Document-Form Identification Using Constellation Matching of Keywords Abstracted by Character Recognition

Hiroshi Sako[1], Naohiro Furukawa[1], Masakazu Fujio[1], and Shigeru Watanabe[2]

[1] Central Research Laboratory, Hitachi, Ltd.
1-280 Higashi-Koigakubo, Kokubunji, Tokyo 185-8601, JAPAN
[2] Mechatronics Systems Division, Hitachi, Ltd.
1 Ikegami, Haruoka, Owariasahi, Aichi 488-8501, JAPAN
sakou@crl.hitachi.co.jp

Abstract. A document-form identification method based on constellation matching of targets is proposed. Mathematical analysis shows that the method achieves a high identification rate by preparing plural targets. The method consists of two parts: (i) extraction of targets such as important keywords in a document by template matching between recogised characters and word strings in a keyword dictionary, and (ii) analysis of the positional or semantic relationship between the targets by point-pattern matching between these targets and word location information in the keyword dictionary. All characters in the document are recognised by means of a conventional character-recognition method. An automatic keyword-determination method, which is necessary for making a keyword dictionary beforehand, is also proposed. This method selects the most suitable keywords from a general word dictionary by measuring the uniqueness of keywords and the stability of their recognition. Experiments using 671 sample documents with 107 different forms in total confirmed that (i) the keyword-determination method can determine sets of keywords automatically in 92.5% of 107 different forms and (ii) that the form-identification method can correctly identify 97.1% of 671 document samples at a rejection rate 2.9%.

1 Introduction

Though e-X [X: banking, cash, commerce, etc.] is very popular now, people are still using paper application forms and banknotes at windows or counters of banks, city halls, etc. This is because direct manipulation of such forms is very easy for most people. Senior people, especially, who are not familiar with personal computers, are not easy beneficiaries of e-X. Therefore, a hybrid system that can accept both electronic applications and paper applications would be most convenient. Such a system would apply document analysis technology to interpret the content information described by characters on a paper application into codes. The codes are then stored together with corresponding information from the electronic application.

This study focuses mainly on technology for document identification. For example, such technology could be applied for discriminating the kind of application

D. Lopresti, J. Hu, and R. Kashi (Eds.): DAS 2002, LNCS 2423, pp. 261–271, 2002.

slip handled at a counter at a bank, and it could work on a system composed of a scanner and a personal computer. Moreover, in Japan, some automated teller machines (ATMs) need a function for remitting money (e.g. to pay the public utility charges) as well as a withdrawal function. These machines must recognise the kind of remittance form submitted. By identifying the form and using the knowledge of its frame structure, the machines can read the information, such as the amount of money to be sent, the time of payment, the remitter and the remittee, which is written at specific places of the form.

Conventional methods for identifying an application form are mainly based on the dimensions of the form and the characteristics of the frame structure. The characteristics [1]-[4] include the number of frame strokes and the relative positional or absolute spatial relationship of the frames. The relationship is sometimes expressed as a tree [2]. One of the advantages of these methods is that the processing speed is very fast because the image processing (such as matching with dictionary frame templates) is very simple. However, these methods might not be applicable to forms with a similar frame structure because their essential principle is based on the structure difference. Other related studies include layout analysis methods [5][6] based on the location of fields and logos. The relationship between fields is sometimes expressed as graphs [5]. While these studies are mainly concerned with layout analysis for reading items in known forms, this paper focuses on the identification of the unknown forms.

To discriminate similar forms, a method based on constellation matching of keywords is proposed. The keywords are abstracted by character recognition and template (character string) matching. One of the advantages of this method is that the system interprets all character images into their category codes, which can be used not only for string matching to extract the keywords but also for initial reading of the written items at the same time. The items include the amount of money and the time of payment on the remittance sheet, for example. In this paper, firstly, the concept of constellation matching is explained. Secondly, a form identification method based on constellation matching of keywords and a method for selecting unique and valuable keywords to recognise each form are explained in Chapter 3. Finally, experimental results on 671 real document-forms are given in Chapter 4.

