A Strategy for Selecting Classes of Symbols from Classes of Graphemes in HMM-Based Handwritten Word Recognition

Cinthia O. A. Freitas[1], Flávio Bortolozzi[2], Robert Sabourin[3]

[1,2]Pontifícia Universidade Católica do Paraná, Programa de Pós-Graduação em Informática Aplicada, Rua Imaculada Conceição 1155, Curitiba, PR, Brasil
cinthia.freitas@pucpr.br , fborto@ppgia.pucpr.br
[3]École de Technologie Supérieure (ETS) 1100, Rue Notre Dame, Ouest, H3C 1K3, Montreal (QC), Canada
robert.sabourin@.etsmtl.ca

Resumo – Este artigo descreve uma metodologia para seleção de classes de símbolos a partir de classes de grafemas em um sistema de reconhecimento de palavras manuscritas do extenso de cheques bancários brasileiros baseado em HMM (*Hidden Markov Models*). Este artigo discute as definições de primitivas, grafemas e símbolos considerando um enfoque Global para o reconhecimento das palavras, o qual evita a segmentação das palavras em letras ou pseudo-letras utilizando HMM. Assim, a entrada para os modelos consiste em uma descrição da palavra a partir de um alfabeto de símbolos gerados a partir dos grafemas extraídos das imagens das palavras, sendo esta a representação visível para o HMM. Portanto, a idéia é introduzir uma conceituação de alto nível, tais como primitivas perceptivas (laços, ascendentes, descendentes, concavidades e convexidades) e fornecer um modo de retro-alimentação rápido e informativo sobre a informação contida em cada classe de grafema, permitindo uma seleção de classes de símbolos. O artigo apresenta o algoritmo com base na Informação Mútua (*Mutual Information*) e HMM, ambos trabalhando em um mesmo processo de avaliação. Os resultados experimentais demonstram que é possível selecionar a partir de um conjunto "original" de grafemas (composto por 94 grafemas) um alfabeto de símbolos (composto por 29 símbolos). O artigo conclui que o poder discriminante dos grafemas é muito importante para a consolidação de um alfabeto de símbolos.

Palavras-chave: Primitivas, Informação Mútua, HMM, Reconhecimento de Palavras.

Abstract - This paper presents a new strategy for selecting classes of symbols from classes of graphemes in HMM-based handwritten word recognition from Brazilian legal amounts. This paper discusses features, graphemes and symbols, as our baseline system is based on a global approach in which the explicit segmentation of words into letters or pseudo-letters is avoided and HMM models are used. For this framework, the input data are the symbols of an alphabet based on graphemes extracted from the word images visible on the Hidden Markov Model. The idea is to introduce high-level concepts, such as perceptual features (loops, ascenders, descenders, concavities and convexities) and to provide fast and informative feedback about the information contained in each class of grapheme for symbol class selection. The paper presents an algorithm based on Mutual Information and HMM working in the same evaluation process. Finally, the experimental results demonstrate that it is possible to select from the "original" grapheme set (composed of 94 graphemes) an alphabet of symbols (composed of 29 symbols). We conclude that the discriminating power of the grapheme is very important for consolidating an alphabet of symbols.

Key-words: Features, Mutual Information, HMM, Handwritten Word Recognition.

Introduction

Following the traditional pattern recognition approach, we divide the recognition task into two steps: first, a set of features is extracted from the images, and then a classifier computes the class-conditional probabilities based on these extracted features. So, the objective of the feature extraction is to capture the most relevant and discriminatory characteristics of the object to be recognized. Accordingly, in our baseline system, the feature extraction algorithms use perceptual pattern recognition techniques, while the classification is based on a statistical approach.

In order to select a subset of the original features by reducing irrelevant and redundant ones, feature selection algorithms have been applied [9,12,15]. In the absence of such algorithms, large training sets are mandatory. Therefore, the problem of finding relevant features from a given feature space is defined as “Feature Selection”.

Feature selection methods attempt to find reduced feature sets, which minimize the probability of error. Most of these methods use evaluation functions and search algorithms to achieve their objective. The evaluation functions measure how good a specific subset is in discriminating between classes, and can be divided in three kinds of feature relevance being assumed [15]:

- Relevance of feature to constructing consistent hypothesis;
- Relevance of feature to improving accuracy;
- Relevance of feature to the concept.

