In Fig.2, we can observe explicitly that the data items listed in Table 1 occupy more or less distinct partitions, corresponding to the category. Moreover, the mutual sequence and topological/geometric locations among data items are not only assigned regularly, but also the data values and data formats are usually determined in advance, except some data items derived from the bibliographic information. At least, we will be able to recognize the cataloging information easily if we interpreted the data items in cataloging cards step by step, using the knowledge about the standard description rules.

## 4. SEGMENTATION OF DATA ITEMS

We often experience that we can understand implicitly the meaning of an entity by our inference, based on some knowledge about the environment of the entity, even if we did not know the entity as it was. This means that the context information takes important roles for the object recognition. From such a

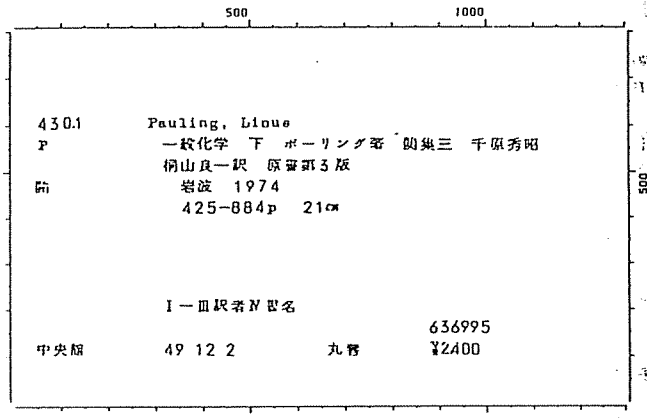



Fig.2 Cataloging card in University Library of Nagoya University

Table 1 Data items of library cataloging card

| | category | | data items | area |
|---|---|---|---|---|
| I | call-number section | | classification-number, usage author-mark | ① |
| II | location section | | location | ② |
| III | bibliography section | heading | author | ③-1 |
| | | entry | title, subtitle, edition, author-statement | ③-2 |
| | | imprint | place-of-publication, name-of-publisher, date-of-publication | ③-3 |
| | | collation | pagination, size, illustration | ③-4 |
| | | series | series | ③-4 |
| | | note | note | ③-5 |
| IV | management section | tracing | subject-heading | ④-1 |
| | | registry | registry-number | ④-2 |
| | | supplement | price, registry-date, bookstore | ④-3 |

recognition point of view, we adapt an experimental approach to extract characters from the cataloging cards on the basis of the context inference. The cataloging cards have their own layout structures, as discussed in the previous section. We can distinguish the compositive partitions of the layout structure hierarchically with the characteristics of data items or the relationships among data items. Such structures may be specified mandatorily by a top-down description. Our procedure of a structure description must be designed with the following requirements:

(1) it is possible to describe topological/geometric relationships among data items;
(2) it is flexible enough to apply a structure description to various types of processing objects;
(3) it is easy to specify layout structures.

Generally, there are variations in the geometric configuration because the length of data items is variable. Thus, we can not represent the layout structures by using the absolute values as the start positions of printing areas, the data length and so on. For example, the strings in some data item may occupy 2 or more lines even if in many cases it was 1 line.

From a viewpoint of these requirements, we designed a hierarchical description method based on the horizontal and vertical relationships among partitions of data items. The information specified in our structure description can be represented as a tree structure of a sort conceptually. For example, the representation of a layout structure is illustrated

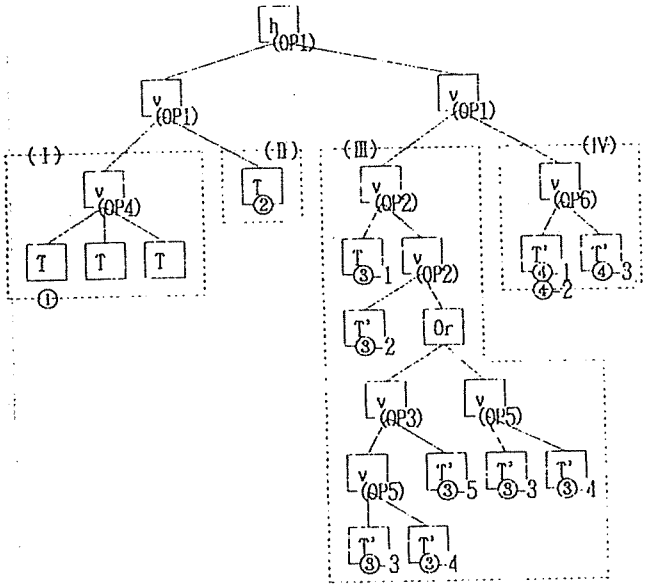
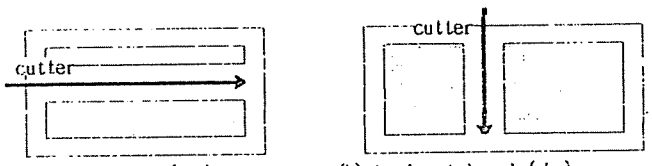