2 Constellation Matching

2.1 Structure

Template pattern matching and point pattern matching are simple but very useful techniques for identifying an input object. The techniques have therefore been implemented into real-time image-processing system in industrial and OCR products in the form of hardware in the 70's and software in the 90's. Generally speaking, the combination of these techniques can achieve good performance with high recognition rate and low error rate, because the template pattern matching can check the existence of characteristic patterns in the image or essential words (keywords) in the document, and the point pattern matching can analyse the geometrical relationship between these patterns or the semantic relationship between these words. In other words, the combination makes it possible to recognise an input image or a document by abstracting essential patterns or important words. This combined method [7] is referred as constellation matching hereafter because targets such as characteristic patterns or essential words are like fixed stars in a constellation.

Constellation matching is composed of two parts as shown in Fig. 1. (1) Template pattern matching (character string matching) extracts important targets such as keywords of the document. (2) Point pattern matching analyses positional or semantic relationship between the targets.

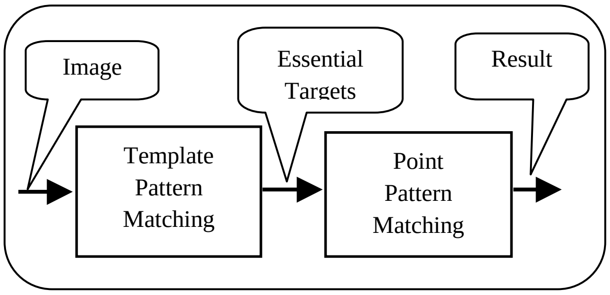


Fig. 1. Constellation matching

2.2 Mathematical Analysis

Generally speaking, constellation matching can statistically achieve good performance with high recognition rate and low error rate, because the number of targets is usually plural and it is possible to make a decision by detected targets even if some of targets are lost. To estimate recognition rate approximately, it is assumed that detection rate q of an individual target is identical for all targets in the object. Thus, the final recognition rate Q is represented as the probability of detecting more than a threshold number of detected targets M out of a number of the expected targets N. Therefore, by using the probability function of a binomial distribution, probability Q is expressed as

$$Q = \Sigma_{i=M,N}\, {}_N\mathrm{C}_i * q^i * (1.0 - q)^{N-i} . \qquad \textbf{(1)}$$

Fig. 2 shows the relationship between Q and N when $M = N / 2$. It can be seen that Q is larger than individual detection rate q since probability Q is the summation of the probabilities for every combination of detected CPs (Characteristic Patterns, i.e., targets or keywords) whose number exceeds M. Thus, constellation matching has the advantage that it has higher recognition rate than conventional methods using a single target.

3 Document-Form Identification

3.1 Structure

As shown in Fig. 3, the document-form identification system is composed of three main parts: keyword determination, keyword dictionary and document-form identi-

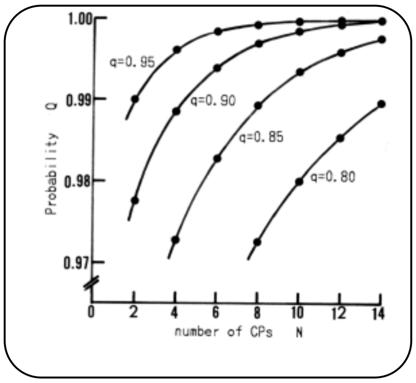


Fig. 2. Relationship between *Q* and *N*

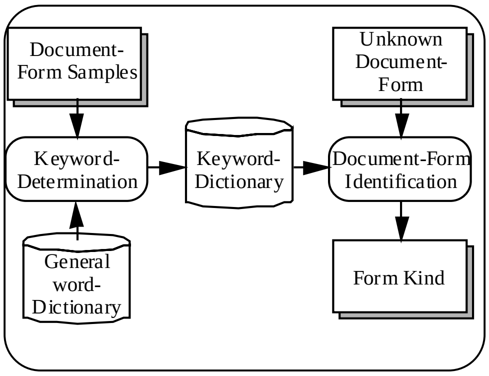


Fig. 3. Structure of document-form identification

fication. In advance, the keyword determination defines the plural keywords in each individual document-form and stores them in the keyword dictionary. For each document-form, the set of keywords must be unique because the difference between sets of keywords is essential for discriminating the form.

A binary image of the form to be identified is put into the form identification part and the form structure is analysed. Then, the binary connected components are detected as character candidates, and they are put into a character classifier. The detected character category codes are matched with all sets of keywords in the keyword dictionary, and the kind of the document-form with the biggest number of matched keywords is regarded as the kind of the input form.