Based on these three relevance metrics the feature selection algorithm can be grouped as shown in Figure 1 [15]. The sub trees are not mutually exclusive. This means that a feature selection algorithm can be a member of more than one sub tree.

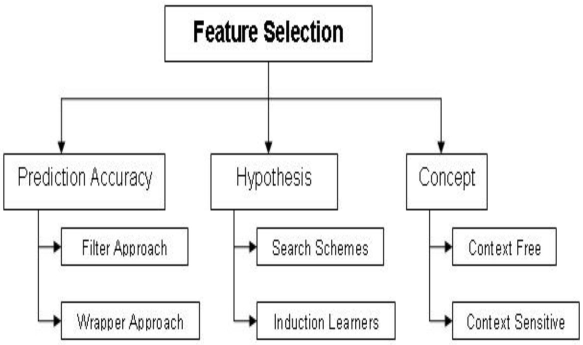

Figure 1 – Feature Selection Algorithms Classification

Filters measure the relevance of feature subsets independently of the classifier, whereas wrappers use the classifier’s performance as the evaluation function [9.15]. Search algorithms, by contrast, are responsible for driving feature selection using a specific strategy, e.g. branch-and-bound, stepwise and genetic algorithm, among others [12].

Usually, the dimension of the resulting feature vector is smaller than the dimension of the original pixel images, facilitating the subsequent classification step. However, it is difficult to introduce high-level concepts, such as loops and strokes, in a robust way. In fact, for the limit of an infinite number of training samples, it can be shown that the best possible performance can be obtained from an unbiased classifier.

In order to introduce high-level concepts for selecting classes of symbols; features, graphemes and symbols are discussed in this paper, since our baseline system is based on a global approach in which the explicit segmentation of words into letters or pseudo-letters is avoided and in which Hidden Markov Models (HMM) are used. For this framework, the input data are the symbols of the alphabet based on graphemes extracted from the word images visible on the HMM. Our symbol approach is based on the concept of Mutual Information (MI), such as in [1,6], but using graphemes to establish the classes of symbols.

For our study, the grapheme notion is directly linked with the feature set extraction method. The perceptual feature set used in this work takes into account the global approach and studies on the human reading-writing process as described in [10,11,16,18]. As a result, we introduce high-level concepts, such as perceptual features (loops, ascenders, descenders, concavities and convexities) and graphemes for selecting classes of symbols. The following questions are discussed in connection with the selection of an alphabet of symbols based on MI:

- What are a perceptual feature, a grapheme and a symbol?
- How can the most discriminatory graphemes be kept, considering a shape-space (S_S) formed by features and graphemes?
- How should an observation-space (S_O) based on shape-space (S_S) be defined?

This paper is divided into 9 sections. In section 2, the handwritten word recognition problem is explained in terms of Brazilian legal amounts. In section 3, the perceptual feature definition and extraction method is presented. In sections 4 and 5, we explain what a grapheme and a symbol are in the context of our study. Section 6 sets out the background of the theory of entropy and of MI. In section 7, the application of MI to the selection of an alphabet of symbols based on classes of graphemes is shown. In section 8, the experimental results are presented; while in section 9 the conclusion and suggestions for future work are presented.

Word Recognition Problem

The scope of this study is limited to the off-line recognition of individual handwritten words from legal amounts. The legal amount corresponds to a numerical value which obeys a known grammar, and the database contains legal amounts 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 subsets of words, such as: “entos”, “enta”, “ten” and “unity”, as shown in Figure 1, and the keywords {“mil”, “reais or real” and “centavos or centavo”} [5]. We can also observe in Figure 2 the similarities among the suffixes and prefixes of the words in the lexicon. This increases the complexity of the recognition problem.

A global approach in the context of our baseline system is one in which the explicit segmentation of words into letters or pseudo-letters is avoided and in which HMMs are used. In this framework, the input data are the symbols of the alphabet based on graphemes which are visible on the HMM.

