Handwriting recognition, authenticity verification, and assurance of integrity of scanned documents

Review of the Literature on Month Word Recognition

Luiz S. Oliveira

Federal University of Parana (UFPR)
Department of Informatics (DInf)
R. Rua Cel. Francisco H. dos Santos, 100, Curitiba, PR, Brazil
lesoliveira@inf.ufpr.br

1 Introduction

Writing, which has been the most natural mode of collecting, storing, and transmitting information through the centuries, now serves not only for communication among humans but also serves for communication of humans and machines. Machine simulation of human reading has been the subject of intensive research for the last four decades. However, the early investigations were limited by the memory and power of the computer available at that time. With the explosion of information technology, there has been a dramatic increase of research in this field since the beginning of 1980s. The interest devoted to this field is not explained only by the exciting challenges involved, but also the huge benefits that a system, designed in the context of a commercial application, could bring.

Each day, billions of business and financial documents have to be processed by computer. The great bulk of them are still processed manually by human operators, the most common and labor-consuming operation being document amount reading and typing. A common way to automate this process is to replace the human operator with an off-line recognition system that is able to do the operator's job.

With respect to document processing, many systems have been developed for information extraction, numerical string recognition, word recognition, and signature verification. Comparatively, there is very limited published work on the processing of date information even though this is a necessity in some application environments, e.g., in some countries it is illegal to process post-dated cheques.

The main objective of this report is to present the state of the art on handwritten month recognition, which is a specific application of handwritten word recognition. Therefore, before discussing the specific application of month recognition, we present an overview of word recognition so that we can have a better understanding of the research problem. We conclude this report by outlining some of the next steps for the current project.

2 Handwritten Word Recognition

The majority of research in handwritten word recognition has integrated the lexicon as constraint to build lexicon-driven strategies in opposite to handwritten digit string recognition. The lexicon is a list of all valid words that are expected to be recognised by the system. There are no established definitions, however, the following terms are usually used [1]:

- small vocabulary - tens of words
- medium vocabulary - hundreds of words
- large vocabulary - thousands of words
- very large vocabulary - tens of thousands of words

The lexicon is a key point to the success of such recognition systems, because it is a source of linguistic knowledge that helps to disambiguate single characters by looking at the entire context [2]. As the number ofwords in the lexicon grows, the more difficult the recognition task becomes, because more similar words are more likely to be present in the lexicon. Koerich et al. [3] show (Figure 1) that, in general, the performance of the system decreases and the lexicon size gets bigger.

Another important issue on handwritten word recognition concerns the approaches for training and recognition, which can be classified into analytic and holistic. In an analytic approach, the segmentation of words into segments that relate to characters is required. Nevertheless, this is not a trivial task due to problems such as touching, overlapping, or broken characters. Moreover, this operation is made more difficult because of the ambiguity encountered in handwritten words. Therefore, most successful analytical methods employ segmentation-based recognition strategies where the segmentation can be explicit [4, 5, 6, 7] or implicit [8, 9, 10]

In an holistic approach, word recognition is performed considering the whole word. In such a case, there is no attempt to split the word image into segments. Still, it is possible that the image would be segmented in order to produce a sequence of observations [11, 12].

Concerning the classification models, the literature clearly shows that the HMM (Hidden Markov Model) is the most used technique for handwritten word recognition. A HMM is a statistical Markov model in which the system being modelled is assumed to be a Markov process with

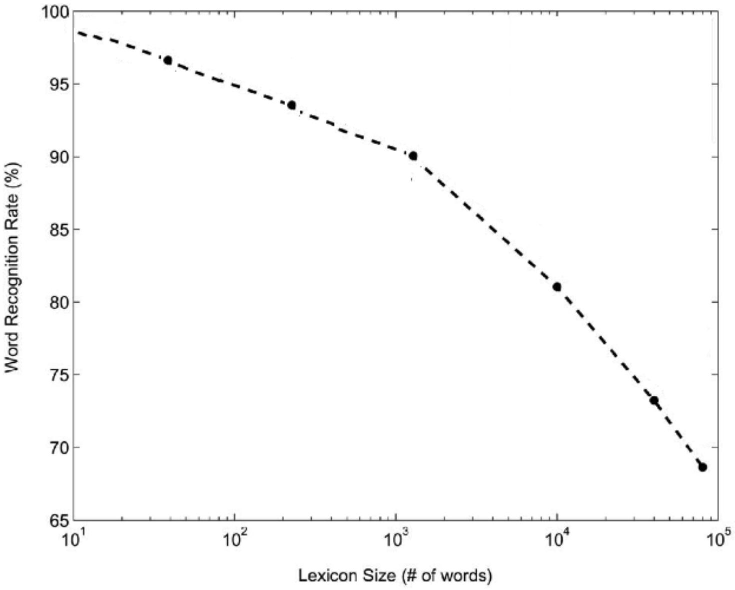

Figure 1: Relationship between performance and the lexicon size (adapted from [3])

unobserved (hidden) states. Even though the states are not visible in an HMM, the output that is dependent on a state is visible. Each state has a probability distribution over its possible output tokens. Therefore, the sequence of tokens generated by an HMM gives some information about the sequence of states [13, 14]. It can be found in several different flavours combining both implicit and explicit segmentation. In Table 1 we summarise some works reported in the literature that use HMM as base classifier to recognise handwritten words.