3.2 Form Identification

The identification of the document-form is one of the new applications of constellation matching [9]. As mentioned before, this task is very important in document-reading systems to realise automation equipments such as special ATMs, because the items to be read might depend on the kind of document. Therefore, to complete the document-reading, the identification of a document-form is indispensable. The developed document identification system is composed of two steps as shown in Fig. 4. (i) Character recognition and string matching to detect keywords (i.e., template pattern matching). (ii) Analysis of locational relationship between the keywords to identify the document-form (i.e., point pattern matching).

In the step (i), the input document image is analysed and decomposed [10][11] into several character-line images surrounded by line frames as shown in Fig. 5. The character-line image in each field is separated into connected components as character candidates, which are all examined by a character classifier and are transformed into the corresponding character category codes. However, in the separation process, it is very difficult to separate the character-line image into characters one by one correctly because the components of a Japanese Kanji character are usually Kanji characters

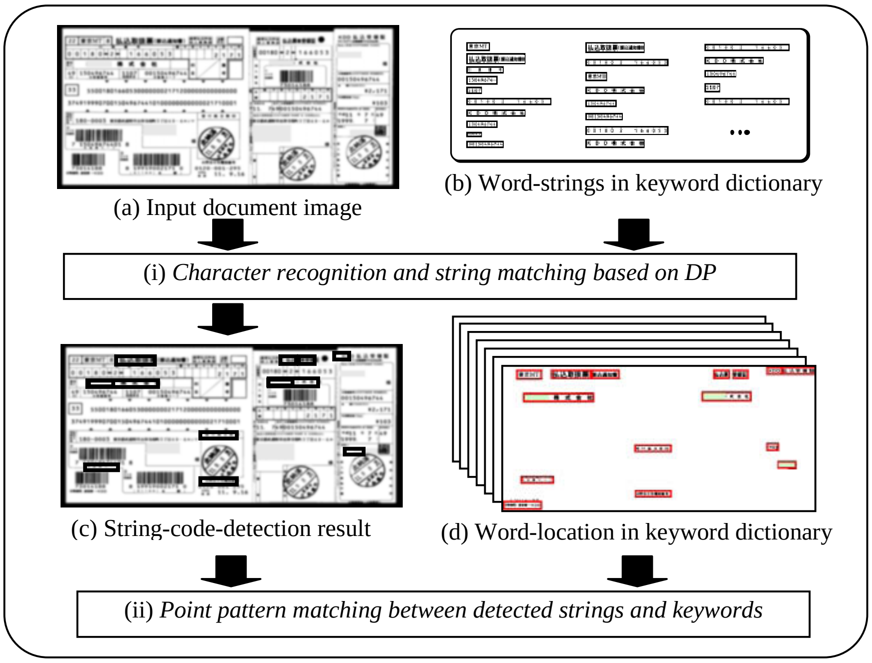

(a) Input document image
(b) Word-strings in keyword dictionary
(i) *Character recognition and string matching based on DP*
(c) String-code-detection result
(d) Word-location in keyword dictionary
(ii) *Point pattern matching between detected strings and keywords*

Fig. 4. Document-form identification

themselves. To cope with this difficulty, many ways of separation are executed so as to be one correct separation within them (so-called "over-segmentation" [8]). As shown in Fig. 6(i-a), one way of separation is expressed by one path of the network. In the process, joined characters, if any, are also separated by the rule-based method [12] using shape analysis of the joined parts of the strokes. Each character-string image in the document-form is mapped onto the network. In the character-recognition process, all connected components in the network are classified into character category codes. The classification is executed under the assumption that the category of an examined character should exist within categories of characters that express dictionary keywords. This assumption can effectively reduce both processing time and the number of misclassified cases. Since the first candidate from the classifier might not be correct, plural candidates are kept in a table (Fig. 6(i-b)). This table is made for every path expressed in the network.

To detect keywords, the detected character category codes (detected strings) are matched with each keyword string in a keyword dictionary prepared as *a priori* knowledge. The keyword-dictionary in Fig. 4(b) is made of popular pre-printed words and proper nouns, which are automatically determined and gathered from sample document-forms in advance. The string matching method is based on a special type of DP (Dynamic Programming) [9], [13]-[15], which allows fluctuations

