Automatic Extraction of Items from Cheque Images for Payment Recognition

Ke Liu[†*], Ching Y. Suen[†] and Christine Nadal[†]

[†]Centre for Pattern Recognition and Machine Intelligence, Suite GM-606, Concordia University
1455 de Maisonneuve Blvd. West, Montreal, Quebec H3G 1M8, Canada
[*]Department of Computer Science, Nanjing University of Science and Technology
Nanjing 210014, P.R. China

Abstract

A novel approach is proposed for the extraction of legal and courtesy amounts and date from cheque images based on the structural description of cheques. A method for the representation of cheques is presented. Several image processing techniques and algorithms have been developed in this approach. Experimental results show that the approach is effective and the proposed techniques and algorithms perform well.

1. Introduction

Automatic reading of bank cheques is an active topic in Document Analysis [1-9]. A machine capable of reading bank cheques will have wide applications in banks and those companies where huge quantities of cheques have to be processed. To realize such a system, many image processing, pattern recognition and OCR techniques must be involved [9].

The most basic step of automatic reading of cheques is to extract the interested items, such as courtesy amounts, legal amounts, dates, and so on. Only after this step has been done that one can start the process of recognizing the legal amount, courtesy amount and date, etc. However, to develop an effective item extraction system is a difficult task, especially when the cheques contain complicated and colourful background pictures. Most existing methods are based on binarization, i.e. thresholding the cheque images first. In our opinion, these methods will be useful in those applications where the cheques have only simple background pictures. However, in real applications, bank cheques may contain a variety of colourful backgrounds. In such cases, it is very difficulty to find a thresholding method which will produce a satisfactory binarized image.

In this paper, an approach has been taken to extract items from grey-scale cheque images. The interested items to be extracted from such an image are the legal amount, courtesy amount and date, respectively. At the present moment, we focus our efforts mainly on the processing of Canadian cheques which may be printed in different styles where a variety of pictures may appear in the background. The proposed image processing techniques should be applicable to cheques other than Canadian. The idea is as follows: first construct the structural description from a cheque image, then find the grey distributions of the pixels which correspond to the strokes of the legal amount, courtesy amount and date, respectively, and finally trace the the legal and courtesy amounts and date based on their respective *searching* and *bounding* regions. An automatic system has been developed based on this idea and related image processing techniques and algorithms. The results of a series of testing experiments will be presented.

2. Preprocessing

Straight lines are used to describe the structure of cheques. Hence, our approach makes use of them to locate the zones which contain the interested items. The first step is to extract all straight lines from an input image. The line extraction method is based on edge images because bank cheques may contain a variety of colourful backgrounds in most practical applications. An automatic edge thresholding technique is proposed to detect candidate edge points. A modified formula is used to calculate edge directions so that the normal directions of local edges point at their corresponding grey regions with lower grey values. An efficient and robust line extraction algorithm has been developed based on a constrained Hough Transform[12-13].

After preprocessing a cheque image, a set LS_P which contains the straight line segments parallel to cheque skew direction α is obtained.

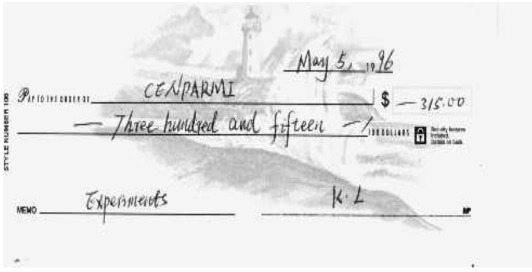

Figure 1. Original cheque image.

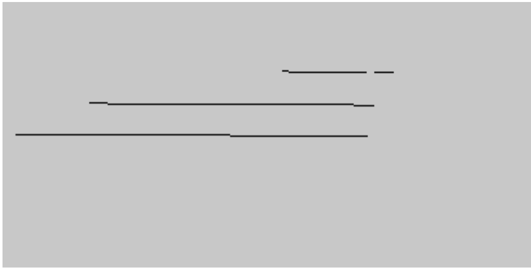

Figure 2. Structural description of cheque.

