IMAGE SEGMENTATION OF HANDWRITTEN DATES ON BANK CHEQUES

Dr. Manish M. Kayasth

ABSTRACT

In this paper Author has described different issues of handwritten date segmentation task briefly along with its techniques. The paper describes a system develop to segments handwritten date information specifically written on bank checks. A system uses a newly adopted segmentation-based strategy. In order to achieve high performance in terms of efficiency and reliability, a knowledge-based module is proposed in order to segments the date. The interaction between the segmentation and recognition stages is properly established by using Segmented-Date generation and evaluation modules. The paper concludes with the current status of the effort made in segmentation of handwritten dates and future enhancement in the same direction.

Keywords: OCR, Segmentation, Pattern-based grammar.

1. INTRODUCTION

Research on OCR (Optical Character Recognition) began in the 1950s, and it is one of the oldest research areas in the field of pattern recognition [6]. Nowadays, most commercial efforts in OCR concentrate for reliably processing cleanly machine-printed text documents with simple layouts i.e. forms with pre-set character boxes, and some successful systems have also been developed to recognize handwritten texts, particularly, isolated hand-printed characters and words. However, the analysis of documents with complex layouts, recognition of degraded machine-printed texts, and the recognition of unconstrained handwritten texts demand further improvements through research. The development of reliable classifiers requires considerable effort, and is still not an entirely solved problem [1, 6].

The current generation of OCR systems can be characterized as a pipeline composed of Preprocessing, Segmentation, Classification, and Identification stages. In traditional handwriting OCR systems, recognition is performed at the character level, using the output of an independent segmentation step. The segmentation stage has to make decisions about the location of segment boundaries early on, which is helpful for later stages of character recognition task [1]. The majority of the existing segmentation algorithms make use of a set of rules that has been extracted from empirically observations of the human writing. Thus, despite the fact that results may often seem satisfactory, there is no guarantee that the whole or optimal rule set has been created [6]. Segmentation is posed as a graph cut problem that incorporates the apriori

information from script structure. It aims to partition a document image into various homogeneous regions such as text blocks, image blocks, lines, words etc. [4].

The ability to recognize the handwritten date information on bank checks is very important and one of the very challenging topics in OCR. It is essential in application environments where checks cannot be processed prior to the dates shown and verified. At the same time in case of similar application area, date information also appears on many other kinds of forms. Therefore, there is a great demand to develop some reliable automatic data processing system.

There are high degree of variability in case of handwritings, there has been no published work on the said topic until the work on the date fields of machine-printed checks was reported in 1996 [3]. In 2001, there was some work noted for a date processing system for recognizing handwritten date images on Brazilian checks system [7]. A segmentation-free method was used in this system, i.e. an HMM (Hidden Markov Model) based approach was developed to perform segmentation in combination with the recognition process [7]. The date processing is to be the most difficult target in check processing, given that it has the worst segmentation and recognition performance [3].

The system addressed in this paper is the only work on processing of date zones written in cursive text on bank checks. The research aims at developing an automatic recognition system mainly highlighted around segmentation task for unconstrained handwritten dates. The word-level segmentation is addressed in this paper. In the proposed system, date images are recognized by a segmentation-based method, that is, a date image is first segmented into day, month, and year, the category of month i.e. alphabetic or numeric is identified, and then an appropriate recognizer is applied for each field. In the following, different challenges are addressed in the development of electronic handwritten date processing, the main modules of the entire system architecture will be discussed then after, and finally the segmentation based approach is used for separating each of the date components together with conclusion and future attempts.

2. CHALLENGES

The main challenge in developing an effective date processing system is from the high degree of variability and uncertainty in the data. In other words, there has been no standard rule or format for writing dates on bank cheques people use their own way of writing dates on cheques as shown in the Figure 1. People

usually write the date zones on check in such free styles that little a priori knowledge and few reliable rules can be applied to define the layout of a date image. For example, the date fields can contain either only numerals or a mixture of alphabetic letters say for month and numerals for day and year, punctuations, suffixes.