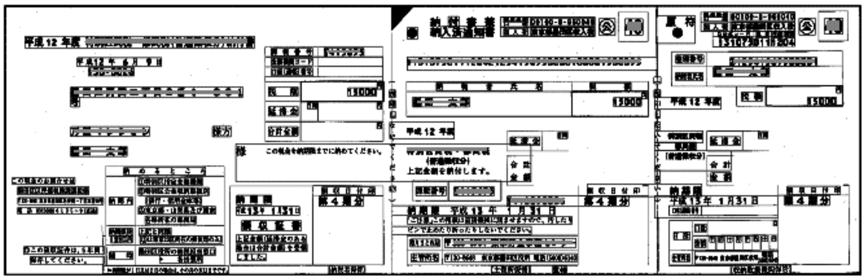
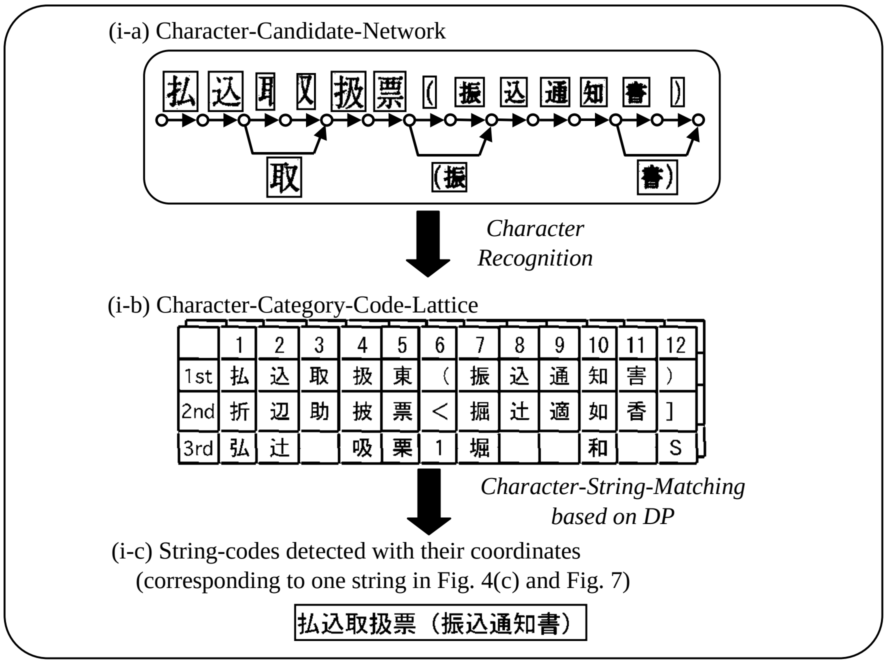


Fig. 5. Detection of character-line images (surrounded by rectangles)



(i-a) Character-Candidate-Network
払 込 耳 又 扱 票 (振 込 通 知 書)
取
(振
書)
Character Recognition
(i-b) Character-Category-Code-Lattice
Character-String-Matching based on DP
(i-c) String-codes detected with their coordinates (corresponding to one string in Fig. 4(c) and Fig. 7)

	1	2	3	4	5	6	7	8	9	10	11	12
1st	払	込	取	扱	東	（	振	込	通	知	害	）
2nd	折	辺	助	披	票	＜	掘	辻	適	如	香	］
3rd	弘	辻		吸	栗	1	堀			和		S

払込取扱票（振込通知書）

Fig. 6. Character recognition and string matching based on DP

such as insertion, deletion and substitution of one character, because the character segmentation and recognition are not always perfect. Penalty *P* of the detected keyword is defined by taking account of both the degree of fluctuations and the rank of candidates of the character recognition. The matching score of keyword S_{kw} is calculated using penalty *P* as follows:

$$S_{kw} = 1 - (P / L) , \quad (2)$$

where *L* is the number of characters of the corresponding dictionary keyword in one kind of form. Fig. 7 shows a magnified image including detected strings, and the image is a part of the detection result in Fig. 4(c).

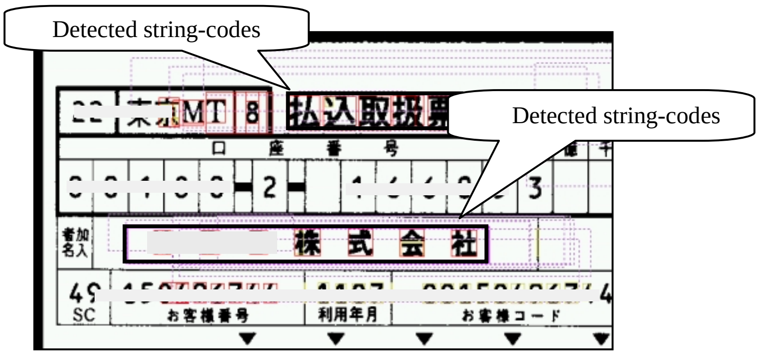


Fig. 7. Result of string code detection

In the step (ii), the detected strings in Fig. 4(c) are estimated by referring to each set of keywords and their locational information as shown in the keyword-dictionary (d). The estimation is based on the average matching scores of the keywords that are detected just at the same positions of the dictionary keywords. The average matching score S^{j}_{c} in the dictionary form j is calculated by

