Handwritten Isolated Word Recognition: An Approach Based on Mutual Information for Feature Set Validation

Cinthia O. A. Freitas[1], Flávio Bortolozzi[1], Robert Sabourin[2]

1-Pontifícia Universidade Católica do Paraná (PUC-PR) Rua Imaculada Conceição, 1155 – Prado Velho – CEP:80215-901 - Curitiba (PR) – Brazil - cinthia, fborto@ppgia.pucpr.br
2-École de Technologie Supérieure (ETS) 1100, Rue Notre Dame - Ouest H3C 1K3 - Montreal (QC) – Canada - sabourin@livia.etsmtl.ca

Abstract

This paper presents the application of Mutual Information criterion to validate feature sets extracted from handwritten words in Brazilian legal amounts. The lexicon includes a subset of short words without ascenders /descenders and subsets of words with the same prefix or suffix. These particularities of the Brazilian lexicon show that is necessary to improve the perpetual feature set with complementary geometric features, and also modeling the prefix and suffix of the words. Finally, the experiments show the viability of our approach.

1. Introduction

The bank check recognition problem arouses great interest in researches, since there is a high level of ambiguity and complexity in such a kind of images, as seen in Figure 1, where the keywords are on evidence.

Figure 1 - Examples of writing styles in Brazilian handwritten bank check images

The interest is also explained by practical applications in the bank check compensation systems, since it is well known that the manual process demands both time and elevated cost, besides not being efficient in given situations.

From the literature two main approaches have been proposed to deal with this challenging problem, such as: *local or analytical* held at the character level [1,2] and *global* held at the word level [3,4]. Moreover, different feature selection and validation strategies can be found. In [5] the conditional perplexity based on the entropy notion from the information theory is used to indicate the discriminative power of different feature sets. In [6] the Mutual Information (MI) is applied to evaluate the information content of each feature and to select an informative subset of features to be used as input data for a neural network classifier. Other example is found in [7], where the MI contributes to handwritten word recognition of French legal amounts by improving the feature set. To this end, a concatenation algorithm selects a subset of relevant graphems from an initial set of available graphems.

The proposed system for handwritten legal amount recognition of Brazilian bank checks can be categorized as a *global approach* that avoids the explicit segmentation of words into letters or pseudo-letters by using word Hidden Markov Models (HMM). In this HMM framework, the MI is used to validate a feature set based on *perceptual features*.

This paper is divided into 8 sections. Section 2 summarizes the relevant concepts of the MI. Section 3 explains the handwritten word recognition problem taking into account the Brazilian legal amounts. Section 4 presents the perceptual feature extraction. Section 5 describes the recognition based on HMM. In Section 6 we present the experimental results. Section 7 shows the application of MI to validate the feature set, while Section 8 presents the conclusion and future works.

2. Background Theory

The entropy is a measure of uncertainty of a random variable defined in [8], as:

$$H(P) = -\sum_{i=1}^{N} P_i \log_2 (P_i) \qquad (1)$$

where $\boldsymbol{P_i}$ is the $\boldsymbol{i^{th}}$ word probability in the training database and $\boldsymbol{N}$ is the lexicon length .

The MI considers more than one random variable. It is a measure of the information content that one variable contains about another random variable. This means a reduction in the uncertainty of one random variable due to the knowledge of the other. So, the MI measures how the content of information is distributed in the feature set extracted from the word images. For this purpose, the Mutual Information I(X,Y), described in [8], is the relative entropy between the joint distribution and the product distribution $p(x)p(y)$:

$$I(X,Y) = \sum_{i=1}^{N}\sum_{j=1}^{M} P(x_i, y_j)\log_2\left(\frac{P(x_i, y_j)}{P(x_i)P(y_j)}\right) \quad (2)$$

In [7] the MI is expressed in the context of words, as:

$$I(M,G_K) = \sum_{i=1}^{N}\sum_{j=1}^{M} P(M = M_i, G_K = j)\log_2\left(\frac{P(M = M_i, G_K = j)}{P(M = M_i)P(G_K = j)}\right) \quad (3)$$

Where, the random variables ***M*** and $\boldsymbol{G_k}$ represent the words in the lexicon and the k^{th} "original" graphem, respectively. Moreover, ***j*** $\hat{I}$ ***{0,...,M}*** corresponds to the number of times the graphem ***k*** occurs inside the observation sequences of the word ***i***. A ¨original¨ graphem set is composed of all combinations of features extracted from the words in the training database.