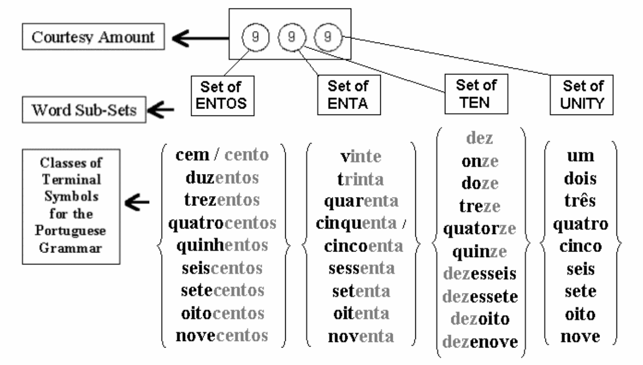

Figure 2 - Subset of words from Brazilian legal amounts

What is a Perceptual Feature?

Feature extraction plays an important role in handwriting recognition systems, as described in [3,7,11,13]. We integrate the relevant aspects of the writing and reading processes, as described in [10,16]. In [10], the authors define perceptual features as the most commonly used characteristics in word form representation (ascenders, descenders and loops, represented by symbol, position and size). In [16], the authors summarize a number of findings in human reading of handwriting. Results reveal a left-to-right strategy in reading; however, extra attention is paid to the initial, left-most parts and the final right-most parts of words in a range of word lengths. This means that shape information (ascenders, descenders, crossings and points of high curvature) is important in handwriting recognition.

In order to minimize the effects of writing variability related to differing styles, to the writer's own particular characteristics and to word slant, a preprocessing treatment is applied [5]. This consists of slant correction as in [19], and smoothing of word images as in [17]. No baseline correction of any kind is used, since a legal amount is written using two printed guidelines in the regular, bank-check pattern.

Three zones are determined based on the horizontal transition histogram: ascender, body and descender. The body of the word is the area located between ±70% of the maximum value of this histogram [5]. Figure 3a shows a feature extracted based on perceptual features extracted from these word zones, called PFCCD (Perceptual Features, Concavities and Convexities Deficiencies). The character # denotes a separator between two graphemes.

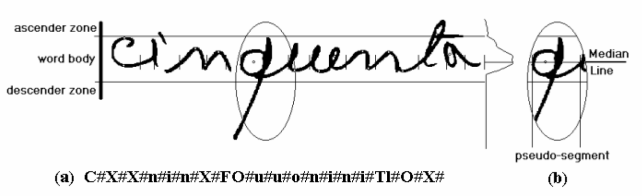

Figure 3 - Feature sets: a) PFCCD; and b) pseudo-segment

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 3b. A segment is delimited between two black-white transitions over the maximum peak of the horizontal transition histogram (Median Line), and a corresponding representation (called *F-symbol*) is designated to represent the features extracted, making up a grapheme. Only the transitions that are not found inside the loops of the word body are considered. In a case where no feature can be extracted in the analyzed segment, an empty symbol is emitted, denoted by *X*.

Concavity and convexity deficiencies in the word body are extracted and labeled, as shown in Figure 3a. These deficiencies are obtained by labeling the background pixels of the input images [5]. So, the PFCCD feature set is a classification capable of representing the ligature between letters and separating graphemes made up of "C", "S", "E" and "Z" or, "u", "n", "r" and "i". Table 1 summarizes the feature set and the corresponding *F-symbol* for each one.

Table 1 - Feature Set

Item	Basic Feature	F-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

What is a Grapheme?

In our study, a grapheme is an entity which can correspond to a part of a letter, a letter or connected letters. The feature set is extracted inside a pseudo-segment from the word obtaining the “original” grapheme set. An “original” grapheme set is composed of all combinations of features extracted from the words in the training database.

Consequently, the word will have a number of graphemes corresponding to the number of pseudo-segments observed over the Median Line, as described previously. Now, suppose that a vector with ***x*** binary elements can represent the basic features, where each element represents the absence (0) or the presence (1) of each feature, as shown in Figure 4. This representation describes the grapheme *FO* extracted from the word “cin**q**uenta”, as shown in Figure 3.

T	t	F	f	l	j	O	o	(	)	C	Z	n	u	a	i	r	X
0	0	**1**	0	0	0	**1**	0	0	0	0	0	0	0	0	0	0	0

Figure 4 - Feature vector

With this description, 2^x = 262,144 different graphemes in the PPCCD set can be codified, where x = 18. In reality, we found 94 different graphemes, called “original” graphemes, from a total of 73,165 graphemes extracted from 7,146 word images in the training database. Figure 5 shows that some graphemes are more common than others, such as: *X, i, u, r, o, O, a, n* and *T*. By contrast, many of them occur only a few times, such as: *flO*, *TFO*, *na* and *Er*.

