Automatic Segmentation and Recognition System for Handwritten Dates on Canadian Bank Cheques

Qizhi Xu†, Louisa Lam†‡ and Ching Y. Suen†
†Centre for Pattern Recognition and Machine Intelligence, Concordia University,
Suite GM 606, 1455 de Maisonneuve Boul. West, Montreal, Quebec H3G 1M8, Canada
{qxu, llam, suen}@cenparmi.concordia.ca
‡Department of Mathematics, Hong Kong Institute of Education,
10 Lo Ping Road, Tai Po, Hong Kong

Abstract

This paper describes a system being developed to recognize date information handwritten on Canadian bank cheques. A segmentation based strategy is adopted in this system. In order to achieve high performances in terms of efficiency and reliability, a knowledge-based module is proposed for the date segmentation and a cursive month word recognition module is implemented based on a combination of classifiers. The interaction between the segmentation and recognition stages is properly established by using multi-hypotheses generation and evaluation modules. As a result, promising performance is obtained on a test set from a real-life standard cheque database.

1. Introduction

The ability to recognize the date information handwritten on bank cheques is very important in application environments (e.g. in Canada) where cheques cannot be processed prior to the dates shown. At the same time, date information also appears on many other kinds of forms. Therefore, there is a great demand to develop reliable automatic date processing systems.

The main challenge in developing an effective date processing system stems from the high degree of variability and uncertainty in the data. As shown in Figure 1, people usually write the date zones on cheques 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 (for *Month*) and numerals (for *Day* and *Year*), punctuations, suffixes, and the article "Le" may also appear. (The dates can be written in French or in English in Canada, and a "Le" may be written at the beginning of a French date zone.)

Perhaps because of this high degree of variability, there has been no published work on this topic until the work on the date fields of machine-printed cheques was reported in 1996 [2]. This reference also considered date processing to be the most difficult target in cheque processing, given that it has the worst segmentation and recognition performance. In 2001, a date processing system for recognizing handwritten date images on Brazilian cheques was presented in [4]. A segmentation-free method was used in this system, i.e. an HMM-based approach was developed to perform segmentation in combination with the recognition process.

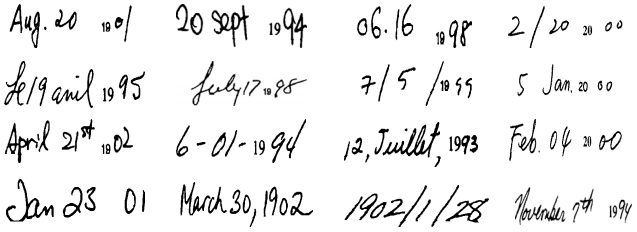

Figure 1. Sample dates handwritten on standard Canadian bank cheques

The system addressed in this paper (which is an extension of a previous work [1]) is the only publication on processing date zones on Canadian bank cheques. In our system, date images are recognized by a segmentation-based method, i.e. a date image is first segmented into *Day*, *Month*, and *Year*, the nature of *Month* (alphabetic or numeric) is identified, and then an appropriate recognizer is applied for each field. In the following, the main modules of the whole system will be discussed, together with some experimental results.

2. System Architecture

The main procedures in our date processing system consist of segmentation and recognition stages, as shown in

Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR 2003)
0-7695-1960-1/03 $17.00 © 2003 IEEE

Figure 2. In addition to these two main stages, a module is designed in the preprocessing stage to deal with simple noisy images, and to detect and process possible appearances of "Le". In the postprocessing stage, a verification module is implemented to accept valid and reliable recognition results, and to reject others.