3. Description of cheques

A cheque is described by a subset of LS_P. For a standard Canadian cheque, the legal amount is always written above a baseline which is longer than half the width of the cheque. This baseline is called the "legal amount baseline". Above this is another baseline which is also longer than half the width of the cheque for filling in the name of the payee. This baseline is called the "payee baseline". The date is written above two short baselines at the same height defined as the "date baselines" (sometimes date is written above only one short baseline, in such a case, we assume that two baselines coincide). The date baselines are located above the payee baseline on the right hand side of the cheque. The courtesy amount is also located on the right hand side of the cheque between the date and legal amount baselines. For some cheques, there may be a short baseline on the right side of payee baseline and both baselines have the same height. In such cases, the courtesy amount is written above this baseline and the baseline is called courtesy baseline. Therefore, a cheque will be described by the legal amount baseline, payee baseline, date baselines, and courtesy baseline, and an algorithm [12-13] has been developed to construct descriptions of cheques from baselines. The algorithm tries to determine the legal amount and payee baselines first based on the cheque layout information and the control of the lengths of baselines, then proceeds to find the date baselines and courtesy baseline (if there is one on a cheque) also based on cheque layout information. Fig. 2 shows the finally derived structural description of the cheque as shown in Fig. 1.

4. Extraction of items from cheques

4.1. Determination of thresholds

In the following part of the paper, for each baseline $BL[i]$, its equation is denoted by:

$$y = a_i x + b_i \tag{1}$$

and its two end points are represented by $(x^i_L,\ y^i_L)$ and (x^i_R, y^i_R), respectively.

The legal and courtesy amounts and date in a cheque image can be considered as thin connected regions with their grey values lower than those of their local background regions. As mentioned before, an edge has a normal always pointing to their corresponding grey regions which possess lower grey values, therefore, for an edge point P of a stroke represented by a thin grey region, along its normal direction there is a corresponding point P^* on the other side of the edge such that the difference between their normal directions is approximately 180^o. Hence, the regions of the strokes of the legal and courtesy amounts and date can be detected by searching for the corresponding edge points which satisfy the above constraint. Once these regions have been detected, the grey distributions of legal amount, courtesy amount and date can be estimated based on the pixel values of these regions. The following summarizes the steps of determining the grey distributions of the legal amount:

(1) Determine a region R_L which contains the strokes (maybe partial or all) of the legal amount.
(2) Detect the thin connected regions with their grey values lower than those of their local background regions within R_L based on the technique of searching for the corresponding edge points.
(3) Calculate histogram $HIST_L$ based on the grey values of the points within the connected thin regions obtained in step (2).

R_L in step (1) can be easily determined and is bounded by the payee and legal amount baselines in the image coordinate system:

$$R_L = \{(x,y) | (x,y) \text{ satisfies inequalities (3) and (4)}\} \tag{2}$$

$$a_{i_p}x+b_{i_p} < y < a_{i_l}x+b_{i_l} \tag{3}$$

$$0 < x < x_R^{i_l} \tag{4}$$

where subscripts i_p and i_l indicate the payee and legal amount baselines, respectively. R_L is called legal amount zone.

$HIST_L$ obtained in step (3) can be considered as the estimation of the grey distributions legal amount, from which, the threshold T_l for getting the binary image of legal amount can be determined based on the following formula:

$$\frac{\sum_{i=0}^{T_l} HIST_L[i]}{\sum_{i=0}^{255} HIST_L[i]} \leq \beta \tag{5}$$

where β is a given value within [0, 1]. In theory, β can be taken as 1. However, in practice, it is recommended that β be taken from 0.75 to 0.95.

The basic idea and steps of determining the grey distributions and thresholds T_c and T_d of the courtesy amount and date are the same as described above.

4.2. Extraction of legal amount, courtesy amount and date

If the legal amount is filled completely inside the legal amount zone R_L, it can be extracted by finding the connected points in R_L of which the grey values are lower than T_l. In real-life cheques, the strokes of a legal amount frequently touch or cross the baselines. In such cases, extraction of the legal amount will be complicated. By observing real Canadian cheque images, it can be concluded that for the legal amount, touching or crossing mainly occurs between its strokes and the legal amount baseline. Note that the strokes may also touch the payee baseline, but the probability of the strokes touching a baseline other than the legal amount baseline is very low. Based on this reasoning, a method of extracting the legal amount is proposed. Suppose regions $R_s^{(l)}$ and $R_b^{(l)}$ are defined by