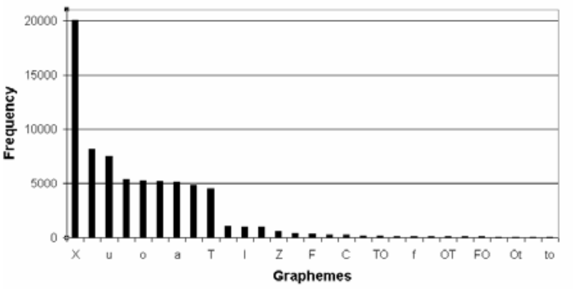

Figure 5 - Appearance frequency of “original” graphemes

In fact, because of the variability of the handwritten word and to reduce the dimensions of the HMM model, we need to improve the robustness of the system. The problem when working with HMM is to answer the following question: *How it is possible to train HMM models using graphemes which have a low apparition frequency in the observation sequences?* It is important not to reduce the quality of the grapheme. So, definition of the *observation-space* is an important step of the symbol selection process.

This improvement is related to an analysis based on the “original” grapheme set. Therefore, we implement a methodology for selecting classes of symbols from classes of graphemes using MI theory and grapheme similarity. This method suppresses the graphemes that have low apparition frequency. It should be remembered that suppressing graphemes does not mean eliminating them, but rather to concatenate them with other, similar graphemes.

What is a Symbol?

A symbol corresponds to the physical input of the system being modeling. We denote the individual symbols as V = {v1, v2,…, v_M}, where M is the number of distinct observation symbols, i.e., the discrete alphabet size.

Considering the feature vector as shown in Figure 4 and knowing that we can extract more than one basic feature inside the pseudo-segment making up a grapheme; we need define the discrete alphabet V for our system HMM-based. For this, we analyze the interaction among three different spaces: feature, grapheme and symbol.

The features and graphemes extracted define a space which we call the *shape-space*. This space represents the shape to be recognized and is linked to letter shape. Figure 6 presents the graphic representation of the *shape-space* (S_S) and can be written as:

$$S_S = f\,(F,G) \qquad (1)$$

where F are the basic features (see Table 1) and G is the grapheme set extracted from the training database.

It is easy for us, as human beings, to relate the basic feature and the grapheme. For instance, the basic features O and F can be extracted from the same pseudo-segment making up the grapheme “*OF*”. This grapheme represents the letter “*q*” in the shape-space (with a big loop and big descender). The same analogy can be made using the features *o* and *F* (letter “*q*”, with a small loop and big descender) and features *O* and *T* (letter “*d*”, with a big loop and big ascender) (see Figure 5).

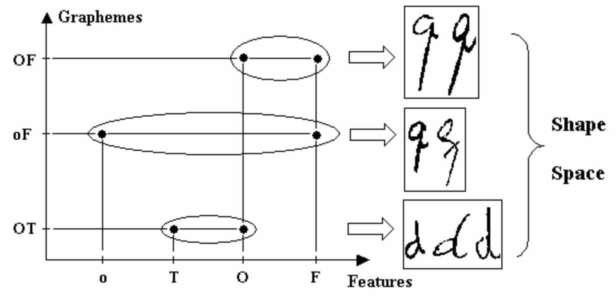

Figure 6 - Shape space: features versus graphemes

As stated above, we need to define the discrete alphabet V for our HMM-based system. Therefore, we need convert the *shape-space* into another space, which we call the *observation-space* (S_O). This space defines the alphabet of symbols V and M which is the number of distinct observation symbols. Figure 7 presents the graphic representation of the *observation-space*, where the graphemes *OF* and *oF* were concatenated to make up a unique symbol v_i.

