3rd IAPR International Workshop on Document Analysis Systems (DAS), Nagano, Japan, October/1998, pp.141-144

A NOVEL APPROACH FOR AUTOMATIC EXTRACTION OF THE USER–ENTERED DATA FROM BANKCHECKS

A. L. KOERICH
Centre for Pattern Recognition and Machine Intelligence, Concordia University,
1455 de Maisonneuve Blvd. West, Montreal, QC, H3G 1M8, Canada
alekoe@cenparmi.concordia.ca

L. L. LEE
Dept. of Communications, State University of Campinas,
CP 6101, 13083–970, Campinas, SP, Brazil
zhao@lrprc.fee.unicamp.br

This paper presents a system for automatic extraction of the user–entered data from Brazilian bankchecks. Experimental results show that this approach is robust for extracting the user–entered items achieving a moderate processing time and excellent accurancy rates

1 Introduction

Millions of handwritten or machine printed bankchecks have to be processed everyday. Since the bankcheck processing is merely a repetitional task it is desirable to realize it in an automatic fashion. Currently, thresholding [2,3,4] and image subtraction [1,6,7] techniques are being used for extracting the user–entered data. The techniques based on image subtraction have shown more robustness to segment the user–entered data. Okada–Shridhar[7] have suggested an approach based on a subtraction operation between one filled–in bankcheck image and the same bankcheck image without the filled–in information. This approach might not be feasible for real–life applications since a large database is needed to keep the sample images of unfilled bankchecks. The main idea of this new approach is to handle only sub–images which contain the user–entered data. A template is used to extract these areas. For each of the five resulting sub–images, the background pattern is eliminated by a subtraction operation, the baselines are detected and eliminated by using an algorithm based on the projection profiles and the character strings are eliminated by a subtraction operation between the sub–image and a generated binary image which contains every make–up character string. Finally, some pixels lost during the foregoing processing are recovered.

2 User–Entered Data Extraction

The proposed system is composed by three main modules: data acquisition, database, and image processing, as shown in Fig. 1. The data acquisition module includes an optical scanner and a MICR scanner. The database stores

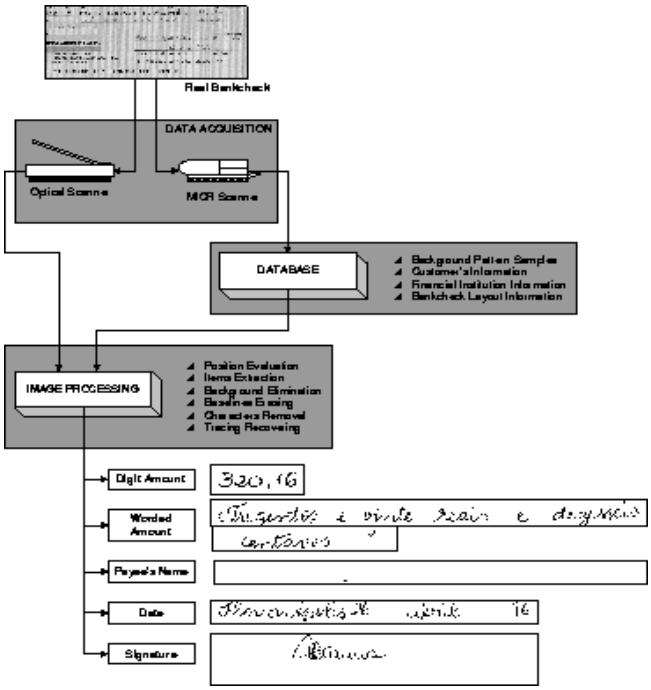

Figure 1: A system overview

samples of the background patterns, customer's information, parameters to adjust the template and other useful data. The information stored in the database is indexed by keywords composed by the bank, agency, and account numbers. The image processing module has several algorithms, such as, position evaluation, items extraction, background pattern elimination, baseline erasing, character elimination and a tracing recovering.

Since the Brazilian bankchecks have a standardized layout structure, it is reasonable to design a template for extracting the user–entered data from any bankcheck. From a basic template, we construct a database with the possible positions of the user–entered data according to the financial institution that has issued the bankcheck. The output of the item extraction algorithm are five sub–images representing the user–entered items: digit amount, worded amount, payee's name, city and date, and signature. The elimination of the background pattern is done through a modified morphological subtraction operation performed between the sub–images and the corresponding background

Table 1: Results for extracting items.