Table 1: Works using HMM to recognise handwritten words

Ref	Lexicon Size	Rec. Rate (%)	Test Set	Database
Freitas [15]	39	70	2380	Legal Amount
Mohamed [10]	100	89	317	City names
Kundu [5]	100	88	3,000	City names
Yacoubi [6]	100	96	4,313	City names
Bunke [16]	150	98	3,000	English words
Arika [4]	1,000	90	2,000	English words
Gimenez [17]	1,117	70	14,000	English words
Koerich [3]	80.000	68	4,674	French city names

A direct comparison is not possible since different lexicon sizes and databases have been considered in these works. What is interesting to notice, though, is that some applications with large lexicon achieve similar of applications with small lexicon, e.g., the first and last lines in Table 1. This can be explained in parts by the similarity of the words that belong to the lexicon and also by the quality of the handwriting.

In spite of the good performance of the HMM as word recogniser, we can find in the literature some attempts to increase the performance of HMM-based systems by implementing hybrid approaches such as HMM-SVM [18] and HMM-Statistical Language Models [19].

3 Month Recognition

Most of the works available in the literature on month recognition are related to Date Recognition on bank checks. Automatic reading of bank checks still is a important application since millions of bank checks issued from thousands of banks and financial institutions are daily used over the world for monetary transactions. Thus, 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 since most of the cheques are still processed manually by human operators. Usually, bank cheque processing is performed in big centers or at branch agencies which are equipped with fast scanners/sorters, archiving systems, and videocoding terminals for operators who make data entry. Operators look at cheque images one by one and enter the cheque amounts.

One may find in the literature dozens of works related to the processing of courtesy and legal amounts, however, the number of works dealing with date recognition is quite limited. A recent review on automatic processing of handwritten bank cheques by Jayadevan et al [20] highlights this difference showing that most of the works on months recognition deal with Brazilian and Canadian bank cheques. One explanation for the lack of works on date recognition may be found in [21], where the authors state that date processing of bank cheques is considered as the most difficult target in cheque processing because of the worst segmentation and recognition performance.

To the best of our knowledge, the first work for month recognition was proposed by Fan et al [22] in 1996. In this work the authors proposed a system based on HMM and neural networks to read the date field in Canadian bank cheques. They reported a very poor recognition rate at the time, around 22% of recognition rate. One justification for such a poor performance was the huge number of ways a date can be written on a Canadian bank cheque (Figure 2).

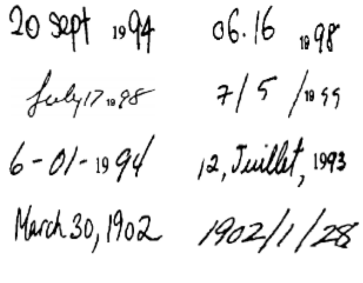

Figure 2: Variability of dates on Canadian bank cheques.

Xu et al [23] pursued the work of Fan et al. and proposed a more robust month word recogniser based on HMM and two different architectures of neural networks (Multi-Layer Perceptron - MLP). They tried different combination rules including Majority Voting, Sum, and Product to fuse the outputs of the different classifies and present a recognition rate of 85.3% on a database of 2,063 Canadian bank cheques. In [24], Xu et al. described a complete data recognition system using the month recogniser presented in [23].

Morita et al [25] presented a system to recognise month words on Brazilian cheques based on a combination of holistic and analytical approaches with a single explicit segmentation technique to provide a grapheme sequence for the HMM of each recogniser. In order to perform the segmentation, first the word was divided into zones (lower, middle and upper) and then segmentation points were detected based on contour information, as depicted in Figure 3. In this work they reached a recognition rate of 83.6% on 2,000 images of Brazilian cheques.

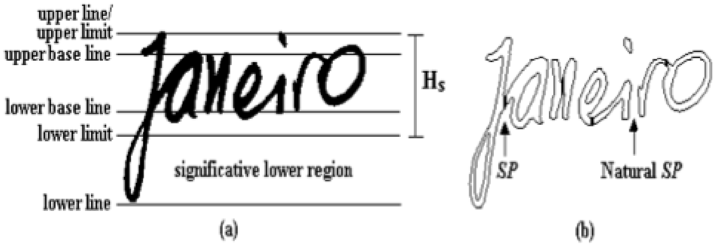

Figure 3: Segmentation used for feature extraction in [25].