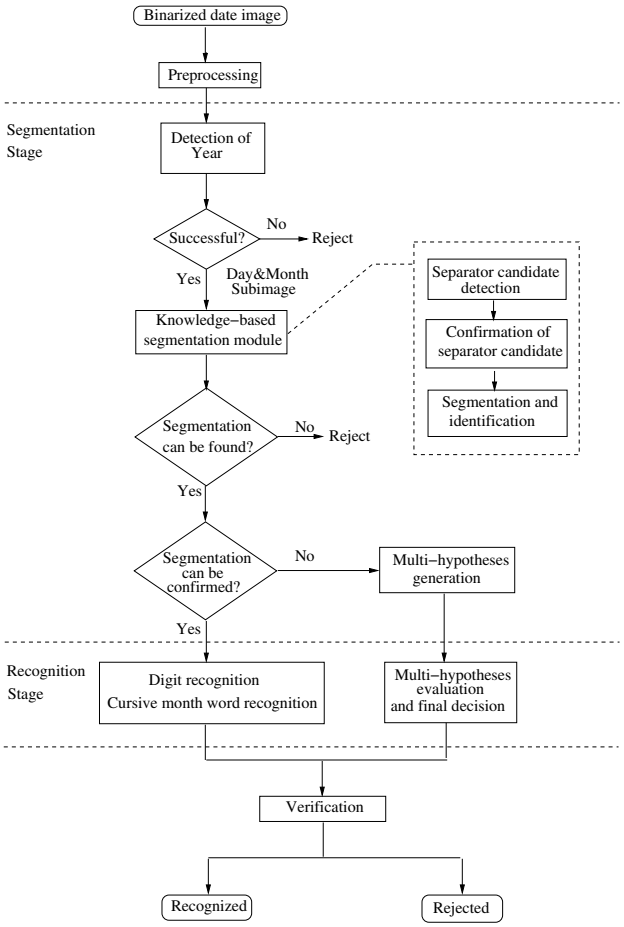

Figure 2. Diagram of date processing system

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. Ambiguous cases are handled by a multi-hypotheses generation module at this stage, for a final decision to be made when more contextual information and syntactic and semantic knowledge are available, i.e., multi-hypotheses evaluation is made in the recognition stage.

3. Date Image Segmentation

3.1. *Year* Detection

As shown in Figure 2, the first step in our date segmentation stage is to separate *Year* from *Day* and *Month* fields (denoted by *Day&Month* subimage). Based on our database analyses (CENPARMI Cheque database and CENPARMI_IRIS Cheque database), the writing styles of date zones on Canadian bank cheque can be grouped into two categories, which are defined as standard format and free format in this paper. The standard format is used when '1' and '9' or '2' and '0' are printed as isolated numerals on the date zone indicating the century, and the free format is adopted when the machine-printed "19" or "20" does not appear on a date zone. Since about 80% of date zones are of the standard format in our databases, *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 "19" or "20". For detecting *Year* with the free format, we first assume: (i) *Year* is located at one end of the date zone; (ii) *Year* belongs to one of the two patterns: $19**$ (or $20**$) and $**$, where $*$ is a numeral; and (iii) a separator is used by writers to separate a *Year* field from a *Day&Month* field. Based on these assumptions, we use a candidate_then_confirmation strategy to detect the *Year* with the free format. *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 4 numerals starting with "19" (or "20") or 2 numerals. Here *Year* candidates are obtained by detecting the separator (a punctuation or a big gap) between *Year* and *Day&Month* fields, and these separator candidates are detected from structural features [1, 6]. In the confirmation stage, the confidence value for a *Year* candidate with the pattern $19**$ (or $20**$) is considered to be higher than that of a *Year* candidate with the pattern $**$.

3.2. Knowledge-Based Module for *Day&Month* Segmentation

The tasks of this 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 nature of the *Month*. For the separator detection, the separators can be punctuations, such as slash '/', hyphen '-', comma ',' and period '.', or big gaps, and a candidate_then_confirmation strategy is used in our system to detect them. Separator candidates are first detected by shape and spatial features [1, 6]. While some of the candidates with high confidence values of the features can be confirmed immediately, others should be evaluated by considering more information.

Based on our database analyses, some relationship between the type of separator and the writing style of *Day&Month* has been found, e.g. slashes or hyphens usu-

ally appear when both *Day* and *Month* are written in numerals. So some separator candidates are easily confirmed or rejected using a set of rules if the knowledge about the writing style can be obtained [6]. Furthermore, these rules can be designed in the training stage based on human knowledge and syntactic constraints, and in general they can be encoded in a pattern based grammar. Here pattern means image pattern, and usually a *Day&Month* subimage can appear in any one of the three patterns: NSN, NSA and ASN, where S denotes the separator, N denotes a numeric string (*Day* or *Month* field) and A denotes an alphabetic string (*Month* field). Some explanations about this pattern based grammar is given in the following table, where *Day&Month* image "Pattern"s are given as the entries. When we try to confirm a separator candidate and the "Condition" is that the separator candidate is a "/" or "−" or..., we can take the corresponding "Action".