The other issue in this regard is automatic cheque recognition system verifies whether date has been written on cheque or not? In other way, the cheque is only significant if it contains date at the proper given area. Further, the automatic date identification is important by the way that only proper date is valid for the clearance of cheque. In other words, the system itself is important that the cheque has not been processed and cleared for the post-dated cheque and for such improper date the system suggest that cheque date and is invalid using appropriate recognizer.

The other unique feature of this electronic processing is cheque seems to be worthless if the date written on the cheque is very old and not within the prescribed time duration. Specifically, the bank declared cheque validity for some period of time in terms of say months and after such duration cheque tends to be null and void. The proposed system therefore is useful to check the cheque date is within the given time span? Therefore the system is used for the cheque validity purpose also.

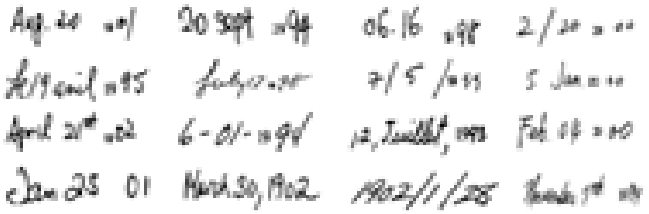

Figure 1: Some Sample handwritten dates on various bank checks.

3. SYSTEM ARCHITECTURE

Since the date image contains fields that may belong to different categories i.e. alphabetic or numeric, it is difficult to process the entire date image at the same time efficiently. Therefore, a segmentation-based strategy is employed in the suggested system. The main procedures in the system consist of segmenting the date image into fields through the detection of the separator or transition between the fields, identifying the nature of each field, and applying an appropriate recognizer for each.

Figure 2 illustrates the basic modules in the suggested date processing system. In addition to the main modules related to the segmentation, a preprocessing is

performed to deal with simple noisy images before the segmentation. Same way, after segmentation, recognition task followed by the post-processing has been designed. In the post-processing stage, in order to further improve the reliability and performance of the system, two-level verification is suggested to accept valid and reliable recognition results, and to reject the others.

The date image segmentation module divides the entire image into three sub-images corresponding to day, month, and year, respectively, and also makes a decision on how month is written so that it can be processed using an appropriate recognizer. Since there is no predefined position for each field, and no uniform or even obvious spacing between the fields, it is difficult to implement the segmentation process with a high success rate by using simple structural features.

In order to improve the performance and efficiency of the system, a knowledge-based segmentation module is used to solve most segmentation cases in the segmentation stage using Pattern based Grammar and the samples of database. An ambiguous case is handled then after by a Segmented-Date generation module. At this stage, for a final decision to be made when more contextual information and syntactic and semantic knowledge are available, this is carrying out by Segmented-Date evaluation.

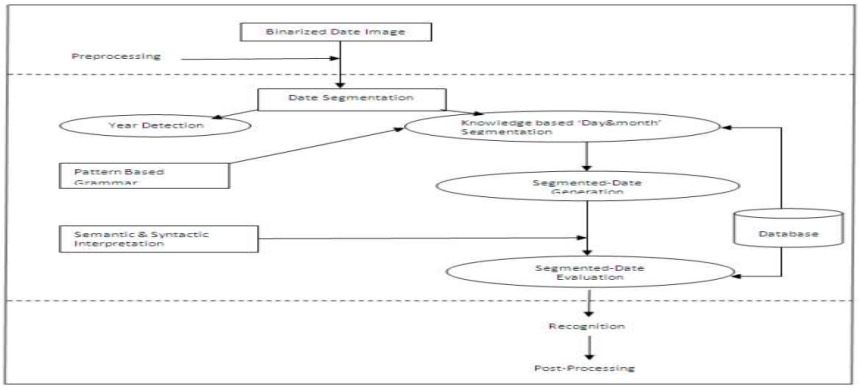

Figure 2: Diagram of date processing system.

4. IMAGE SEGMENTATION OF DATE

The first step of segmentation module is to separate year from 'day&month' based on structural features and the characteristics of the Year field, and then the knowledge-based segmentation module and the Segmented-Date generation and evaluation modules are applied in the 'day&month' segmentation stage.

