Handwritten Month Word Recognition on Brazilian Bank Cheques

M. Morita[1,2], A. El Yacoubi[1], R. Sabourin[1−3], F. Bortolozzi[1] and C. Y. Suen[3]
[1]PUCPR Pontifícia Universidade Católica do Paraná (PPGIA-LARDOC)
Rua Imaculada Conceição 1155, 80215-901 - Curitiba, PR - BRAZIL
{marisa,yacoubi,fborto}@ppgia.pucpr.br

[2]ETS - Ecole de Technologie Supérieure (LIVIA)
1100, rue Notre Dame Ouest, Montreal, H3C 1K3, CANADA
sabourin@gpa.etsmtl.ca

[3]CENPARMI Centre for Pattern Recognition and Machine Intelligence
1455 de Maisonneuve Blvd. West, Suite GM 606 - Montreal, H3G 1M8, CANADA
suen@cenparmi.concordia.ca

Abstract

This paper describes an off-line system under development to process unconstrained handwritten dates on Brazilian bank cheques in an omni-writer context. We show here some improvements on our previous work on isolated month word recognition using Hidden Markov Models (HMMs). After preprocessing, a word image is explicitly segmented into characters or pseudo-characters and represented by two feature sequences of equal length, which are combined using HMMs. The word models are generated from the concatenation of appropriate character models. In addition to the small date database, we also make use of the legal amount database to increase the frequency of characters in the training and the validation sets. Although this study deals with a limited lexicon, the many similarities among the word classes can affect the performance of the recognition. Experiments show an increase in the average recognition rate from 84% to 91%. Finally, we present our perspectives of future work.

1 Introduction

Various studies have been dedicated to the processing of off-line handwritten legal and courtesy amounts on cheques. However, the literature shows that only a few have been made on handwritten date recognition in the last decade [1, 7]. Although in these publications, the dates come from Canadian cheques and our research focuses on Brazilian cheques, they have similar problems, e.g., in both contexts the system must process a complex entity composed of different data types such as words and digits. This is one of the reasons why the development of a date recognition engine is one of the most challenging parts of a cheque processing system.

In our application, the date information from left to right can consist of: city name, separator (comma), day (numerical), separator ("de"), month (alphabetical), separator ("de") and year (numerical).

Due to the complexity of processing the entire date information at the same time, we started working on the month word recognition. This study deals with a limited lexicon of 12 classes, but still there are a lot of similarities among some word classes which can affect the performance of the recognition, e.g., the termination in "embro" for the classes "Setembro", "Novembro" and "Dezembro". To face this problem, we presented in our last work [6] a combination strategy involving both holistic and analytical HMM approaches, where an explicit segmentation technique was applied. Using this, we obtained the satisfactory average recognition rate of 84% given the small size of our date database and the limitations of our feature set based only on the detection of loops, ascenders and descenders, namely global features.

This paper reports some improvements on our previous work to recognize isolated month words. Because our date database is too small, we also added the legal amount database in order to increase the frequency of characters in training and validation sets. For this reason, we consider only an analytical approach. A feature set based on concavity analysis was used to improve the discrimination among several writing styles and then combined with global features through HMMs. In this way, a word image is rep-

resented by two feature sequences of equal length to feed our HMMs. Experiments show an increase in the average recognition rate from 84% to 91%.

In section 2, we detail the feature set based on concavities. In section 3, we describe our HMM approach and we show the new results in section 4. Finally, section 5 presents our conclusions and perspectives of future work.

2 Preprocessing, Segmentation and Feature Extraction

As mentioned earlier, two feature sets are used: global and concavity features. Only the second feature set is not invariant to the slant correction. To overcome potential problems, we decided to correct the slant. After preprocessing, a word image is explicitly segmented into graphemes that can represent characters or pseudo-characters. Both feature sequences of equal length are extracted from the grapheme sequences and combined using HMMs.

2.1 Slant Correction

Our technique is based on [8], which uses the external contour of the image. The external contour of each connect component is divided into segments with the same length as the stroke thickness. The global slant is obtained by the median slant of all segments and its correction is made using a shear transformation.

2.2 Segmentation of Words into Characters

The generation of a segmentation point (SP) results from the validation of a local minimum on the upper contour (MP). This validation is based on the detection of the lower contour from the MP vertical projection. The SP is shifted from the original MP when its projection finds a loop, contour tangency or an unacceptable width. If some unacceptable width conditions are found in the MP neighborhood, we validate the upper contour point that minimizes the vertical width in order to avoid a character under-segmentation. A word length contributes to the discrimination between the word classes (Figure 1a, for more details see [6]). For this reason, in this work we permit a SP to be cut in the vertical direction followed by the horizontal direction (Figure 1b).