Table 1. Examples of condition-action rules

Pattern	Condition	Action
NSN	−	"−" candidate is confirmed
NSA	/	New "long" features are checked
ASA	NULL	Separator candidate is not confirmed
...	...	...

For this knowledge-based method, two approaches have been developed in our system to determine the writing styles. One method is based on a *distance_to_numeral* measure [6], while the other module uses ensembles of neural networks [3]. The effectiveness of these two methods has been proven in the experiments. However, their results may be inconclusive in some ambiguous segmentation cases, where the multi-hypotheses generation and evaluation are introduced.

After the separator detection step, *Day&Month* segmentation and identification 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 [1, 6].

3.3. Multi-hypotheses 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 the gap separator usually occurs at the transition between numeric and alphabetic fields, the subimages on both sides of slash or hyphen are often numeric, and a period separator is usually used in *ASN* pattern. Otherwise the multi-hypotheses generation module is activated. This module produces and places multiple hypotheses in a multi-hypotheses list, where each hypothesis consists of a possible segmentation of *Day&Month* field.

3.4. Multi-hypotheses Evaluation

Each possible segmentation in the multi-hypotheses list includes a separator candidate and segments on both sides of the separator candidate. For each such hypothesis, the multi-hypotheses evaluation module estimates its confidence values for the three writing styles or types (NSN, NSA or ASN) as the following weighted sums:

$$\begin{aligned} Confidence_{Type1} &= w_1 * DigitConfidence_{Left} + \\ &\quad w_1 * DigitConfidence_{Right} + \\ &\quad w_3 * SeparatorConfidence \\ Confidence_{Type2} &= w_1 * DigitConfidence_{Left} + \\ &\quad w_2 * WordConfidence_{Right} + \\ &\quad w_3 * SeparatorConfidence \\ Confidence_{Type3} &= w_2 * WordConfidence_{Left} + \\ &\quad w_1 * DigitConfidence_{Right} + \\ &\quad w_3 * SeparatorConfidence \end{aligned}$$

$DigitConfidence_{Left}$ and $DigitConfidence_{Right}$ are confidence values from a digit recognizer for the left and right sides of the separator candidate respectively. $WordConfidence_{Left}$ and $WordConfidence_{Right}$ are confidence values from a cursive month word recognizer. $SeparatorConfidence$ is the confidence value of the separator candidate, which is derived from the segmentation stage. The weights w_1, w_2, and w_3 are determined in the training stage. For example, $w_1 = 0.8$, and $w_2 = 1$ because the distribution of confidence values from the digit recognizer is different from that of the word recognizer, and these weights are set to make the confidence values of the digit and word recognizers comparable.

For each hypothesis in the list, usually the Type with the maximum $Confidence_{Type}$ value is recorded to be compared with the corresponding information from other hypotheses in the list. In application, some semantic and syntactic constraints can be used to improve the performance of this multi-hypotheses evaluation module. First, based on semantic constraints, if the recognition result from a Type is not a valid date, the corresponding $Confidence_{Type}$ would be reduced by a small value (α) before $Confidence_{Type}$ values are compared in Type selection. In addition, as we discussed above, "common styles" have been determined based on database analyses. These "common styles" can be used as syntactic constraints to modify the Type selection procedure, that is, if the interpretation of the first choice is not a "common style" and the difference between the confidence values of the top two choices is very small, the second choice which has the second largest $Confidence_{Type}$ value should be the final selection if the interpretation of this choice is a "common style" and is a valid date.