A. Year Detection

Based on the analyses, the writing styles of date zones in Indian bank checks system can be grouped into two categories, which are defined as standard format and free format. The standard format is used when "2" and "0" or "1" and "9" are printed as isolated numerals on the date zone indicating the century and the free format is adopted when the machine-printed "20" or "19" does not appear on a date zone. Since about 80% of date zones are of the standard format, year with the standard format is first detected in the year detection module. If the detection is not successful, year with the free format is detected.

The detection of year for the standard format is based on the detection of the machine-printed "20" or "19." For detecting year with the free format, there are three possibilities: (i) year is located at one end of the date zone; (ii) year belongs to one of the two patterns i.e. 4 digits 20∗∗ or 19∗∗, or 2 digits ∗∗, where ∗ is a numeral; and (iii) a separator is used by writers to separate a year field from a 'day&month' field.

By considering the said likelihood for writing year with the free style format, use a candidate and then confirmation strategy to detect the year. Year candidates are first detected from the two ends of the date zone, and then one of the candidates is confirmed as the year by using the recognition results from a digit recognizer and by using the assumption that a year field contains either four numerals starting with "20" or "19", or two numerals. Here year candidates are obtained by detecting the separator i.e. punctuation or a big gap between year and 'day&month' fields, and these separator candidates are detected from structural features. In the confirmation stage, the confidence value for a year candidate with the pattern 20∗∗ or 19∗∗ is considered to be higher than that of a year candidate with the pattern ∗∗.

B. Knowledge-Based 'Day&Month' Segmentation

The tasks of knowledge-based segmentation module include (i) detecting the separator between day and month; and (ii) segmenting the 'day&month' field into day and month, and identifying the category of the month.

For the separator detection, the separators can be punctuations, such as slash "/," hyphen "-," comma "," and period "." or big gaps. A candidate and then confirmation strategy is used to detect separators. Separator candidates are first detected by shape and spatial features [2, 8]. Some of the candidates of separators with high confidence values of the features can be confirmed immediately while others should be evaluated by considering more information.

Based on the analyses of the writing styles of different users, some relationship between the type of separator and the writing style of 'day&month' has been found. For example, slashes or hyphens usually appear when both day and month are written in numerals. So some separator candidates are easily confirmed or rejected using the known knowledge based rules about the writing style obtained. Furthermore, these rules can be designed in the training stage based on human knowledge and syntactic constraints, and in general they can be defined in a pattern-based grammar. Pattern here means image pattern, and usually a 'day&month' sub-image can appear in any one of the three patterns: NSN, NSA, and ASN, where S denotes the separator, N denotes a numeric string which may be either day or month or both field, and A denotes an alphabetic string for month field only. The pattern-based grammar used to detect the separator can be expressed as

Input Image Pattern:
If < condition > then < action >
< condition > ⇒ the separator candidate (S) in the image Pattern is P, $P \in \sum$
< action > ⇒ Confidence Adjustment| Add New Feature| Adjust Segmentation Method etc.

Here $\sum$ represents the set of separator types. The pattern based grammar is given in Table 1, where 'day&month' image patterns are given as the entries. To confirm a separator candidate and the condition is that the separator candidate is a "/" or "–" or ..., take the corresponding action.

For this knowledge-based method, two approaches have been developed in the system to determine the writing styles. The first method adopts a distance to numeral measure to represent the likelihood of a subimage in day&month field being numeric, and it is based on feedbacks of a digit recognizer and structural features [2]. A system of combining multiple multilayer perceptron (MLP) networks is the second method to realize this writing style analysis task [5]. The effectiveness of these two methods and their results may be inconclusive or uncertain in some ambiguous segmentation cases, and so the Segmented-Date generation and evaluation modules are introduced.
After the separator detection step, day&month segmentation can be conducted. Based on the separators detected, a set of rules have been developed to segment the day&month field into day and month fields and to determine whether the month field is written in numeric or alphabetic form.