Equation (3) permits to compute the information content of each "original" graphem. Normally, the number of "original" graphems is very high. Therefore, it is necessary to chose a method to look for likeness among the "original" graphems in order to concatenate them and validate the concatenations assumed. The idea is to reduce the number of graphems keeping the most discriminative of them, which contain the most significant part of the whole information associated. For this purpose, four methods can be used: Hamming Distance, Hamming Distance Weighed, Hierarchical [7,9] and Entropy [5,7]. In this work we consider the hierarchical approach and the MI is used to validate the perceptual feature set. The feature set validation is done by applying the MI associated with the α criterion as shown in [7]:

$$\frac{I(M,G^{"})}{\max(I(M,G_1),I(M,G_2))} > a \quad (4)$$

In other words, the validation occurs when the relation between the information of the concatenated graphem I(M,G") and the maximum of isolated graphems is greater than a fix threshold, α =1 [7]. In the sequence of this paper we describe the application of MI to validate the feature set and we discuss the difficulties and problems related to words in the Brazilian legal amounts.

3. Handwritten Word Recognition Problem

The legal amount corresponds to a numerical value that obeys a known grammar. The database comprises values between R$ 0,01 ("um centavo") and R$ 999.999,99 ("novecentos e noventa e nove mil, novecentos e noventa e nove reais e noventa e nove centavos").

From the numerical value it is possible to define five subset of words, such as: "entos", "enta", "ten", "unity", as shown in Figure 2, and the keywords "mil", "reais or real" and "centavos or centavo" (see Figure 1). We can also observe in Figure 2 the similarity among the suffix and prefix of the words in the lexicon. This fact increases the complexity of the recognition problem.

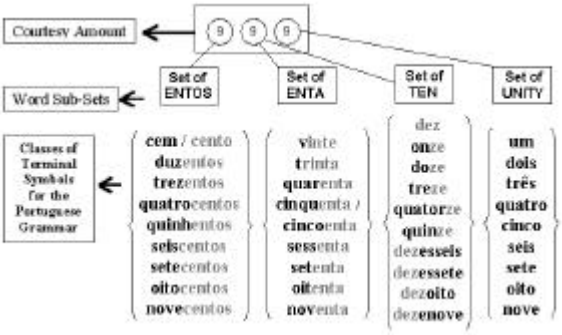

Figure 2 - Subset of words for Brazilian legal amounts

The *a priori* measure of the difficulty related to the recognition task can be obtained by the Entropy, defined in Equation (1). The lexicon in question is composed of 39 isolated words. The H(P) calculated using the training database is equal 4.77 bits. This value permits to compare the recognition of 39 words with a problem containing 27.28 equiprobable word classes. In this study, we have defined one model to represent each isolated word.

4. Feature Extraction Method

A preprocessing is used to minimize the effects of the writing variability related to the different styles, the writer's particular characteristics and the word slant [15]. It consists of slant correction [9] and smoothing [10] of the word images. No kind of baseline correction is used, since the legal amount is written between two printed guidelines in the regular check pattern.

The feature extraction plays an important role on handwriting recognition systems. Two questions are important in the feature selection: 1) *What are the relevant features (perceptual features) in the handwritten recognition process*? and 2) *How to represent cursive words without the presence of perceptual features*? To answer these questions we integrate the relevant aspects of the writing and reading processes as described in [11,12]. In [11], the authors define perceptual features as the most used characteristics in the word form representation (ascenders, descenders and loops, represented by symbol, position and size) [15].

The perceptual features occur in the great majority of the words in the lexicon. However, we can find subsets of words that do not present this kind of features, such as: words without ascenders/descenders ("um", "cinco", "seis", "nove", "cem", "reais"), and short words ("dois", "três", "quatro", "sete", "oito", "onze", "doze", "mil"). To deal with these subsets of words we propose a second and complementary set of features, which takes into account concavities and convexities presents in the word body.