$$S^{j}_{c} = (1 / D_j)\ \Sigma_{i=1, Dj}\ S^{ij}_{kw}\ , \tag{3}$$

where D_j is the total number of detected keywords located correctly on the keyword-dictionary of the form j, and S^{ij}_{kw} indicates the matching score of the ith keyword in the keyword-dictionary of the form j. Finally, the kind of the dictionary form having the maximum S^{j}_{c} determines the kind of form.

3.3 Keyword Determination

A keyword dictionary, which must be prepared in advance, is a kind of abstract of content such as pre-printed titles of fields and pre-printed words in notes and instructions on each kind of document. Each abstract is expressed by a combination of such unique keywords or popular words that have a unique positional relationship. In this section, the method for determining these keywords is explained.

The requirement is that the method must determine the keywords: (1) which can identify each document-form and (2) which can be easily recognised by the character classifier when the form identification is being executed. The approach [16] to satisfy requirements (1) and (2) is to measure the degree of uniqueness of each keyword and to measure the degree of recognition stability of the keyword.

In the measurement of the uniqueness, a similar procedure of the form identification (Section 3.2), which is composed of the form structure analysis and the recognition of the character string in each field, is executed. The difference is that this measurement uses a general word dictionary, which stores several thousand general words, in order to detect a unique set of keywords for each document-form. Samples with different format are collected at first. The degree of uniqueness U of a word is then defined by

$$U = 1 / N_u\ , \tag{4}$$

where N_u is the number of forms at the same position at which the very word is located. A word with $U = 1$ is very unique to all forms to be discriminated; therefore, it must become a keyword candidate for a particular form.

To measure the stability of keyword detection, physically different but the same kind of samples are collected. The number of forms, N_s, where the keyword candidate can be recognised at proper position is counted. The stability S is defined by

$$S = N_s \,/\, N_t\,, \tag{5}$$

where N_t is the total number of examined samples with the same format. This definition is very useful to determine the stable keywords because the recognition rate of the keyword depends on the document image. The image usually changes according to scanning and printing conditions even if the sample has the same format. The keyword candidate, whose uniqueness is 1 and stability S is higher than a certain value, is selected as one of keywords of the form, and they are stored to the keyword-dictionary in advance. Note that each document-form can have plural keywords in the dictionary. The desirable number of keywords is estimated from the final recognition rate Q formulated in Equ. 1.

A serious problem encountered in the real world is that requirements (1) and (2) must also be satisfied under the circumstance that we cannot collect many samples but only single sample of a document-form because hundreds of new document-forms are produced in public and private organisations every year, and it's impossible to ask them to collect many samples in every different kind of form. To solve this problem, the perturbation method is used to increase the number of samples virtually. The perturbation can be realised by binarising the grey image of the document-form at several levels around an optimum threshold level and by slightly shifting and skewing the image of the form intentionally. These kinds of images of the form are added and used in the keyword determination.

4 Experiments Using Real Document Samples

To evaluate the effectiveness of the perturbation method and to measure the correct identification rate, 671 document samples were prepared in total and they include 107 kinds of document-forms. Also, about 4500 words were prepared in the general word dictionary to evaluate the accuracy of keyword determination.