Fig.3 Example of tree structure

graphically in Fig.3, as the tree structure. In Fig.3, we perceive that individual non-terminal nodes are characterized by some attribute symbols such as "h", "v", "Or", "T" and "T'". Moreover, non-terminal nodes accompany additional strings such as "OP1", "OP2" and so on. On the other hand, terminal nodes are indicated by the numbers such as ①, ②, ③-1, ···· and ④-3. These numbers represent distinct areas in Table 1. These symbols and strings characterize the meaning of each node. The attribute symbols define the layout structure of objects to be partitioned hierarchically. While, the additional strings point out the segmentation operators to divide partition objects under the topological/geometric relationships.

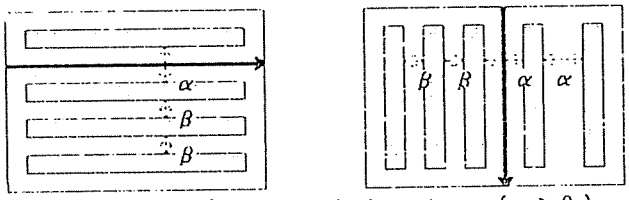
The non-terminal node "h", which represents the horizontal mode, indicates that one processing object must be divided into the left-to-right partitions by the vertical cutter, while the node "v" as the vertical mode divides one processing object into the top-to-down partitions by the horizontal cutter. The cutters work effectually with help of the segmentation operators ( e.g. "OP1", "OP2", ··· ), attended to their own nodes. The node "Or", which stands for the selection mode, indicates several candidate strategies for the next segmentation. Namely, this informs the strategy procedure of the possibility of different division ways without performing the practical segmentation. We show the roles of these modes in Fig.4. The terminal nodes "T" and "T'" indicate the end of the segmentation process, and represent the area of one data item( by "T") and 1 or more data items( by "T'").

The individual non-terminal nodes hold 4 information cells: ( MOD, SNUM, OP, CO ). The field "MOD" indicates the node type such as "h", "v" and "Or" as explained above. The field "SNUM" counts the number of the following lower nodes. The field "OP" represents the practical segmentation operator based on the topological/geometric relationships among partitions. Finally, the field "CO" accommodates the coordinate values for the partition size in which the segmentation operator must evaluate the partition structure. The segmentation operators are shown in Fig.5, under the topological/geometric characteristics among partitions. These operators inform "CO"s in the lower nodes of the sizes of newly segmented partitions. The operator in the lower node judges the next partition structure in the partitioned area indicated by its own "CO". These processes are repeated

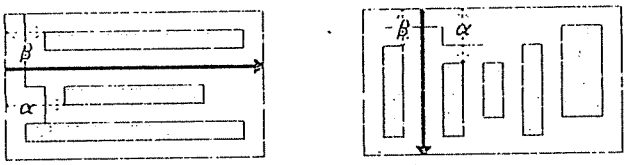



(a) vertical node( v ) (b) horizontal node( h )

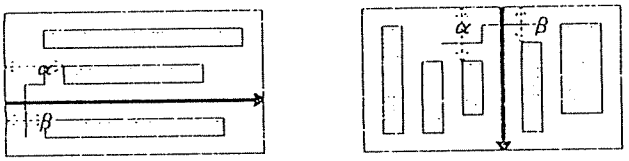
Fig.4 Non-terminal nodes "h" & "v"



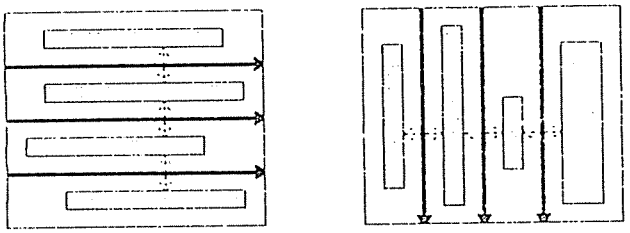

(a) OP1: generation of 2 partitions by largest space( $\alpha > \beta$ )