Figure 3-a shows the feature set based on perceptual features that are extracted from the three word zones (ascender, body and descender), called Set 1 - PF. The character # denotes a separator between two graphemes.

These three zones are determined based on the horizontal transition histogram. The body of the word is the area located between ±70% from the maximum value of this histogram [16].

Figure 3 - Features sets: a) perceptual features (PF), b) perceptual features, concavities and convexities deficiencies (PFCCD) c) segmentation

The features are extracted over the word images and a pseudo-segmentation process is applied to obtain a sequence of corresponding observations, as seen in Figure 3-c. A segment is delimited between two black-white transitions over the maximum peak of the horizontal transition histogram (median line), and a corresponding symbol is designated to represent the extracted set of features, making up a graphem. Only the transitions that are not found inside the loops of the word body are considered. In case of no feature can be extracted in the analyzed segment, an empty symbol is emitted.

Since the proportion of the graphems with empty symbol in conformity to the perceptual features is about 67.3%, we can conclude that the representation of cursive words needed to be completed through the ligature between letters, in order to make up the graphems. In this manner, we established a classification capable of separating graphems made up of "C", "S", "E" and "Z" or, "u", "n", "r" and "i".

The second feature set, called PFCCD (Perceptual Features, Concavities and Convexities Deficiencies), is implemented for a better graphem discrimination. It takes into account the concave and convex deficiencies in addition to the perceptual features [13]. Concavities and convexities deficiencies in the word body are extracted and labeled, as shown in Figure 3-b. These deficiencies are obtained by labeling the background pixels of the input images [16].

The symbol alphabet was defined based on the hierarchical occurrences of the basic feature types, as well as, on the occurrences of the combination of these features in the same pseudo-segment (see Table 1). The entire and definitive hierarchical alphabet is composed of 29 different symbols.

Table 1 - Basic feature

Item	Feature	Symbol
01	Large and small ascender	T , t
02	Large and small descender	F , f
03	Superior and inferior loop	l , j
04	Large and small loop in word body	O, o
05	Open right and open left concave	(,)
06	Open right and open left convex	C , Z
07	Open down and open up convex	n , u
08	False loop in word body	a
09	Ligature down	i
10	Ligature up	r
11	Empty	X

5. Word Recognition Method

The Hidden Markov Model (HMM) theory has been successfully used to model the writing variability. The theoretic formulation of HMM is beyond the scope of this paper. An excellent introduction to this subject can be found in [14]. Our interest in the HMM lies in its ability to efficiently model different knowledge sources. It correctly integrates different modeling levels (morphological, lexical, syntactical), and also provides efficient algorithms to determine an optimum value for the model parameters.

Our word HMMs are based on a left-to-right discrete topology (*Bakis Topology*), where each state can skip at most one state. The lexicon size permits to consider one model for each class. The word models are independent of the handwriting style or the orthography for the same word, for example: 1 - "um" and "hum", 14 - "quatorze" and "catorze", 50 - "cinquenta" and "cincoenta". Both possibilities, in each case, are considered correct. In the current system, an unique model for each pair "reais and real" and "centavos and centavo" is used. The reason is that the words "real" and "centavo" are frequently not found in financial applications. They are just found in the following courtesy amounts: "R$ 1,00 - um real" and "R$ XX,01 - um centavo", respectively.

The model training is based on the Baum-Welch Algorithm and the Cross-Validation process [14]. The objective of the Cross-Validation process is to monitor the general outcome during the training process. It is done over two sets of data: training and validation. After each iteration of the Baum-Welch Algorithm on the training data, the likelihood of the validation data is computed using the Forward Algorithm [14]. During the experiments, the matching scores between each model and an unknown observation sequence is carried out using the Forward Algorithm. Moreover, we also evaluate the use of the known probability of each word in the training database during the recognition process. For this purpose, $k = argmax_i\ [pr(O / l_i) \cdot p(l_i)]$, where $\boldsymbol{k}$ is the index $\boldsymbol{i}$ that maximize the fuction ($i = 1,..,N_c$), $\boldsymbol{N_c}$ is the number of word classes in the lexicon (39 words), $\boldsymbol{O}$

is the word observation sequence and $p(l_i)$ is the relative frequency of word i in the training database.