4.1 Evaluation of Keyword Determination

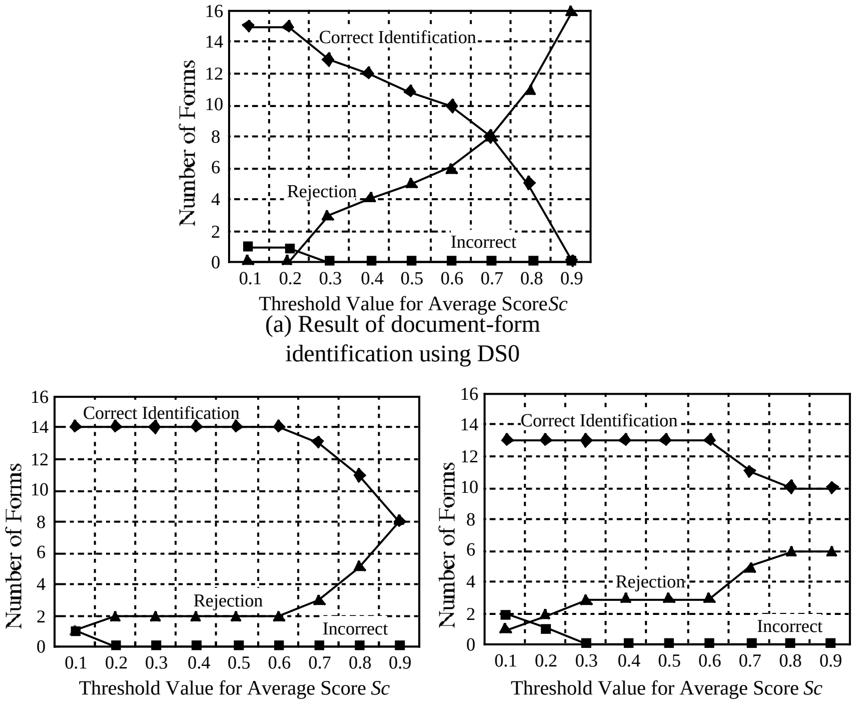
To measure the effectiveness of the perturbation method, the following four data subsets were prepared from original samples. Each dataset includes 16 kinds of forms. The datasets are listed below:

DS0: one sample for each kind of form (1x16 samples in total),
DS1: nine randomly selected samples for each kind of form (9x16 samples),
DS2: nine samples generated by perturbing the sample for each kind of form in dataset DS0 (9x16 samples),
DS3: another one randomly selected sample for each kind of form (1x16 samples).

Figs 8(a), (b) and (c) show the results of the document-form identification for the samples in DS3 by using sets of the keywords that are determined from datasets DS0, DS1 and DS2, respectively. The figures show the number of forms that can be identified correctly as well as the number of rejected forms. They are plotted when the threshold value for the average score S^j_c changes from 0.1 to 0.9. Though the number of samples is small, it is possible to conclude the followings.

(1) Increasing the number of samples for each kind of form increases the identification rate and decreases the rejection rate and error rate even when the threshold for the average score S^j_c is high [Figs. 8(b) vs. 8(a)].
(2) The comparison of the identification rate using the keywords determined by DS2 with that by DS0 [Figs. 8(c) vs. 8(a)] shows that the perturbation method improves the identification rate and its stability significantly, and the identification rate using the keywords determined by DS2 is comparable with that by DS1 [Figs. 8(c) vs. 8(b)].

The keyword determination rate is 92.5%, which means that the sets of keywords can be determined in 99 kinds out of 107 kinds of document-forms. The determination failure is mainly caused by the failure to detect the character string images by the structure analysis. The keywords in these failure forms are prepared manually to complete the evaluation of the document-form identification in the next section.



(a) Result of document-form identification using DS0

(b) Result of document-form identification using DS1

(c) Result of document-form identification using DS2

Fig. 8. Effect of perturbation method

4.2 Evaluation of Document-Form Identification

An evaluation of document-form identification based on 671 document samples showed 97.1% correct identification with 2.9% rejection and 0% error rates. The main reason for the rejection is the detection failure of line frames, which makes it difficult to detect the character-line images correctly. Apart from this rejection, the method can realise reliable document identification without any errors in this sample size.

5 Concluding Remarks

A document-form identification method (based on the constellation matching of keywords) and a method for determining the keywords (based on the uniqueness of keywords and the stability to their recognition) are proposed. To increase the identification rate in the case that plural samples for each document-form cannot be collected, a perturbation method is also proposed. Experiments using 671 sample documents with 107 different forms in total confirmed that (i) the keyword-determination method can determine sets of keywords in 92.5% of 107 different forms and (ii) the form-identification method can correctly identify 97.1% of 671 document samples with the rejection rate 2.9%.

These results show the effectiveness of document-form identification method. This method can be applied to natural images and artificial documents and can be very useful in position measurement [17] as well as object (document-form) identification. The currently available large-memory capacity and high-speed CPUs will enable the method to be applied widely. That is, the large memory capacity makes it possible to store massive volumes of cases (keywords) as *a priori* knowledge, and a high-speed CPU can process a huge number of simple matching instructions in a short time. It is thus considered that the memory-based approach based on constellation matching can be applied to many applications in the real world.