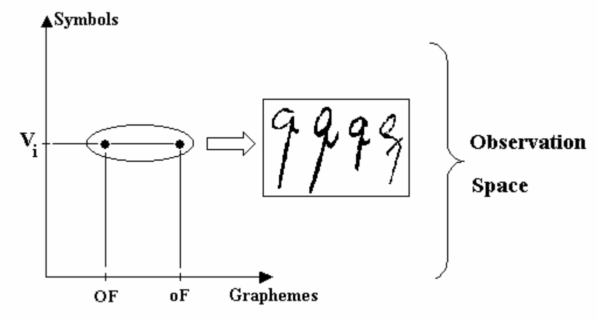

Figure 7 - Observation space: graphemes versus symbols

For this, we need to choose a method capable of computing the following function:

$$S_O = g\ (S_S\ ,V) \qquad (2)$$

where S_S is the *shape space*, $V = \{v_1 , v_2 ,…, v_M\}$, i =1, 2, …,M and $v_i \in V$. However, to compute S_O is not a trivial task, especially when many graphemes are involved. The idea is to keep a relevant and discriminatory grapheme set. In fact, we need to choose a subset of the "original" graphemes by removing irrelevant and redundant ones.

In the next section, MI theory is applied to studying the "original" grapheme set and implementing an algorithm based on the discriminatory power of the graphemes through their similarities.

Mutual Information Theory

The *a priori* measure of the difficulty of the recognition task can be obtained by a measure of entropy, which is a measure of the uncertainty of a random variable, defined in [4] as:

$$H(P) = -\textstyle\sum_{i=1}^{N} P_i \log_2 (P_i) \qquad (3)$$

where P_i is the ith word probability in the training database and N is the lexicon length. So, the lexicon in question is composed of 39 isolated words. The *H(P)* calculated using the training database is equal to 4.77 bits. With this value, it is possible to compare the recognition of 39 words and a problem containing 27 equally probable word classes. When the lexicon is smaller, such as a French lexicon, the problem complexity is around 12 equally probable word classes, as demonstrated in [1].

Given random variables X and Y, the MI Theory measures the amount of information in X that can be predicted when Y is known. Uncertainty means a reduction in one random variable due to knowledge about the other. So, MI measures how the amount of information is distributed in the "original" grapheme set. For this purpose, the *I(X,Y)*, described in [4], is the relative entropy between the joint distribution and the product distribution *p(x)p(y)*:

$$I(X,Y) = \sum_{i=1}^{N}\sum_{j=1}^{M} P(x_i, y_j)\log_2(\frac{P(x_i, y_j)}{P(x_i)P(y_j)}) \qquad (4)$$

In [1], MI is expressed in terms of words, as:

$$I(C,G_K) = \sum_{i=1}^{W}\sum_{j=1}^{X} P(C = C_i, G_K = j)\log_2(\frac{P(C = C_i, G_K = j)}{P(C = C_i)P(G_K = j)}) \qquad (5)$$

where the random variables C_i and G_k represent the i^{th} word class in the lexicon and the k^{th} "original" grapheme respectively. Moreover, $j \in \{0,...,X\}$ corresponds to the number of times the grapheme G_k occurs inside the observation sequences from word class i. W is the lexicon length corresponding to 39 classes of words.

Equation (5) enables computation of the amount of information in each "original" grapheme. Normally, the number of "original" graphemes is very high. Therefore, it is necessary to choose a method to look for similarities among these "original" graphemes of the set (S_S) in order to concatenate them and validate the concatenations obtained. The idea is to reduce the number of graphemes while keeping the most discriminating of them, i.e. those which contain the most significant part of the totality of associated information. The result of this process will provide the classes of symbols (S_O).

Four methods can be used to achieve this: Hamming Distance, Weighted Hamming Distance, Hierarchical [1] and Entropy [1,2,6]. In [1], the author did a hierarchical analysis based on the grapheme/letter shapes. In [6], the conditional perplexity based on the entropy notion from the information theory is used to indicate the discriminating power of different feature sets. In [2], MI is applied to evaluate the information contained in each feature and to select an informative subset of features to be used as input data for a neural network classifier. Another example is found in [1], where MI contributes to handwritten word recognition of French legal amounts by improving the feature set. For this purpose, a concatenation algorithm selects a subset of relevant graphemes from the "original" set.

In the next section, we discuss the MI criterion based on the amount of information contained in each extracted grapheme for selecting an informative alphabet of symbols to be used as input data for HMM.