6. Experimental Results

The database used in the present work is composed of 11,936 isolated words. This database is divided into 3 subsets, called: Training (60%), Validation (20%) and Testing (20%). The cursive represents the most frequent writing style, with 72% in the training database [15].

The results with 39 models, considering PF and PFCCD sets are shown in Table 2. A significant contribution to the recognition performance is observed by using the concavity and convexity deficiencies features of the PFCCD set. This shows as expected, a better word representation especially for the words with an absence of perceptual features. Through an error analysis, we observed specific problems with families "enta" and "entos", in the following words: "cinquenta/cincoenta", "sessenta", "setenta", "oitenta", "duzentos", "quatrocentos", "quinhentos", "novecentos" and "centavos".

Table 2 - Recognition Experiment

Lexicon = 39 words	Without pwc		With pwc	
	PF	PFCCD	PF	PFCCD
TOP1	**57.22**	67.66	61.52	**70.61**
TOP2	73.16	80.42	76.34	82.44
TOP3	80.32	86.65	83.24	88.08
TOP4	84.78	89.94	86.85	90.53
TOP5	88.10	92.21	90.27	92.84

pwc = probability of each word class

Comparing results is not easy since the works refer to different databases, so we have to view the comparison on that basis. It seems that at present our results are comparable to others, especially because our lexicon is rather than the others (English – 32 words or French - 25, 27 or 29 words with according to the different authors), i.e. 39 words. We remind the Figure 2 that shows the similarity among the suffix and prefix of the words in the Portuguese lexicon. Unfortunately, in the literature, only few studies report results on Portuguese lexicon. This lack of studies in the literature makes the comparison of results rather difficult.

Table 3 presents a comparison with others existing works. These studies consider global approach, one model to represent each isolated word at different lexicon.

Table 3 – Comparison of word recognition results (% correct in top 1 choice)

Authors	E	F	P
Côte [3]	73.6	-	-
Guillevic ADS [4]	72.6	-	-
Guillevic AD [4]	63.9	-	-
Avila [7]	-	62.2	-
Guillevic AD [4]	-	78.3	-
Ollivier [17]	-	75.0	-
Gomes [18]	-	-	50.0
Freitas PP [19]	-	-	57.2
Freitas PPCCD [19]	-	-	70.6

E = English, F = French, P = Portuguese

7. Feature Set Validation

The validation process based on MI starts by computing the $I(M,G_k)_{93}$ for the 93 "original" graphemes extracted from the training database, applying the Equation (3). Afterwards, we compute $I(M,G_k)_{29}$ for the 29 hierarchical graphems. Moreover, we compute the α values using Equation (4). Some examples have their results shown in Table 4, where the Hierarchical Grapheme "*Ot*" (presence of a loop in the word body + a small ascender) corresponds to the concatenation of four symbols "tO", "Ot", "ot" and "to" graphemes into one symbol. The resulting symbol "*Ot*" does not make any difference if the ascender comes before or after the loop. In this manner, we concluded that the 29 Hierarchical Graphems are validated based on a significant increase of mutual information $I(M,G_k)_{29}$ compared to the "original" graphems.

Table 4 – Examples of concatenated graphem validated with Equation (4)

Original Graphems	max $I(M,G_k)$	Hierarchical Graphem	$I(M,G_k)$	a value
Ot,tO,ot,to	0.4102425	*Ot*	4.198101	10.23
OF,FO,oF, Fo	0.4101486	*OF*	4.207193	10.26
JO,Oj,oj,jo	0.4104498	*Oj*	4.206434	10.25

8. Conclusion and Future Works

This paper presents the isolated word recognition in the context of Brazilian legal amounts. It shows the similarity among the words in the lexicon by considering that the handwriting of the legal amount is based on a numerical value. The MI and α criterion were applied for hierarchical feature set validation. The advantage is that a fast informative feedback about the information contained in each graphem is provided.