4. Date Image Recognition

In the date recognition stage, if one segmentation hypothesis can be confirmed in the segmentation stage, an appropriate recognizer (digit recognizer or cursive month word recognizer) is invoked for each of *Day*, *Month* and *Year* fields. Otherwise, the multi-hypotheses evaluation module which makes use of the results from the digit and word recognizers is invoked. The digit recognizer used in our date processing system was originally developed for processing the courtesy amount written on bank cheques [5], and a 74% recognition rate (without rejection) was reported for processing these courtesy amounts. For the recognition of cursive month words, a new combination method with an effective conditional topology has been implemented, and more discussions are given below on this method.

4.1. Cursive Month Word Recognition

Altogether 33 English/French month word classes have been observed in the CENPARMI Cheque database and the CENPARMI_IRIS Cheque database including full and abbreviated forms. Based on analyses of the databases, many similarities have been found among the word classes, which give rise to a challenge in designing an effective classifier. Since each pair of classes, "September" and "Septembre", "October" and "Octobre", "November" and "Novembre", and "December" and "Décembre", are very similar in shape and they represent the same month (one is in English and the other is in French), they are assigned to the same class. So only 29 classes will be considered.

A segmentation based grapheme level Hidden Markov Model classifier (HMM), and two Multi-Layer Perceptron classifiers (MLPA and MLPB) with different architectures and different features have been developed in CENPARMI for the recognition of month words [7]. In order to enhance the recognition performance, combinations of the three individual classifiers have been considered. Based on considerations of both speed and accuracy, as well as experimental findings, a conditional combination topology is proposed here to combine the three classifiers. In this architecture, MLPA and MLPB are first applied (in that order) using the serial strategy in the first stage; for samples rejected by both MLPA and MLPB, the decisions of all three classifiers (in parallel) are combined in the second stage. Here a new modified Product rule has been proposed to combine the decisions of the three classifiers in the second stage [7]. We found that this rule is superior to general Majority Vote, Sum, and Product rules in the combination system for the cursive month word recognition.

The performances of these individual classifiers and the combination system using the modified Product rule have been tested on a test set of 2063 month word samples extracted from the CENPARMI_IRIS Cheque database (consisting of real-life standard Canadian personal cheques), as shown in Table 2. In this table, the 12 outputs correspond to the 12 months. In the combination system, the rejection conditions of both MLPA and MLPB in tht first stage have been set very strictly (total error rate introduced in the first stage is 0.6% on the test set), and about 33% of the samples can be recognized in this stage.

Table 2. Performances of cursive month word classifiers

Classifier	Recognition rate(%)	
	29 outputs	12 outputs
MLPA	76.44	78.87
MLPB	75.42	77.27
HMM	66.89	69.70
Combination system	85.36	87.06

5. Experimental Results

Several experiments have been designed to test the performance of our date processing system. Since the performance of the cursive month word recognition has been discussed above, the performances of only the date segmentation module and the entire system will be given below.

5.1. Date Image Segmentation

The test set is derived from the CENPARMI_IRIS Cheque database, and it contains 3399 date images, of which 1219 samples are written in English, and the other 2180 samples are written in French. The segmentation results based on different rejection thresholds are given in Table 3 for both English and French samples. Some discussions on these results are given below.

Table 3. Performances of date segmentation system for the English and French sets

English	Correct	Rejection	Error
rejection rate 1	90.40%	1.81%	7.79%
rejection rate 2	74.57%	21.16%	4.27%
French	Correct	Rejection	Error
rejection rate 1	82.94%	4.22%	12.84%
rejection rate 2	65.69%	26.97%	7.34%

1. The rejection

Rejection rate 1 is obtained by trying to recognize every date sample, and a rejection is made when at least one of the three fields corresponding to *Year*, *Month* and *Day* cannot be found, e.g. *Day&Month* is written or binarized as one component or the *Year* field cannot be found. If strong noise such as a big blob of ink (due to improper binarization) or too many components have been detected in the preprocessing stage, a rejection is also made. For rejection rate 2, the

recognition result should be a valid date and the average confidence value of the three fields should exceed a threshold; otherwise a rejection is made.