Aiming at improving the results presented in [25], the authors [26, 27] introduce some modifications in their approach. First, in order to cope with the limited number of samples available for training, they added the legal amount database to increase the frequency of characters in training and validation sets. For this reason, in this work they have opted by the 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 represented by two feature sequences of equal length to feed the HMMs (Figure 4). Experiments show that this new approach reached an recognition rate of 91% on the same 2,000 images of Brazilian bank cheques.

Kapp et al [28] assessed two architectures of artificial neural networks for the classification of handwritten month words on Brazilian bank cheques. The performances of conventional and class-modular MLP architectures are compared. Using global features (holistic approach) like perceptual features and characteristics based on concavities/convexities, it has been found that the class-modular architecture is superior to the conventional MLP architecture. The authors reported

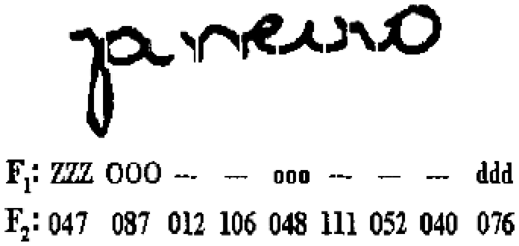
F1: ZZZ OOO -- — ooo -- — — ddd
F2: 047 087 012 106 048 111 052 040 076

Figure 4: Pair of feature sequences representing a word image [26]

an accuracy of 81.7% on a database composed of 6,000 month word images. Table 2 summarises the aforementioned works on month recognition.

Table 2: Summary of works on month recognition

Ref	Data set Size	Cheque Type	Techniques	Recognition Rate (%)
Fan [22]	4,565	Canadian	HMM,MLP	21.8
Xu [23]	2,063	Canadian	HMM,MLP	85.3
Morita [25]	2,000	Brazilian	HMM	83.6
Morita [26]	2,000	Brazilian	HMM	91.0
Kapp [28]	6,000	Brazilian	MLP	81.7

4 Conclusions and Further Developments

In this report we have reviewed the literature on handwritten word recognition with focus on month word recognition. As mentioned before, most of the works on month word recognition are presented in the context of date recognition on bank cheques. The recent review presented by Jayadevan et al in [20] shows that the amount of works on month recognition is quite scarce when compared with the works related to other parts of the bank cheque, such as legal amount, courtesy amount, and signature verification. It does not mean, though, that this is a less important application. On the contrary, it is very important in application environments were cheques cannot be processed prior to the dates shown.

In spite of the limited number of works, this review give us some insights about the further developments of this project. In a first moment we intend to implement the system described in Morita et al. in [26, 27], which includes the following steps:

- Preprocessing: Binarization, Slant Correction, Line Detection/Elimination, and Contour Smoothing.

- Word Segmentation into Graphemes: Segmentation step that provides a sequence of graphemes where each one consists of a correctly segmented, an under-segmented, or a over-segmented character.

- Feature Extraction: Two feature extractors will be implemented - Global Features and Concavity descriptors. It has been proved that both descriptors combined provide good results.
- Classification: The classification will be performed through HMMs. However, other classifiers will be assessed and combined with the HMMs so that we can benefit from the different classifier's architecture.

References

[1] A. Koerich, R. Sabourin, and C. Y. Suen, "Large vocabulary off-line handwriting recognition: A survey," *Pattern Analysis and Applications*, vol. 6, no. 2, pp. 97–127, 2003.

[2] F. Kimura, M Shridhar, and Z. Chen, "Improvements of a lexicon directed algorithm for recognition of unconstrained handwritten words," in *International Conference on Document Analysis and Recognition*, 1993, pp. 18–22.

[3] A. Koerich, R. Sabourin, and C. Y. Suen, "Recognition and verification of unconstrained handwritten words," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 10, pp. 1509–1522, 2005.

[4] A. Arika and F. T. Vural, "Optical character recognition for cursive handwriting," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 6, pp. 801–813, 2002.

[5] A. Kundu, Y. He, and M. Chen, "Alternatives to variable duration hmm in handwriting recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1275–1280, 2002.

[6] A. El Yacoubi, M. Gilloux, R. Sabourin, and C. Y. Suen, "An hmm-based approach for off-line unconstrained handwritten word modeling and recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 8, pp. 752–760, 1999.

[7] F. Grandidier, R. Sabourin, M. Gilloux, and C. Y. Suen, "An a priori indicator of the discrimination power of discrete hidden markov models," in *International Conference on Document Analysis and Recognition*, 2001, pp. 350–354.