$$R_s^{(l)} = R_L \tag{6}$$

and

$$\mathrm{R}_b^{(l)}=\{(\mathrm{x,y})|(\mathrm{x,y}) \text{ satisfies inequalities (8) and (9)}\} \tag{7}$$

$$a_{i_p}\mathrm{x}+b_{i_p} < \mathrm{y} < a_{i_l}\mathrm{x}+b_{i_l}+\mathrm{D}_{pl}/(2\sin(\mathrm{a})) \tag{8}$$

$$0 < \mathrm{x} < x_R^{i_l}+\mathrm{T}_{bl} \tag{9}$$

where R_L is defined by (2)-(4), $D_{pl} = D(BL[i_p], BL[i_l])$ is the distance between the legal amount and payee baselines and T_{bl} is a given value.

Extraction of the legal amount is done by finding the connected components of which the grey values of their connected points are lower than T_l within a restricted region. All the initial searching points for tracing out the connected points must be chosen from $R_s^{(l)}$ and all the connected points must be restricted within the bounding region $R_b^{(l)}$.

After the above process, further analysis is needed to remove those connected components which may not belong to the strokes of the legal amount.

The method of extracting the date is similar to that of the legal amount except that the initial searching region $R_s^{(d)}$ and the bounding region $R_b^{(d)}$ here are given by

$$\mathrm{R}_s^{(d)}=\{(\mathrm{x,y}) \mid (\mathrm{x,y}) \text{ satisfies inequalities (11) and (12)}\} \tag{10}$$

$$a_{i_d^1}\mathrm{x}+b_{i_d^1}-D_{bl}/\sin(\mathrm{a}) < \mathrm{y} < a_{i_d^1}\mathrm{x}+b_{i_d^1} \tag{11}$$

$$x_L^{i_d^1}-T_{sd} < \mathrm{x} < x_R^{i_d^2}+T_{sd} \tag{12}$$

and

$$\mathrm{R}_b^{(d)}=\{(\mathrm{x,y}) \mid (\mathrm{x,y}) \text{ satisfies inequalities (14) and (15)}\} \tag{13}$$

$$a_{i_d^1}\mathrm{x}+b_{i_d^1}-D_{bl}/\sin(\mathrm{a}) < \mathrm{y} < a_{i_d^1}\mathrm{x}+b_{i_d^1}+D_{bl}/(2\sin(\mathrm{a})) \tag{14}$$

$$x_L^{i_d^1} - D_{bl}/2 < \mathrm{x} < x_R^{i_d^2} + D_{bl}/2 \tag{15}$$

where T_{sd} in (12) is a given constant and subscripts i_d^1 and i_d^2 indicate the two date baselines.

The method of extracting the courtesy amount is also based on the determination of the initial searching region $R_s^{(n)}$ and the bounding region $R_b^{(n)}$. Although there is a bounding box for filling the courtesy amount on most cheques, the bounding edge lines cannot be extracted easily due to the low grey contrast against the background. The courtesy amount is sometimes positioned on the right hand side of the legal amount baseline. It may also be located on the right of the payee baseline. However, no matter where it is positioned, there is always a machine printed dollar sign '$' in front of the courtesy amount. Therefore, a courtesy amount item extractor **Num_Extr_D** is developed based on finding '$' in a cheque image.

In real applications, the strokes of courtesy amount may even touch or cross '$' occasionally. In such cases, **Num_Extr_D** cannot be directly used. After examining a large number of real cheques, we have derived the following rules which can be applied to a standard Canadian cheque:

(1) If there is a short baseline on the right hand side of the payee baseline and it has the same height as that of the payee baseline, the courtesy amount lies above this short baseline;
(2) If the right end points of both legal amount and payee baselines almost reach the right edge of the cheque, the courtesy amount lies above the right part of the payee baseline;
(3) If the conditions corresponding to above two rules are not satisfied, the courtesy amount may lie on the right hand side of either the legal amount or payee baseline.