Pattern	Condition	Action
NSN	–	"–" candidate is confirmed.
NSA	/	New features are checked.
ASN	NULL	Separator candidate is not confirmed.

Table 1: Samples of condition-action rules

C. Segmented-Date Generation

In the knowledge-based 'day&month' segmentation module, only the separators with high confidence values and the writing styles with high confidence values to indicate "common styles" would be confirmed. Here "common styles" are determined based on database analyses, including that (i) the gap separator usually occurs at the transition between numeric and alphabetic fields, (ii) the subimages on both sides of slash or hyphen are often numeric, and (iii) a period separator is usually used in ASN pattern. Otherwise, the Segmented-Date generation module is activated. This module produces and places multiple hypotheses in a Segmented-Date list, where each assumption consists of a possible segmentation of 'day&month' field.

D. Segmented-Date Evaluation

Each possible segmentation in the Segmented-Date list includes a separator candidate and segments on both sides of the separator candidate. For each such hypothesis, the Segmented-Date evaluation module estimates its confidence values for the three writing styles or types (NSN, NSA, or ASN). In this system, some semantic and syntactic constraints can be used to improve the performance of this Segmented-Date evaluation module. In addition, "common styles" have been determined based on database analyses. These "common styles" can be used as syntactic constraints.

5. CONCLUSION & FUTURE ATTEMPTS

The author proposes system for automatically segmenting the handwritten date information on bank checks. In the system, the date segmentation is implemented at different levels using knowledge obtained from different sources. Simple segmentation cases can be efficiently solved by using the knowledge-based segmentation module, which makes use of some contextual information provided by writing style analyses. For ambiguous segmentation cases, the Segmented-Date generation and evaluation modules are invoked to make the final decision semantic and syntactic constraints. As a result, promising

performance may be obtained on a various test set data from a real-life standard check database.

The author has concentrates on image segmentation step only and hence the image preprocessing, recognition and post-processing steps are out of scope of this paper. A system uses a newly adopted segmentation-based strategy which can be further useful in recognition stage i.e. cursive word recognition for month is based on a combination of classifiers. Still work is going on and soon it may come up with the entire solution that will fulfill rest of the image processing steps. The said task would be attempted by the researcher as his future endeavors.

REFERENCES:

1. Andr´as Kornai, K.M. Mohiuddin et al., An HMM-Based Legal Amount Field OCR System for Checks, IEEE International Conference on Systems, Man and Cybernetics, Vancouver BC, Vol. 3, PP. 2800-2805, 1995.
2. C. Y. Suen, Q. Xu et al., Automatic recognition of handwritten data on cheques—fact or fiction? Pattern Recognition Letters, PP. 1287–1295, 1999.
3. G. F. Houle, D. B. Aragon et al., A multi-layered corroboration-based check reader, In Proceedings of IAPR Workshop on Document Analysis Systems, USA, PP. 495–546, October 1996.
4. K.S. Sesh Kumar, Anoop M. Namboodiri, et al., Learning Segmentation of Documents with Complex Scripts, Vol. 4338, PP. 749-760, 2006.
5. L. Lam, Q. Xu et al., Differentiation between alphabetic and numeric data using ensembles of neural networks, In Proceedings of the 16th International Conference on Pattern Recognition, Canada, Vol. 4, PP. 40–43, August 2002.
6. M. Maragoudakis, E. Kavallieratou et al., An Effective Stochastic Estimation of Handwritten Character Segmentation Bounds
7. M. Morita, A. El Yacoubi et al., Handwritten month word recognition on Brazilian bank cheques, In Proceedings of the 6th International Conference on Document Analysis and Recognition, USA, PP. 972–976, September 2001.
8. R. Fan, L. Lam et al., Processing of date information on cheques, Progress in Handwriting Recognition, World Scientific, Singapore, PP. 473–479, 1997.

AUTHOR'S PROFILE

Dr. Manish M. Kayasth (Ph. D., MCA) is an HOD at Udhna Academy College Of Computer Application & IT. He is having 9 years academic and 2 years industrial experience. His interest area includes AI, Software Engineering and Web Development.