2.3 Concavity Measurements and Feature Extraction

The basic idea of concavity measurements [3] is the following: for each white pixel in the grapheme, we verify in each possible direction (Figure 2a), if a black pixel can be reached. The number of times as well as the directions leading to the black pixels are computed and stored in a vector.

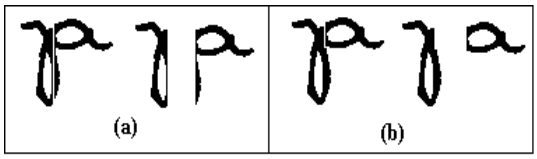

Figure 1. (a) SP that minimizes the vertical width only, (b) SP cut in the vertical and horizontal directions

When black pixels are reached in four directions (e.g., point x_1 in Figure 2c), we branch out in four auxiliary directions (s_1 to s_4 in Figure 2b) in order to confirm if the current white pixel is really inside a closed contour.

To reduce the feature vector as well as the handwriting variability, we are considering only the more relevant concavities. We are labeling the white pixels which have black pixels in three or four directions (5 configurations). The white pixels with two or less directions are often located on the grapheme borders (e.g., point x_2 in Figure 2c). The labeling leads to a vector with 9 components for each grapheme: s_1 to s_4 and 5 configurations. Since we are zoning the grapheme into 2 equal parts along the vertical axis, the final feature vector is composed of 18 components normalized between 0 and 1.

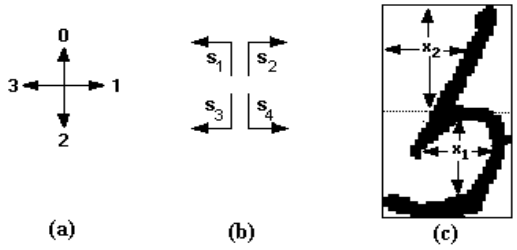

Figure 2. Concavity features

The feature vector sequence of each image is turned into a sequence of symbols to feed our discrete HMMs. In such a case, clustering is done by a vector quantization algorithm. The codebook size adopted in each experiment (Section 4) is chosen after several tests carried out on the validation set.

Figure 3 shows an example of a word image represented by two feature sequences of equal length.

3 Markovian Modeling of Handwritten Words

This section describes the justification behind the proposed model architecture as well as the training and recog-

F_1: ZZZ OOO --- --- ooo --- --- --- ddd

F_2: 047 087 012 106 048 111 052 040 076

Figure 3. Pair of feature sequences representing a word image

nition phases used in our system. The markovian modeling remains the same as in our previous work [6].

3.1 Character Model Architecture

In this application we observe that the characters can be represented by a maximum of three graphemes. Therefore, four states are needed for each character model as shown in Figure 4. In such a case, the observation sequences are emitted from the model transitions in order to take advantage of the explicit segmentation. Table 1 describes the several configurations provided by our segmentation. The model architecture adopted is based on [8, 9, 10].

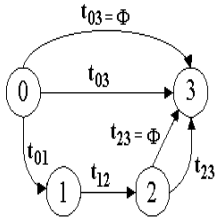

Figure 4. Character model architecture

Table 1. Transitions allowed by the character models

Transition	Description
$t_{03} = \Phi$	character under-segmentation
t_{03}	no segmentation inside a character
t_{01}-t_{12}-($t_{23} = \Phi$)	character over-segmented into 2 graphemes
t_{01}-t_{12}-t_{23}	character over-segmented into 3 graphemes

The month alphabet comprises 20-character classes. In this manner, we have 40 HMMs considering uppercase and lowercase characters.

3.2 Training and Recognition

Since the writing style (uppercase/lowercase) of each training word image is available, the word model is generated from the concatenation of appropriate character models. The last state of a character model becomes the initial state of the next model, and so on (see Figure 5). We are using an embedded Baum-Welch algorithm with the Cross-Validation procedure [10]. In such a case, the character boundaries do not need to be manually identified, in opposite to the works presented in [2] and [5].

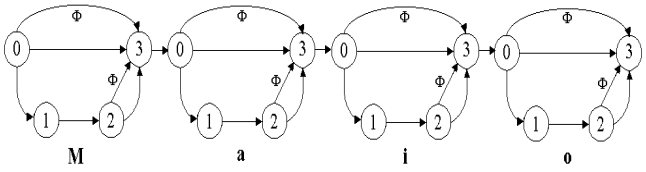

Figure 5. Training model of class "Maio" (May)

The word model construction in recognition remains basically the same as in training. However, as the writing style (uppercase, lowercase) of an unknown word is not available, we adopted two character models in parallel (uppercase, lowercase) (see Figure 6) [9, 10]. The word model consists of an initial state (I) and a final state (F), and two consecutive character models linked by four transitions: two uppercase characters (UU), two lowercase characters (LL), one uppercase character followed by one lowercase character (UL) and one lowercase character followed by one uppercase character (LU). The probabilities of these transitions are estimated by their occurrence frequency in the training set. In the same manner, the probabilities of beginning a word by an uppercase character (0U) or lowercase character (0L) are also estimated. This architecture handles the problem related to the mixed handwritten words detecting implicitly the writing style during recognition using the Backtracking procedure of the Viterbi algorithm.