[8] W. Cho, S. W. Lee, and J. H. Kim, "Modeling and recognition of cursive words with hidden markov models," *Pattern Recognition*, vol. 28, no. 12, pp. 1941–1953, 1995.

[9] A. M. Gillies, "Cursive word recognition using hidden markov models," in *Proc. Fifth U.S. Postal Service Advanced Technology Conference*, 1992, pp. 557–562.

[10] 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*, vol. 18, no. 5, pp. 548–554, 1996.

[11] K. Han and I. K. Sethi, "An off-line cursive handwritten word recognition system and its application to legal amount interpretation," in *International Journal of Pattern Recognition and Artificial Intelligence*, S.Impedovo et al, Ed., pp. 757–770. World Scientific, 1997.

[12] G. Dimauro, S. Impedovo, G. Pirlo, and A. Salzo, "Automatic bankcheck processing: A new engineered system," in *International Journal of Pattern Recognition and Artificial Intelligence*, S.Impedovo et al, Ed., pp. 467–503. World Scientific, 1997.

[13] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proc. of IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[14] F. Jelinek, *Statistical Methods for Speech Recognition*, MIT Press, 1997.

[15] C. Freitas, F. Bortolozzi, and R. Sabourin, "Handwritten isolated word recognition: An approach based on mutual information for feature set validation," in *International Conference on Document Analysis and Recognition*, 2001, pp. 665–669.

[16] H. Bunke, M. Roth, and E. G. Schukat-Talamazzini, "Off-line cursive handwriting recognition using hidden markov models," *Pattern Recognition*, vol. 28, no. 9, pp. 1399–1413, 1995.

[17] A. Gimenez and A. Juan, "Embedded bernoulli mixture hmms for handwritten word recognition," in *International Conference on Document Analysis and Recognition*, 2009, pp. 896–900.

[18] A. R. Ahmad, C. Viard-Gaudin, and M. Khalid, "Lexicon-based word recognition using support vector machine and hidden markov model," in *International Conference on Document Analysis and Recognition*, 2009, pp. 161–165.

[19] U. Marti and H. Bunke, "Using a statistical language model to improve the performance of an hmm-based cursive handwriting recognition system," *International Journal of Pattern Recognition and Artifical Intelligence*, vol. 15, no. 1, pp. 65–90, 2001.

[20] R. Jayadevan, S. R. Kolhe, P. M. Patil, and U. Pal, "Automatic processing of handwritten bank cheque images: a survey," *International Journal on Document Analysis and Recognition*, vol. 15, pp. 267–296, 2012.

[21] G. F. Houle, D. B. Aragon, R. W. Smith, M. Shridhar, and F. Kimura, "A multi-layered corroboration-based check reader," in *Workshop on Document Analysis Systems*, 1996, pp. 495–546.

[22] R. Fan, L. Lam, and C. Y. Suen, "Processing of date information on cheques," in *Proc. 5^{th} International Workshop on Frontiers of Handwriting Recognition (IWFHR)*, Colchester-U.K, September 1996, pp. 207–212.

[23] Q. Xu, J. H. Kim, L. Lam, and C. Y. Suen, "Recognition of handwritten month words on bank cheques," in *Proc. 8^{th} International Workshop on Frontiers of Handwriting Recognition (IWFHR)*, Niagara on the Lake-CA, August 2002, pp. 111–116.

[24] Q. Xu, J. H. Kim, L. Lam, and C. Y. Suen, "Automatic segmentation and recognition system for handwritten dates on canadian bank cheques," in *International Conference on Document Analysis and Recognition*, August 2003.

[25] M. Morita, E. Lethelier, A. El Yacoubi, F. Bortolozzi, and R. Sabourin, "An hmm-based approach for date recognition," in *Proc. Fourth IAPR International Workshop on Document Analysis Systems (DAS)*, Rio de Janeiro-Brazil, December 2000, pp. 233–244.

[26] M. Morita, A. El Yacoubi, R. Sabourin, F. Bortolozzi, and C. Y. Suen, "Handwritten month word recognition on Brazilian bank cheques," in *Proc. 6^{th} International Conference on Document Analysis and Recognition (ICDAR)*, Seattle-USA, September 2001, pp. 972–976.

[27] M. Morita, R. Sabourin, F. Bortolozzi, and C. Y. Suen, "Segmentation and recognition of handwritten dates: an hmm-mlp hybrid approach," *International Journal on Document Analysis and Recognition*, vol. 6, no. 4, pp. 248–262, 2004.

[28] M. Kapp, C. Freitas, and R. Sabourin, "Methodology for the design of nn-based month-word recognizers written on brazilian bank checks," *Image and Vision Computing*, vol. 25, pp. 40–49, 2007.