Methodology for Selecting Classes of Symbols from Classes of Graphemes: Entropy Alphabet

We analyze the *F-symbols*, and the most frequent combinations of them, and place them in a set of “original” graphemes, establishing 5 *similarity-classes* based on letter shape (S_S) and their graphemes, as follows:

- **Class 01:** perceptual features (*O,T,F,X*) – these graphemes cannot be concatenated with others because they represent the most frequent *F-symbols* and have the most discriminatory power among graphemes;
- **Class 02:** concavity (*i,u*), convexity (*r,n*) and loop (*o,a*) – the graphemes making up this class can be concatenated with others;
- **Class 03:** small ascenders and descenders (*t,f*) without or with a loop (*l,j*) – the graphemes making up this class can be concatenated with others;
- **Class 04:** concavity (*C,E,S,(*) and convexity (*),Z*) – the graphemes making up this class can be concatenated with others;
- **Class 05:** graphemes composed of 3 *F-symbols* (one per zone) (for example: *TOF, Tj, Tf*) – these graphemes can be concatenated preferentially.

In order to improve the recognition results, an algorithm was implemented to provide fast informative feedback on the information contained in each grapheme. Then, a decision is made as to whether the grapheme must be kept as is or concatenated, within the *similarity-class*, with other graphemes and, if so, which one.

The algorithm implemented considers the MI (Equation 5) associated with the α *criterion* as shown in [1]:

$$\frac{I(C,G^{''})}{\max(I(C,G_1),I(C,G_2))} > \alpha \qquad (6)$$

Grapheme validation occurs when the relation between the information on the concatenated grapheme *I(C,G”)* and the $\max(I(C,G_1),I(C,G_2))$ from isolated graphemes is greater than a fixed threshold, $\alpha = 1$ [1]. Other values for α were tried, but, when $\alpha < 1$, more graphemes were concatenated, which was not satisfactory (low recognition rate). By contrast, when $\alpha > 1$, some graphemes are not concatenated, this yields an even worse result, because in this case the training HMM is prejudiced. Appling, for instance, $\alpha = 1.5$ the concatenation process obtained two different symbols: *Ot*,*tO* and *ot*,*to*. But, the occurrence of these symbols in the database is not appropriately for HMM training.

Figure 8 presents the algorithm implemented relating to MI, *similarity-class*, α *criterion* and HMM training models. Running this process, we can establish the number of symbols, M = 29, from the 94 “original” graphemes. We call this alphabet of symbols *V* the Entropy Alphabet.

```
MI Algorithm
1-Compute the $I(C,G_k)$ of each grapheme (starting with the “original” grapheme set)
2-Reduce the alphabet dimension, making $N$ concatenations (go to Reducing Stage)
3-Train HMM models based on the new alphabet
4-Recognition of the testing database with trained HMM models
5-If the recognition rate (%) is worse than before, return to step 2 and try the other possible concatenations, otherwise replace the alphabet
9-Return to step 1 while the recognition improvement (%) remains positive

Reducing Stage
2-Repeat X times with all combinations of pairs of graphemes (taking into account Classes 02,...,05):
    2.1-Choose two graphemes with low appearance frequency within the same Class (02,...,05)
    2.2-Compute the $I(C,G_k^{''})$ (graphemes concatenated)
    2.3-Validate the acceptable concatenations (apply the $\alpha$ criterion)
    2.4-Reduce the alphabet taking into account the validated concatenations
```

Figure 8 - MI algorithm

In this way, we improve the discriminating power of the feature set, and we improve system performance. Note that the process guarantees that any loss which occurs is related to grapheme quality. This is because the process does not avoid the information contained in the grapheme set. An important role was played by the *similarity-class*, which provides a search algorithm within the “original” grapheme set. Table 2 presents some samples of the normalized $I(C,G_k)$ for the “original” graphemes (94 symbols) and symbols (29 symbols). Observe that $I_{entropy}(C,G''_k)$ is higher than $I_{original}(C,G_K)$.

The MI process started by computing the $I(C,G_k)_{94}$ based on an “original” grapheme set (94 graphemes), applying Equation (5). We then applied the MI algorithm, computing $I(C,G_k^{''})_X$ for each *X* possible grapheme concatenation based on *similarity-class*. Moreover, we compute the α values using Equation (6). The results of some examples are shown in Table 3. For instance, the “original” grapheme *Ot* was concatenated with three other graphemes, *to*, *tO*, and *ot*, making up a new symbol $(G''_k)_{entropy}$. It makes no difference in the resulting symbol whether the ascender comes before or after the loop, because, for our lexicon, we do not differentiate between “d” and “b”. This is because we have no words containing the letter “b” (see Figure 2). In the sequence, we describe the word recognition method (HMM) and the experimental results.