We have developed other four procedures **Num_Extr_1**, **Num_Extr_2** , **Num_Extr_3**, and **Num_Extr_4** for the extraction of the courtesy amount, which are used in different

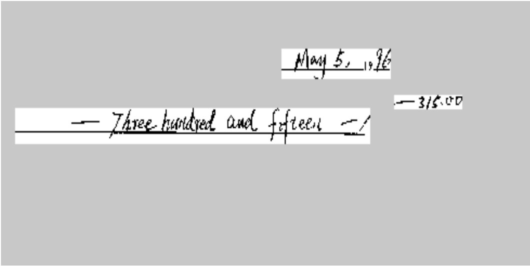

Figure 3. Result of item extraction.

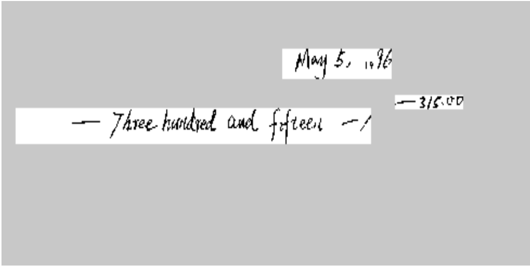

Figure 4. Result of item extraction after removing touching lines.

situations. **Num_Extr_1** is applied if there is a short baseline on the right hand side of the payee baseline and it is also based on the principle of determining the initial searching and bounding regions. **Num_Extr_2** is used when the condition corresponding to rule (2) applies. **Num_Extr_3** and **Num_Extr_4** correspond to rule (3) above and are used to extract the courtesy amount located on the right hand side of payee baseline and the legal amount baseline, respectively. **Num_Extr_2 , Num_Extr_3** and **Num_Extr_4** are all developed based on the same principle. The idea is as follows: first extract the connected components of pixel points of which the grey values are lower than T_c within a specified region, then determine the bounding rectangle of each extracted connected component, and the courtesy amount is finally extracted by grouping the bounding rectangles from the right to the left of the specified region according to the distance between two consecutive bounding rectangles. The algorithms for **Num_Extr_3** and **Num_Extr_4** are similar. However, in the algorithm for **Num_Extr_2** , the pixels on the payee baseline should be excluded to determine the bounding rectangles of the connected components.

A general method of extracting the courtesy amount can be obtained by combining **Num_Extr_D**, **Num_Extr_1**, **Num_Extr_2**, **Num_Extr_3** and **Num_Extr_4**. Fig. 3 shows the legal and courtesy amounts and date extracted from the image in Fig. 1 by the proposed method.

4.3. Separation of strokes connected lines

The connected components extracted from a cheque image may contain both strokes and lines which either touch or cross each other, or the baselines, or occasionally the bounding box lines of the courtesy amount. Therefore, post-processing is needed to remove these lines, which do not belong to the strokes of the legal and courtesy amounts, and date, from the connected components. A new and effective technique of separating strokes from connected baselines has been developed. A new and effective technique has also been developed to separate strokes from connected lines of the bounding boxs of courtesy amounts on some cheques[12]. Fig. 4 is the result after removing lines.

5. Experimental results

An automatic cheque item extraction system has been developed in the environment of X-Window and SUN/SPARC based on the proposed approach, which is a sub-system of an **Automated Payment Recognition System** developed by **CENPARMI** (Centre for Pattern Recognition and Machine Intelligence, Concordia University). The input of the system is a scanned grey level cheque image which has a TIFF file, and the output provides the legal amount, courtesy amount and date items. The system has been applied to process real bank cheques and the results indicate that the proposed approach is effective and its performance is encouraging.

Table 1 summarizes the results of extracting the items from 494 real cheque images which are printed in different styles and contain a variety of colourful backgrounds, where **l_amount** and **c_amount** indicate legal and courtesy amounts, **C_ITEMS** and **M_ITEMS** correspond to the items which can or cannot be correctly located and extracted, respectively, and **Uns_ITEMS** stands for the extracted items with unsatisfactory image quality for their corresponding recognizers, i.e. the legal amount recognizer, courtesy amount recognizer, and date recognizer. The factors which affect the quality of these images can be divided into two categories. The first comes from the item extraction system, i.e. caused by the binarization process due to the quality of original images, by baseline removal algorithm, by un-recognized '$', and so on. The second category mainly comes from poor handwritings and complicated background pictures, i.e. caused by the handwritings which touch the machine printed characters and '$' sign on the cheques, by the handwritings which are filled in the wrong regions, and by various noise components generated from the background pictures.