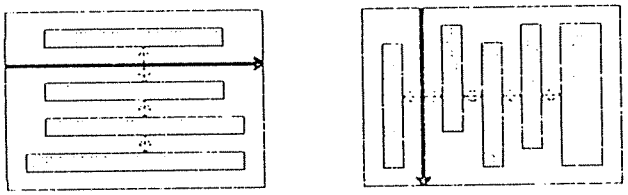

(b) OP2: generation of 2 partitions in $\alpha > \beta$



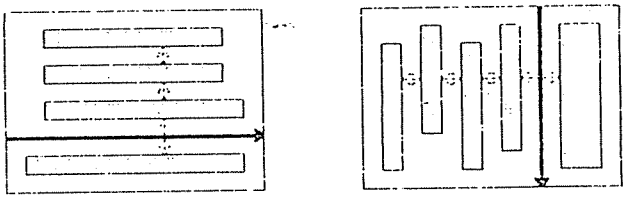

(c) OP3: generation of 2 partitions in $\alpha > \beta$



(d) OP4: generation of several partitions by spaces



(e) OP5: generation of 2 partitions by first space



(f) OP6: generation of 2 partitions by last space

Fig.5 Segmentation operator

for the leaf directions until the tree node becomes terminal. Our segmentation strategy is heuristic. The structure description must be specified so as to be able to interpret various kinds of cataloging cards effectively. We show an example in Fig.6. This is a practical application of the tree structure in Fig.3 for the cataloging card in Fig.2.

## 5. IDENTIFICATION OF DATA ITEMS

The structure description is not always so powerful to distinguish each data item from the cataloging cards completely, because the representation ability is based only on the topological/geometric relationships among partitions. It is necessary to investigate another method which extracts data items from each partition under the constructive relationships of data items. For example, the terminal nodes "T"s, which are composed of several data items yet, can not be divided explicitly by the topological/geometric information. These data items in 1 or more lines can be separated by the space and characters such as "著", "訳" and so on. We call these discriminative symbols the keywords.

From such a discussion point of view, we propose another method to describe constructive relationships among data items. Our method is basically an application of the derivation rules. The derivation rules are composed simply of the form "if --then -- else --". The derivation rules for the bibliography section are shown in Fig.7. For example, some rules in the entry are interpreted as follows:

(ex.1)
( 版 ), edition
=> IF character is "版" THEN data item is "edition";
go to rule-2;
ELSE go to rule-2;

(ex.2)
( 著/N ; 上 ; 中 ; 下 ; 巻 ; 集, E%J ), author-statement, 8
=> IF character in first pointed data item is "著"
and character in second pointed data item is
number(N), "上", "中", "下", "巻" or "集"
THEN data item is "author-statement";
IF data attribute is English or Japanese character
THEN go to rule-4;
ELSE call normalization-routine;
go to rule-4;
ELSE go to rule-8;

These rules are applied to the partition from the last data item area. This is because the keywords are attached to the last position of each string, in many cases. We explain the matching mechanism by using the sample card in Fig.2. The matching control process is illustrated in Fig.8. As shown in Fig.8(a) conceptually, the rule actions indicate 3 ways:

(1) In the complete concidence of 2 surrounded keywords( "訳" and "著" ), the data item is determined as the translator completely;
(2) In the concidence of only one keyword( "訳" ), the next term must be examined to search the keyword "著";
(3) In the disagreement of the first keyword( "訳" ), another rule must be selected to determine the data item.

In this phase, we must detect the particular keywords such as " "(space), "著", "訳" and so on in order to separate strings in each data item. So, it is necessary to pick up the characters by image processing routines in advance. However, it is no problem even if the routine for character pattern sizes did not necessarily work completely. The derivation rules also point out to modify the predetermined pattern sizes when the results are not consistent with the terms in the derivation rules. We show it in Fig.8(b). In this case, all character pattern sizes are not the same though the rules tell to the procedure that the scanned characters must be Japanese attributes. Therefore, the normalization routine must be called to modify the wrong character pattern sizes with the attribute information "K%J" attended to the rule. Here, "K" indicates the kana-character, "E" does the English-character and "J" does the kanji-character.