4 Experiments and Analysis

From our laboratory date database that contains 2,000 images, we used 1,188, 408 and 402 images for training, validation and testing respectively, considering an omni-writer context. Since we are dealing with a small database, we also use the legal amount database to train our character models. The legal amount database contains images of isolated words. The training and validation sets have as a whole about 10,200 images. The images in both databases have a resolution of 300 dpi.

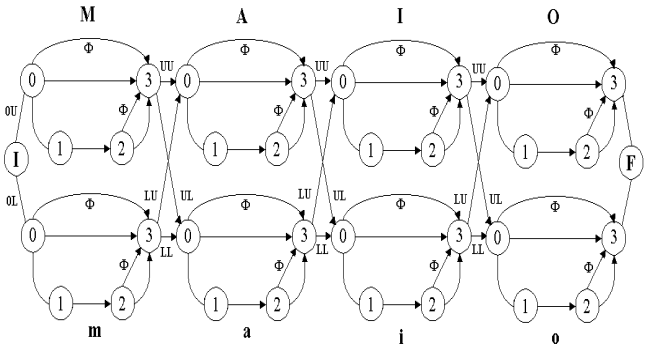

Figure 6. Recognition model of class "Maio" (May)

The experiments presented in Table 2 were carried out on the test set using the Forward procedure. The second column shows the database(s) used in each experiment. The date and legal amount databases are represented by D_1 and D_2 respectively. The feature sets are described in the third column where the set F_1 corresponds to the global features and the set F_2 is related to the concavity features. The combination of both feature sets using HMMs (HMMs with multiple-codebooks) is represented by $\{F1, F2\}$. The fourth column shows the codebook size used in each experiment. Considering the concavity features, the codebook size was chosen after several tests carried out on the validation set as we mentioned earlier. The last column details the average recognition rate (RR) of each experiment with no rejection.

Table 2. Average recognition rates obtained in the experiments

Exp.	Database	Feature Set	Codebook Size	RR %
I	D_1	F_1	20	82.1
II	D_1	F_2	20	80.1
III	D_1	$\{F_1, F_2\}$	{20,20}	90.0
IV	$D_1 \cup D_2$	F_1	20	85.8
V	$D_1 \cup D_2$	F_2	150	87.8
VI	$D_1 \cup D_2$	$\{F_1, F_2\}$	{20,150}	91.0

From Table 2 we notice that using the legal amount database, we were able to increase the recognition rates by 3.7% (Exp IV vs I), 7.7% (Exp V vs II) and 1% (Exp VI vs III). The last result can be explained by the fact that the combination of both feature sets in the last experiment became less complementary than in the third.

Table 3 shows the five best recognition results for ranks 1 to 5 (R(1) to R(5)) carried out in the last experiment in order to show the real performance of this system in the context of a multi-hypothesis analysis. Moreover, we show the results achieved in our previous work [6] in order to compare with the new results obtained in this paper. The improvements related to this experiment increased the average recognition rate by 6.8% and 4.0% for R(1) and R(2) respectively.

Table 3. The five best recognition rates on the test set

Exp.	R(1)	R(2)	R(3)	R(4)	R(5)
VI	91.0%	97.0%	98.0%	99.0%	99.5%
Ref. [6]	84.2%	93.0%	95.5%	98.0%	99.0%

Table 4 details the recognition rate for each word class, for the third and sixth experiments respectively. The last line denotes the RR for each experiment.

Table 4. Recognition results on the test set

Class	n^0 of Images	Exp. III	Exp. VI
Jan*eiro*	39	92.3%	100.0%
Fever*eiro*	32	93.7%	90.6%
Março	36	80.6%	75.0%
Abril	39	92.3%	94.9%
Maio	38	84.2%	81.6%
Junho	29	89.7%	89.7%
Julho	32	81.2%	81.2%
Agosto	28	92.9%	92.7%
Set*embro*	31	93.5%	100.0%
Outubro	30	100.0%	100.0%
Nov*embro*	34	97.1%	97.1%
Dez*embro*	34	85.3%	91.2%
Total	**402**	**90.0%**	**91.0%**

Although this study deals with a limited lexicon of 12 classes, there are classes that contain a common sub-string that can affect the performance of the recognition such as:

- the termination in "eiro" for "Janeiro" and "Fevereiro";
- the termination in "embro" for "Setembro", "Novembro" and "Dezembro";
- almost all characters between "Junho" and "Julho" and between "Maio" and "Março".