Table 2 - MI *I(C,G$_k$)* for “original” and *I(C,G”$_k$)* for entropy alphabets (5 most common graphemes)

G_k (original)	$I_{original}(C,G_k)$	$I_{entropy}(C,G''_k)$
X	$2.66785\cdot10^{-04}$	$7.87530\cdot10^{-04}$
i	$4.35851\cdot10^{-03}$	$1.38081\cdot10^{-02}$
u	$8.80659\cdot10^{-03}$	$2.79295\cdot10^{-02}$
r	$1.20051\cdot10^{-02}$	$3.80493\cdot10^{-02}$
o	$1.31538\cdot10^{-02}$	$4.13278\cdot10^{-02}$

Table 3 - Grapheme concatenation

G_k(original)	max $I(C,G_k)$	G''_k	$I(C,G''_k)$	α
Ot,tO,ot,to	$1.135\cdot10^{-02}$	*Ot*	$3.57\cdot10^{-02}$	10.23
OF,FO,oF,Fo	$1.139\cdot10^{-02}$	*OF*	$3.58\cdot10^{-02}$	10.26
jO,Oj,oj,jo	$1.135\cdot10^{-02}$	*Oj*	$3.59\cdot10^{-02}$	10.25

MI Methodology Applied to Handwritten Word Recognition: Experimental Results

Hidden Markov Model (HMM) theory has been used successfully to model writing variability; however, the theoretical formulation of HMM is beyond the scope of this paper. An excellent introduction to this subject can be found in [14]. 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 HMM word models are based on a left-to-right discrete topology (*Bakis Topology*), where each state can skip at most two states. The size of the lexicon makes it possible to consider one model for each class, as explained in [5].

The database used here is composed of 11,936 isolated words. This database was divided into 3 subsets, called the Training (60%), Validation (20%) and Testing (20%) subsets. The most common writing style is cursive, representing 72% of the training database [5].

The results with 39 models, considering PFCCD sets and the different symbol alphabets obtained when we apply the MI algorithm (Figure 8), are shown in Table 4. Figure 9 shows some examples of misrecognized images.

Comparing results is not easy, since the different works refer to different databases, so the comparison has to be viewed on that basis. It seems that at present our results are comparable to those of others, in particular because our lexicon is rather longer than the others, i.e. 39 words (English - 32 words and French - 25, 27 or 29 words, depending on the authors). We must recall Figure 2, which shows the similarity between the suffixes and prefixes of the words in the Portuguese lexicon, thereby increasing the complexity of the recognition task. Unfortunately, in the literature, only a few studies report results on the Portuguese lexicon. This limited literature makes the comparison of results rather difficult. In Table 5, a comparison with other published work is presented. These studies consider a global approach, and one model to represent each individual word.

Table 4 - MI algorithm application and recognition results – PFCCD feature set

Number of Symbols (M)	Recognition Rate (%)	Comments
94	**67.13**	**Original Alphabet**
62	67.55	---
32	67.54	---
30	67.49	---
29	**67.66**	**Entropy Alphabet**

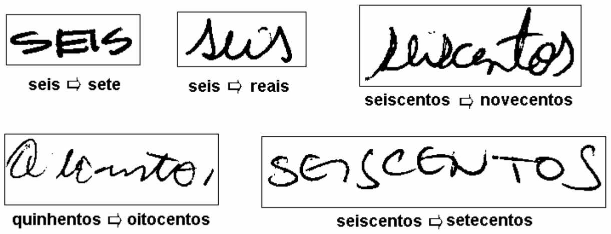

Figure 9 - Examples of misrecognized images

Table 5 - Comparison of word recognition results (Recognition rate (%) in TOP1 choice)

Authors	E	F	P
Côte [3]	73.6	-	-
Guillevic ADS [7]	72.6	-	-
Guillevic AD [7]	63.9	-	-
Avila [1]	-	62.2	-
Guillevic AD [7]	-	78.3	-
Ollivier [13]	-	75.0	-
Gomes [8]	-	-	50.0
Freitas PPCCD [5]	-	-	70.6