Acknowledgements. The authors wish to thank Hisao Ogata and his group in the Hitachi Mechatronics Systems Division for their support in the implementation and evaluation of this system. Special thanks go to Dr. H. Fujisawa of the Hitachi Central Research Lab. for his valuable suggestions and encouragement.

References

1. M. Asano and S. Shimotsuji, "Form Document Identification Using Cell Structures," *Technical Report of IEICE*, PRU95-61, pp. 67–72, 1995 (in Japanese).
2. Q. Luo, T. Watanabe and N. Sugie, "Structure Recognition of Various Kinds of Table-Form Documents," *Trans. of IEICE*, Vol, J76-D-II, No. 10, pp. 2165–2176, 1993 (in Japanese).
3. M. Ishida and T. Watanabe, "An Approach to Recover Recognition Failure in Understanding Table-form Documents," *Technical Report of IEICE*, PRU94-35, pp. 65–72, 1994 (in Japanese).
4. T. Watanabe and T. Fukumura, "A Framework for Validating Recognized Results In Understanding Table-form Document Images," *Proc. of ICDAR '95*, pp. 536–539, 1995.

5. F. Cesarini, M. Gori, S. Marinai and G. Soda, "INFORMys: A Flexible Invoice-Like Form-Reader System," *IEEE Trans. on PAMI*, Vol.20, No. 7, pp. 730–745, 1998.
6. S.L. Lam and S. N. Srihari, "Multi-Domain Document Layout Understanding," *Proc. of ICDAR '93*, pp. 497–501, 1993.
7. H. Sako, M. Fujio and N. Furukawa, "The Constellation Matching and Its Application," *Proc. of ICIP 2001*, pp. 790–793, 2001.
8. H. Fujisawa, Y. Nakano and K. Kurino, "Segmentation Methods for Character Recognition: From Segmentation to Document Structure Analysis," *Proc. of the IEEE*, Vol. 80, No. 7, pp. 1079–1092, 1992.
9. N. Furukawa, A, Imaizumi, M. Fujio and H. Sako, "Document Form Identification Using Constellation Matching," *Technical Report of IEICE*, PRMU2001-125, pp. 85–92, 2001. (in Japanese)
10. H. Shinjo, K. Nakashima, M. Koga, K. Marukawa, Y. Shima and E. Hadano, "A Method for Connecting Disappeared Junction Patterns on Frame Lines in Form Documents," *Proc. of ICDAR '97*, pp. 667–670, 1997.
11. H. Shinjo, E. Hadano, K. Marukawa, Y. Shima and H. Sako, "A Recursive Analysis for Form Cell Recognition," *Proc. of ICDAR 2001*, pp. 694–698, 2001.
12. H. Ikeda, Y. Ogawa, M. Koga, H. Nishimura, H. Sako and H. Fujisawa, "A Recognition Method for Touching Japanese Handwritten Characters," *Proc. of ICDAR '99*, pp. 641–644, 1999.
13. F. Kimura, M. Shridhar and Z. Chen, "Improvements of a lexicon directed algorithm for recognition of unconstrained handwritten words," *Proc. of ICDAR '93*, pp. 18-22, 1993.
14. F. Kimura, S. Tsuruoka, Y. Miyake and M. Shridhar, "A lexicon directed algorithm for recognition of unconstrained handwritten words," *IEICE Trans. Info. & Syst*., Vol. E77-D, No. 7, pp. 785–793, 1994. (in Japanese)
15. H. Bunke, "A fast algorithm for finding the nearest neighbor of a word in a dictionary," *Report of Institut fur Informatik und Angewandte Mathematik, Universitat Bern*, 1993.
16. M. Fujio, N. Furukawa, S. Watanabe and H. Sako, "Automatic Generation of Keyword Dictionary for Efficient Document Form Identification," *Technical Report of IEICE*, PRMU2001-126, pp. 93–98, 2001 (in Japanese).
17. H. Sakou, T. Miyatake, S. Kashioka and M. Ejiri, "A Position Recognition Algorithm for Semiconductor Alignment Based on Structural Pattern Matching," *IEEE Trans. Acoustic, Speech, Signal Processing*, ASSP-37, pp. 2148–2157, Dec. 1989.