Figures 7a, 7b and 7c show some examples of well-recognized images and Figures 7d, 7e and 7f show some

recognition errors. The errors in our system correspond to under-segmentation due to the lack of local minima especially in uppercase words (Figure 7e), confusion of the classifier (Figure 7d), high character distortion (Figure 7f), etc.

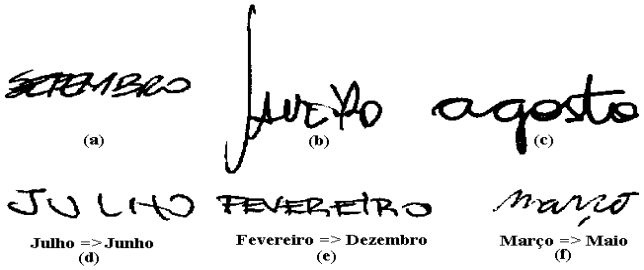

Figure 7. (a),(b) and (c) Examples of well-recognized images, (d), (e) and (f) Examples of misrecognized images

Few studies have been made on handwritten date recognition as mentioned earlier as well as for month word recognition. The closest research to our current work would be Kim et al [4]. They carried out some experiments using cursive month words written in English (21 classes, considering some month abbreviations). In such a work, a methodology of combining an HMM-based approach using an explicit segmentation and a MLP-based approach (multilayer perceptron) using an implicit segmentation was presented. The average recognition rates achieved were 77.6% and 86.2%, for the first and second classifiers respectively. The weighted multiplication combination method reached an average recognition rate of 87.3%. The database of CENPARMI was used in these experiments, which contains 4,413 images for training and 2,152 images for testing.

In our experiments, the average recognition rates obtained were 90% and 91%, considering the date database and both databases respectively. In these experiments unconstrained handwritten month words were considered.

5 Conclusion and Future Work

The challenge of this study remains in the problematic scheme of processing as a whole several data types such as words and digits in an omni-writer context. Due to the complexity of processing the entire date information at the same time, we started working on the month word recognition. This paper is an extension of our previous work, which deals with a small date database and a feature set more adapted to cursive words. To face these problems, in this work we are using the legal amount database to increase the frequency of characters in both the training and validation sets. Moreover, we are using a feature set based on concavities to improve the discrimination of uppercase characters as well as lowercase characters. Both feature sets are combined using HMMs. Experiments showed an increase in the average recognition rates from 84% to 91% and from 93% to 97% for R(1) and R(2) respectively.

Our future work will be dedicated to consider the date field as a whole in order to process the day, the month and the year in the same system.

Acknowledgements

The authors wish to thank Pontifícia Universidade Católica do Paraná and Paraná Tecnologia which have supported this work.

References

[1] R. Fan, L. Lam, and C. Y. Suen. Processing of date information on cheques. In 5^{th} *IWFHR*, pages 207–212, September 1996.

[2] A. M. Gillies. Cursive word recognition using hidden markov models. In 5^{th} *U.S. Postal Service Advanced Technology Conference*, pages 557–562, 1992.

[3] L. Heutte, J. Moreau, B. Plessis, J. Plagmaud, and Y. Lecourtier. Handwritten numeral recognition based on multiple feature extractors. In 2^{nd} *ICDAR*, pages 167–170, 1993.

[4] J. H. Kim, K. K. Kim, C. P. Nadal, and C. Suen. A methodology of combining HMM and MLP classifiers for cursive word recognition. In 15^{th} *ICPR*, volume 2, pages 319–322, Barcelona-Spain, September 2000.

[5] M. A. Mohamed and P. Gader. Handwritten word recognition using segmentation-free hidden markov modeling and segmentation-based dynamic programming techniques. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(5):548–554, May 1996.

[6] M. Morita, E. Lethelier, A. E. Yacoubi, F. Bortolozzi, and R. Sabourin. An hmm-based approach for date recognition. In 4^{th} *IAPR International Workshop on DAS*, pages 233–244, Rio de Janeiro-Brazil, December 2000.

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

[8] A. E. Yacoubi. *Modlisation Markovienne de L'criture Manuscrite Application la Reconnaissance des Adresses Postales*. PhD thesis, Universit de Rennes 1, France, Septembre 1996.

[9] A. E. Yacoubi, M. Gilloux, R. Sabourin, and C. Y. Suen. An hmm-based approach for off-line unconstrained handwritten word modeling and recogtion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(8):752–760, August 1999.

[10] A. E. Yacoubi, R. Sabourin, M. Gilloux, and C. Y. Suen. Off-line handwritten word recognition using hidden markov models. In L. Jain and B. Lazzerini, editors, *Knowledge Techniques in Character Recognition*. CRC Press LLC, April 1999.