E = English, F = French, P = Portuguese

Conclusion

This paper presents an algorithm based on Mutual Information for selecting an informative set of symbols from a grapheme set to be used as input data for HMM. This process was implemented and applied on our database. We started with 94 different graphemes and, through a process of reduction, validated an Entropy Alphabet composed of 29 symbols (*S$_O$*). The advantage is that we combine *I(C,G$_k$)* computing, the reduction stage and HMM into a single algorithm. The concatenations were validated during the process by means of the informative feedback provided about the information contained in each grapheme and by applying 5 *similarity-classes* based on letter shape (S_S) and their graphemes. Finally, our next efforts will be focused on the development considering a new complementary feature set.

Acknowledgements
The authors wish to thank the Pontifícia Universidade Católica do Paraná (PUCPR, Brazil), the École de Technologie Supérieure (ETS, Canada) and Fundação Araucária (Brazil) which have supported this work.

References

1. Avila, M. *Optimisation de Modeles Markoviens pour la Reconnaissance de L'ecrit*, PhD thesis, Université de Rouen, France, (1994)
2. 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
3. Côte, M. *Utilisation d'un modèle d'accès lexical et de concepts perceptifs pour la reconnaissance d'images de mots cursifs*, Ph.D. thesis, École Nationale Supérieure des Télécom, France, (1997)
4. Cover, T. M., Thomas, J. A. *Elements of Information Theory*. Wiley Series in Telecommunications, (1991)
5. Freitas, C. O. A., El Yacoubi, A., Bortolozzi, F., Sabourin, R. *Isolated word recognition in Brazilian bank check legal amounts*. In Proc. of 4th Workshop on Document Analysis and Systems, (2000) 279-290
6. 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*. In Proc. of CIFED (2000)
7. Guillevic, D. *Unconstrained handwriting recognition applied to the processing of bank cheques*, Ph.D. thesis, Department of Computer Science at Concordia University, Canada, (1995)
8. Gomes, N.R. *Reconhecimento de Palavras Manuscritas Baseado em HMM e no Emprego de Características Topológicas e Geométricas*, PhD thesis, UNICAMP, Brazil, (2000)
9. Lazzerini, B., Marcelloni, F. *Feature selection based on similarity*. Electronics Letters, Vol 38, Issue: 3, (2002) 121-122
10. Madhavanath, S., Govindaraju, V. *Perceptual features for off-line handwritten word recognition: a framework for prediction, representation and matching*. Advances in Pattern Recognition, (1998) 524-531
11. Madhavanath, S., Govindaraju, V. *The role of holistic paradigms in handwritten word recognition*. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.23, no.2, (2001) 149-164
12. L.C. Molina, L. Belanche, A. Nebot. *Feature selection algorithms: A survey and experimental evaluation*. IEEE Proceedings of the International Conference on Data Mining, (2002) 306-313
13. Ollivier, D. *Une appproche économisant les traitements pour reconnâitre l'écriture manuscrite: application à la reconnaissance des montants littéraux de chèques bancaires*, PhD thesis, Université de Paris XI Orsay, France, (1999)
14. Rabiner, L., Juang, B.H. *Fundamentals of speech recognition*. Prentice-Hall, Englewood Cliffs, N.J., USA, (1993)
15. Raman, B., Ioerger, T.R. *Enhancing Learning using Feature and Example Selection*. Journal of Machine Learning Research (submitted for publication) (2003)
16. Schomaker, L., Segers, E. *A method for the determination of features used in human reading of cursive handwriting.* In Proc. of IWFHR, (1998) 157-168
17. Strathy, N.W. *A method for segmentation of touching handwritten numerals*, Master's thesis, Concordia University, Montreal, Canada, (1993)
18. Suen, C.Y. *Réflexions sur la reconnaissance d´écriture cursive*. In Proc. of CIFED´98. Quebec, Canada, May, (1998) 1-8
19. El Yacoubi, A., Gilloux, M., Sabourin, R., Suen, C.Y. *Unconstrained handwritten word recognition using hidden markov models*. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.2, no.8, (1999) 752-760