Additionally, this identification procedure will be adaptable to the partitions of data items, even if the cataloging cards are dirty with noises, and/or the printed characters are indistinct and blurred. For example, we show normalized character pattern sizes in Fig.9. This result includes noises yet. As these noise areas are inconsistent with every derivation rule, the identification procedure can neglect them easily. Even if one rule matched, another rule in the next step finds that the interpretation is contradictory. Also, some characters in this cataloging card are very blurred. However, in some cases we can recognize them on the basis of the context information. For example, a character "書" in Fig.9( denoted by the circle ) will be extracted correctly with the knowledge that this data item is the subject-heading and the character is one of several restrained characters such as "訳", "著", "巻", "書" and so on. This is because the meaning of each data item can be heuristically determined by the segmentation procedure and by the identification procedure.

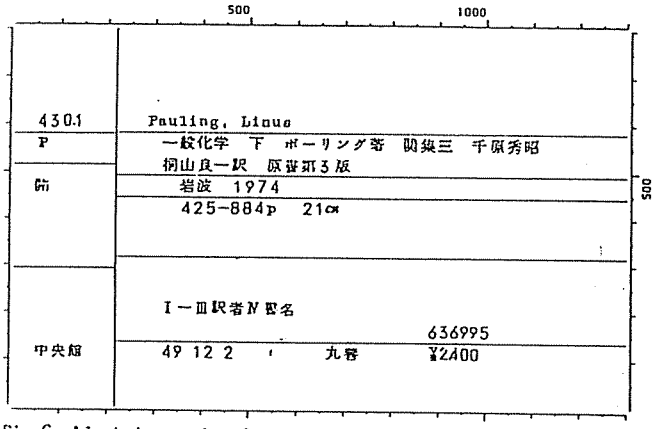



Fig.6 Adapted example of structure description

1) ALL,author
2) END
(a) heading

1) (版),edition
2) (訳/著,K%J),translator
3) (著/N;上;中;下;巻;集,E%J),author-statement,8
4) (N(1,2);上;中;下;上巻;下巻;第N(1,2)巻;第N(1,2)集) ,volume
5) (-/-,K%E%J),subtitle
6) REST,title
7) END
8) (著/-,E%J),author-statement,0
9) (-/-,K%E%J),subtitle
10) REST,title
11) END
(b) entry

1) (N4;明治N(1,2);大正N(1,2);昭和N(1,2)) ,date-of publication
2) REST-2,name-of-publisher,5
3) REST,place-of-publication
4) END
5) REST,name-of-publisher
6) END
(c) imprint

1) (( / ),K%E%J),series
2) (N(1,2)cm) ,size
3) (N/P,・・・),illustration
4) (N(2,3)P;N3;N2P;N3-N3P) ,pagination
5) END
(d) collation and series

1) ALL,note
2) END
(e) note

Fig.7 Example of derivation rules for bibliography section

## 6. DISCUSSION

Our approach does not use directly the character recognition

technique as well as many traditional approaches, but applies stepwise various kinds of knowledge such as information of the layout structure, information of the topological/geometric relationships among partitions, information of the constructive relationships among data items and so on, concerning cataloging cards. Moreover, the procedures and routines for extraction of data items keep mutual interrelations, exchange their information, and control cooperatively the other procedures/routines. Therefore, the following procedure can detect the contradictory processing results at all, even if each procedure/routine could not decide the properties of objects completely.

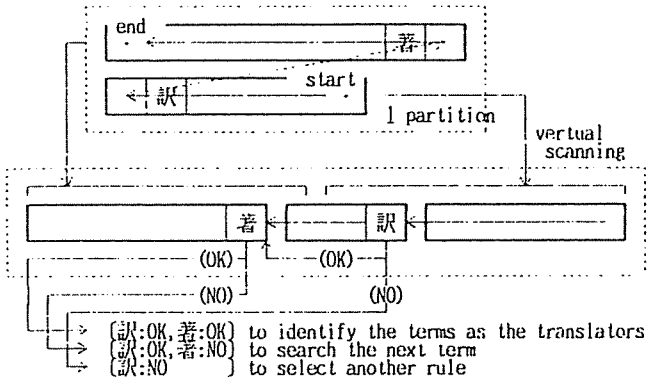
Our approach based on the artificial intelligence technology is more flexible and adaptable than the traditional approaches whose main subject is only the character recognition technique. Namely, the extraction process of data items or characters is performed throughout with various kinds of knowledge. Therefore, even if the local processing module could not recognize objects correctly, the following module can modify easily the wrong result so as to be consistent with its own knowledge. While, the traditional approaches are one-way processing, and can not alter the result in general if the correct object was not extracted.