The obtained results motivate the continuity of the system development considering a new complementary geometric feature set. Our future work consists of

looking for other discriminative features and also improving word modeling in order to optimize the word recognition phase. Moreover, the studies include the suffix and prefix modeling.

References

[1] Kim, G. Recognition of offline handwritten words and its extension to phrase recognition, PhD Thesis, University of New York at Buffalo, USA, 1996.

[2] Lecolinet, E. Segmentation d'images de mots manuscrits: application à la lecture de chaînes de caractères majuscules alphanumériques et à la lecture de l`écriture cursive, Ph.D. thesis, Universite Pierre et Marie Curie (Paris VI), France, 1990.

[3] Côte, M. Utilisation d'un modèle d'accès lexical et de concepts perceptifs pour la reconnaissance d'images de mots cursifs, Ph.D. thesist, École Nationale Supérieure des Télécom., France,1997.

[4] Guillevic, D. Unconstrained handwriting recognition applied to the processing of bank cheques. Ph.D. Thesis, Department of Computer Science at Concordia University, Canada, 1995.

[5] Grandidier, F., Sabourin, R., Suen, C.Y., Gilloux, M. Une nouvelle stratégie pour l´amélioration des jeux de primitives d´un système de reconnaissance de l´écriture. CIFED2000.

[6] Battiti, R. Using Mutual Information for Selecting Features in Supervised Neural Net Learning. IEEE Transactions on Neural Networks, Vol.5, no.4,1994,537--550.

[7] Avila, M. Optimisation de Modeles Markoviens pour la Reconnaissance de L'ecrit. Thèse de doctorat, Université de Rouen, France, 1994.

[8] Cover, T.M., Thomas, J.A. Elements of Information Theory. Wiley Series in Telecommunications, 1991.

[9] El Yacoubi, A., Gilloux, M., Sabourin, R. and Suen, C.Y. Unconstrained handwritten word recognition using hidden markov models. IEEE Transaction on Pattern Analysis and Machine Intelligence, Vol.2, no.8,1999,752--760.

[10] Strathy, N.W. A method for segmentation of touching handwritten numerals. Master's thesis, Concordia University, Montreal-Canada, 1993.

[11] Madhvanath, S., Govindaraju, V. Perceptual features for off-line handwritten word recognition: a framework for heuristc prediction, representation and matching. Advances in Pattern Recognition, Sidney, Australia, 1998.

[12] Schomaker, L. and Segers, E. A method for the determination of features used in human reading od cursive handwriting. IWFHR, Korea, 1998, 157--168.

[13] Parker, J.R. Algorithms for Image. Processing and Computer Vision. Ed. John Wiley & Sons, Inc. 1997, 310--315.

[14] Rabiner, L., Juang, B.H. Fundamentals of speech recognition. Prentice Hall Inc., 1993.

[15] Freitas, C. O. A., El Yacoubi, A., Bortolozzi, F., Sabourin, R. Isolated word recognition in brazilian bank check legal amounts. In Proc. of the 4º Workshop on Document Analysis and Systems-DAS2000, Rio de Janeiro, Brazil, 10-13 december 2000, pp 279-290.

[16] Freitas, C. O. A., El Yacoubi, A., Bortolozzi, F., Sabourin, R. Brazilian bank check handwritten legal amount recognition. In Proc. of the XIII Brazilian Symposium on Computer Graphics and Image Processing - SIBGRAPI2000, Gramado, Brazil, 17-20 october 2000, pp 97-104.

[17] Ollivier, D. *Une appproche économisant les traitements pour reconnâitre l'écriture manuscrite: application à la reconnaissance des montants littéraux de chéques bancaires.* Thèse de doctorat, Université de Paris XI Orsay, France, 1999.

[18] Gomes, N. R. *Reconhecimento de Palavras Manuscritas Baseado em HMM e no Emprego de Características Topológicas e Geométricas*. PhD Thesis, UNICAMP, Brazil, 2000.

[19] Freitas, C. O. A. *Hidden Markov Models to Handwritten Word Recognition (Uso de Modelos Escondidos de Markov para Reconhecimento de Palavras Manuscritas).* PhD Thesis, PUCPR, Brazil. 2001.