2. The performance

The performance for the English set is better than that for the French set. Several reasons can account for this difference. First, based on the database analysis, it was found that more French cheques have the free format. With this format, more variations exist and detecting the handwritten *Year* is more difficult than detecting the machine-printed "19" or "20". Therefore more errors and rejections occur. In addition, the article "Le" sometimes used at the beginning of French date zones, together with freer writing styles on French cheques, also increase the difficulty of segmentation.

3. Error analysis

Based on our experiments in the training stage, the errors made can be categorized into two classes. The first class contains date images of very poor quality, and it includes: (i) date images having touching fields; (ii) strong noise introduced by improper binarization; and (iii) incomplete dates with one of the three fields missing. Our current segmentation module cannot process this type of date image, so a rejection is often made when rejection rate 2 is imposed. Based on the experiments, about 40% of the errors belong to this type when rejection rate 1 is adopted. The second class of errors consist of errors generated by all segmentation modules in the system.

4. The efficiency

The date segmentation is divided into two stages in order to improve the performance and efficiency of the system. The knowledge-based segmentation module is used in the first stage to solve most segmentation cases. Ambiguous cases are handled by multi-hypotheses generation and evaluation modules in a later stage. Experimental results show that 74.19% of the date images in English and 71.41% of the date images in French are processed in the knowledge-based segmentation module, and the others are processed by the multi-hypotheses generation and evaluation modules.

5.2. Overall Performances

The overall performances are given in Table 4, where the rejections are made under the same conditions as in Table 3. The errors of the date processing system mainly come from segmentation, month word misrecognition and/or numeral misrecognition. Currently a more effective verification module in the postprocessing stage is being developed to reduce the error rates and improve the recognition rates.

Table 4. Performances of date processing system for the English and French sets

English	Correct	Rejection	Error
rejection rate 1	62.34%	1.81%	35.85%
rejection rate 2	61.69%	21.16%	17.15%
French	Correct	Rejection	Error
rejection rate 1	57.75%	4.22%	38.03%
rejection rate 2	53.67%	26.97%	19.36%

6. Concluding Remarks

This paper proposes a system for automatically recognizing the date information handwritten on Canadian bank cheques. 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 multi-hypotheses generation and evaluation modules are invoked to make the final decision based on the recognition results and semantic and syntactic constraints. In addition, a new cursive month word recognizer has also been implemented based on a combination of classifiers. The complete system has produced very promising performances on a test set from a real-life standard cheque database.

References

[1] R. Fan, L. Lam, and C. Y. Suen. Processing of date information on cheques. *Progress in Handwriting Recognition*, pages 473–479, eds. A. C. Downton and C. Impedovo, World Scientific, 1997.

[2] G. F. Houle, D. B. Aragon, R. W. Smith, M. Shridhar, and D. Kimura. A multi-layered corroboration-based check reader. In *Proc. of the Int. Association for Pattern Recognition Workshop on Document Analysis Systems*, pages 495–546, Malvern, Pennsylvania USA, Oct. 1996.

[3] L. Lam, Q. Xu, and C. Y. Suen. Differentiation between alphabetic and numeric data using ensembles of neural networks. In *Proc. of 16th Int. Conf. on Pattern Recognition, vol.IV*, pages 40–43, Quebec City, Canada, August 2002.

[4] M. Morita, A. E. Yacoubi, R. Sabourin, F. Bortolozzi, and C. Y. Suen. Handwritten month word recognition on Brazilian bank cheques. In *Proc. of 6th Int. Conf. on Document Analysis and Recognition*, pages 972–976, Seattle, USA, September 2001.

[5] N. W. Strathy. Handwriting recognition for cheque processing. In *Proc. of the 2nd Int. Conf. on Multimodal Interface*, pages 47–50, Hong Kong, January 1999.

[6] C. Y. Suen, Q. Xu, and L. Lam. Automatic recognition of handwritten data on cheques – fact or fiction? *Pattern Recognition Letters*, 20(11-13):1287–1295, 1999.

[7] Q. Xu, J. Kim, L. Lam, and C. Y. Suen. Recognition of handwritten month words on bank cheques. In *Proc. of 8th Int. Workshop on Frontiers in Handwriting Recognition*, pages 111–116, Niagara-on-the-Lake, Canada, August 2002.