Items	Correctly Extracted(%)	Unused(%)
Digit Amount	99.2	0.0
Worded Amount	95.8	2.5
Payee's Name	96.7	1.7
City and Date	95.0	4.2
Signature	90.8	5.8

pattern sample sub–image. Since the baselines generally have horizontal and vertical length close to the proper sub–image dimension, they can be easily identified through the positions where the projection profile has relatively high values. Then, these positions are filled–in with white pixels. The information previously stored in the database is used to generate a binary image to eliminate the printed character strings presented under signature baseline. This generated image contains exactly the same character strings that must be eliminated. A subtraction operation between this generated image and the corresponding sub–image eliminates the character strings. The elimination of baselines can also eliminates some pixels of the user–entered data. To recover this pixels, the positions of the extracted baselines are used and connections between the neighbors are established. For each pixel that may be fitted to a erased baseline pixel, their 8–neighbors are analyzed and filled with black pixels if some neighbors were black.

3 Experimental Results

The proposed system was tested by 120 real–life Brazilian bankchecks. The bankchecks were issued by 10 different financial institutions and were filled–in by 20 different writers. Tab. 1 shows the results for extracting the user–entered data. The images were classified according to the visual quality, which is related to the physical integrity of the extracted items. A comparison among the results provided by the proposed method and other methods are shown in Tab. 2. The results provided by this comparison rely on different databases and application domains since the papers have focused different kinds of bankchecks such as American, French, Brazilian, and Canadian bankchecks.

4 Discussion and Conclusion

This paper presents a system for automatic extracting the user–entered data from bankchecks. A complete solution for data acquisition, layout analysis, MICR data identification and recognition, user–entered data extraction, and tracing recovering is presented. The input of the system is real–live bankchecks

Table 2: A comparison of different extraction methods.

Method	Database	Correctly Extracted Items(%)					
		DA	WA	PN	PL	DT	SG
Proposed Method	120	99.2	95.8	96.7	95.0	95.0	90.8
Koerich–Lee[1]	200	98.3	95.3	97.5	95.2	96.1	88.7
Liu[4] et al.	400	97.5	97.0	NA		97.0	NA
Heutte[5] et al.	3374	98.0	NA				
Dimauro[2] et al.	300	95.4					
Santos[3] et al.	20	95.0					
DA=Digit Amount, WA=Worded Amount, PN=Payees Name, PL=City of Issue, DT=Date, SG=Signature, NA=Non Available Data							

and the output provides digital images of the isolated filled–in items. This novel approach has reached similar rates as our previous approach in terms of accuracy in extracting the user–entered data, but in terms of processing time, it has reached a reducing of 300% [1]. Our system has addressed a complete solution for identifying and extracting the information from bankchecks. The other approaches focus mainly the extraction of the digit and worded amounts. By combining this system with digit recognition, word recognition and signature verification systems, an automatic bankcheck processing system might be feasible for practical applications [6].

References

1. A. L. Koerich, L. L. Lee. Automatic Extraction of Filled–in Information from Bankchecks Based on Prior Knowledge About Layout Structure. In *First Braz. Symp. on Doc. Image Anal.*, pp.322–333, Curitiba, Nov. 1997.
2. G. Dimauro, S. Impedovo, G. Pirlo, A. Salzo. Automatic Bankcheck Processing: A New Engineered System. *IJPRAI*, 11(4):467–504, 1997.
3. J. E. B. Santos, F. Bortolozzi, R. Sabourin. A Simple Methodology to Bank Cheque Segmentation. In *First Braz. Symp. on Doc. Image Anal.*, pp.334–343, Curitiba, Nov. 1997.
4. K. Liu, C. Y. Suen, M. Cheriet, J. N. Said, C. Nadal, Y. Y. Tang. Automatic Extraction of Baselines and Data from Check Images. *IJPRAI*, 11(4):675–697, 1997.
5. L. Heutte, P. B. Pereira, J. V. Moreau, B. Plessis, P. Courtellemont, Y. Lecourtier. Multi–Bank Check Recognition System: Consideration on the Numeral Amount Recognition Module. *IJPRAI*, 11(4):549–570, 1997.
6. L. L. Lee, M. G. Lizarraga, N. R. Gomes, A. L. Koerich. A Prototype for Brazilian Bankcheck Recognition *IJPRAI*, 11(4):549–570, 1997.
7. M. Okada, M. Shridhar. Extraction of User Entered Components from a Personal Bankcheck Using Morphological Subtraction. *IJPRAI*, 11(4):699–715, 1997.