Our approach has not yet completed. Some issues are imposed on the knowledge representation, the control mechanism, the description ability and so on. However, our method is not always limited to only cataloging cards, but will be adaptable to documents with some layout structures. Of course, we must improve or extend our basic framework so as to make the adaptability successful.

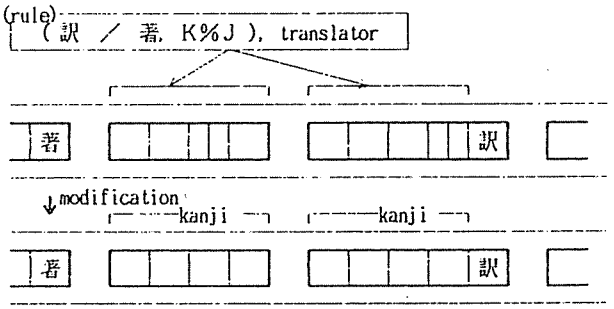
## 7. CONCLUSION

In this paper, we mentioned our experimental approach for extraction and classification of data items from the library cataloging cards. At present, our objective has not yet been attained completely. However, we think that our approach is adaptable to our objective sufficiently because the cooperative control mechanism among procedures/routines is very flexible. For the purpose of the accomplishment of our objective, we must investigate some remained issues attached to our approach: the relationship representation for different knowledge, the cooperative control mechanism among independent processing modules, the specification methods of knowledge, the refinement and reorganization mechanism for conflicted knowledge, and so on. Additionally, we will improve or extend our framework so as to be adaptable to more general processing objects such as documents with layout structures.

Acknowledgements -- The authors are grateful to Prof.T.TSUGE, Mr.A. BANDO, Mr.J.SAKAGUCHI and the staffs of the University Library of Nagoya University for their much assistance and cooperation, and also wish to thank Prof.T.FUKUMURA of Chukyo University and Prof.J. TORIWAKI of Nagoya University for their perspicacious remarks.




(a) matching process of data item




(b) modification of character pattern sizes by attributes

Fig.8 Matching mechanism by constructive derivation rules

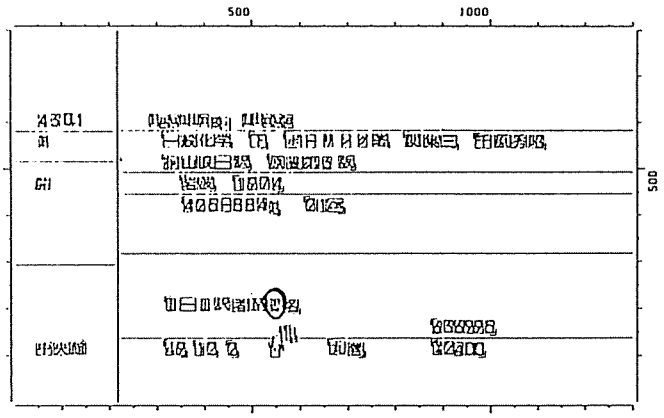



Fig.9 Extraction for dirty cards and/or indistinct characters

References

1) H.HASE, M.YONEDA, M.SAKAI & J.YOSHIDA: "On an Automatic Item Classification for Book Cards", EIC trans., Vol.J70-D, No.8, pp. 1579-1588(1987). (in Japanese )

2) H.HASE, M.YONEDA, M.MATSUDA & M.SAKAI: "Interactive Item Classification System for Book Cards", Tech. Report of EIC Japan PRU88-25, pp.33-39(1988). (in Japanese )

3) E.CHARMICK & D.MCDERMOTT: "Introduction to Artificial Intelligence", P.701, Addison-Wesley Publishing Company(1984).

4) K.YOSHIMOTO, K.ABE & S.KOMORI: "A Printed Characters Recognition Algorithm Using the Peripheral Structures of Characters", Tech. Report of EIC Japan PRU88-13, pp.61-68(1988). (in Japanese )

5) B.S.WYNAR: "Introduction to Cataloging and Classification, 5th edition", Library Science Text Series, P.426, Libraries and Limited Inc.(1976).

6) C.Y.SUEN: "Character Recognition by Computer and Applications", Handbook of Pattern Recognition and Image Processing( ed.by T.Y. YONG & K.S.FU ), P.705, Academic Press Inc.(1986).