6. Conclusions

One of the most important research issues in automatic reading of cheques is to extract interested items such as le-

Items	C_ITEMS	M_ITEMS	Uns_ITEMS
l_amount	99.60%	0.40%	11.33%
c_amount	97.77%	2.23%	18.83%
date	98.99%	1.01%	13.16%

Table 1. Testing results on real bank cheque images.

gal amount, courtesy amount and date etc. from the cheque images. However, to develop an effective item extraction system is a difficult task, especially when the processed cheques have different styles and contain complicated background pictures.

This paper presents a novel approach to cheque item extraction based on the determination of baselines of cheques and a-priori information about the positions of the legal and courtesy amounts and date. Several new image processing techniques and algorithms are proposed, which have been implemented in the environment of X-Window and SUN/SPARC. The results of a series of experiments showed that the proposed approach, techniques and algorithms are effective. However, to realize a robust cheque item extraction system for real applications, further study is needed in improving the performance of the individual algorithms so that the extracted items possess better image quality. The study of post-processing the extracted items is also meaningful.

Acknowledgements

This work was supported by the Canada International Fellowship Program from the Natural Sciences and Engineering Research Council of Canada, the National Networks of Centres of Excellence of Canada, the Ministry of Education of Quebec, and Bell Quebec.

References

[1] J. V. Moreau, B. Plessis, O. Bougeois and J. L. Plagnaud, " A postal check reading system," Proc. of the First International Conference on Document Analysis and Recognition (ICDAR'91), Saint-Malo, France, 1991, pp. 758-766.

[2] M. Leroux, J. C. Salome and J. Badard, "Recognition of cursive script words in a small lexicon," Proc. of the First International Conference on Document Analysis and Recognition (ICDAR'91), Saint-Malo, France, 1991, pp. 774-782.

[3] E. Lethelier, J. F. Couthouis and M. Lerour, " Automatic processing of numerical amounts on postal cheques," Proc. of the First European Conference Dedicated to Postal Technologies, Nantes, France, 1993, pp. 697-704.

[4] T.M. Ha and H. Bunke, "Model-based analysis and understanding of cheque forms," International Journal of Pattern Recognition and Artificial Intelligence, Vol. 8, No. 5, 1994, pp. 1053-1080.

[5] D. Guillevic and C.Y. Suen,"Cursive script recognition applied to the processing of bank cheques," Proc. of the Third International Conference on Document Analysis and Recognition (ICDAR'95), Montreal, Canada, 1995, pp. 11-14.

[6] J.P. Dodel and R. Shinghal, "Symbolic/neural recognition of cursive amounts on bank cheques," Proc. of the Third International Conference on Document Analysis and Recognition (ICDAR'95), Montreal, Canada, 1995, pp. 15-18.

[7] A. Agarwal, L. Granowetter, K. Hussein, A. Gupta, and P.S.P. Wang,"Detection of courtesy amount block on bank checks, " Proc. of the Third International Conference on Document Analysis and Recognition (ICDAR'95), Montreal, Canada, 1995, pp. 748-751.

[8] E. Lethelier, M. Lerour, and J. F. Couthouis, "An automatic reading system for handwritten numeral amounts on French checks," Proc. of the Third International Conference on Document Analysis and Recognition (ICDAR'95), Montreal, Canada, 1995, pp. 92-97.

[9] L. Lam, et al, "Automatic processing of information on cheques," Proc. IEEE Int. Conf. on Systems, Man, Cybernetics, Vancouver, Oct. 1995, pp. 2353-2358.

[10] A.K. Jain, Fundamentals of Digital Image Processing, Prentice Hall, Inc., Englewood Cliffs, New Jersey 07632, USA, 1989.

[11] K. Liu, The study of image processing techniques for the optical testing of flow field, Ph.D Thesis, Nanjing University of Science and Technology, May 1989.

[12] K. Liu, et al, "The development of cheque image segmentation technique," Technical Report, CENPARMI, Concordia University, Montreal, Canada, Sept. 1995.

[13] K. Liu and C.Y. Suen, "Robust method for the extraction and identification baselines of cheques," Technical Report, CENPARMI, Concordia University, Montreal, Canada, March, 1995.