International Journal of Pattern Recognition
and Artificial Intelligence
Vol. 18, No. 7 (2004) 1321–1338

World Scientific
www.worldscientific.com

STUDY OF PERCEPTUAL SIMILARITY BETWEEN DIFFERENT LEXICONS

CINTHIA O. A. FREITAS* and FLÁVIO BORTOLOZZI†

Pontifícia Universidade Católica do Paraná (PUCPR),
Rua: Imaculada Conceição, 1155, Prado Velho, 80215-901 Curitiba (PR), Brazil
**cinthia@ppgia.pucpr.br*
†*fborto@ppgia.pucpr.br*

ROBERT SABOURIN

École de Technologie Supérieure (ETS),
1100, Rue Notre Dame, Ouest, H3C 1K3 Montreal (QC), Canada
robert.sabourin@etsmtl.ca

The study investigates the perceptual feature similarity between different lexicons based on visual perception of the words and their representation through an observation sequence. We confirm that it is possible to use databases, which are similar in terms of morphological/perceptual features to improve the recognition performance. In this work, we demonstrated through experimentation, that it is possible to improve the recognition rate of handwritten Portuguese words by adding samples of French words in the training set. Experimental results show the efficiency of this strategy reducing the error rate.

Keywords: Perceptual similarity; visual perception; Portuguese lexicon; French lexicon; feature extraction; handwritten word recognition.

1. Introduction

Reading has been described as a language code picked up through the visual or tactile system and then processed further, and a procedure involving an assembly of human activities. These activities include visual sensory input, accurate eye movements and higher cognitive aspects of comprehension, reflecting the complexity of reading. Perceptual features, which are visually important features of the word shape, have been cited in reading studies and utilized in fluent reading.[5] These features have been used for word recognition when the lexicon is small and static.

Usually, for recognition experiments, we do not use different databases collected in different situations, collected through different machines, with writers of different nationalities. Our study investigates the perceptual feature similarity between different lexicons. These similarities were studied based on word representation through an observation sequence. We confirm that it is difficult, during these

experiments, to apply some databases, which are different in terms of morphological or perceptual features.

Having selected Portuguese and French for this study, our idea is to demonstrate, based on experiments, that it is possible to improve the recognition rates of handwritten Portuguese words by adding samples of French words in the training database. In this light, the French database is applied during HMM (Hidden Markov Model) training. The scope of this study is limited to the offline recognition of individual handwritten words from legal amounts. Comprehensive experiments validate the proposed hypothesis.

The paper is organized as follows. Section 2 presents the most important visual perception concepts and aspects of linguistic background to our study. Section 3 summarizes the handwritten Portuguese word recognition problem. Section 4 explains the hypothesis to be analyzed in this paper and discusses the perceptual feature similarity of Portuguese and French words. Section 5 presents the databases considered in the experiments. Section 6 describes the word recognition method based on HMM. Section 7 discusses the experimental results and Sec. 8 concludes this work.

2. Visual Perception Concepts and Linguistic Background

In this section we introduce a summary about the visual perception concepts related to handwritten word recognition. The Gestalt theory describes the principles of organization, which tend to encourage the emergence of perceptual forms and promote the grouping of those forms, segregated from their surroundings. This theory is beyond the scope of this paper. However, an excellent introduction to this subject can be found in Ref. 10.

Visual perception is a dynamic process and involves the observer and the object being observed. When the people are looking at an object or a scene, they analyze the structure, solve ambiguities and make connections. Generally speaking, people organize what they see.

The human eye is the sense organ responsible by lecture, except for visually impaired people who use the fingers to "read" the Braille text. For our study, the eyes and visual perception are important elements. Summarizing the visual process, the eyes work as a sophisticated photographic machine carrying the visual information to the brain, where it will be processed.

When people perceive a visual field, the pattern that emerges as a figure, and not as the background, depends on the characteristics from the field and the relation among the objects inside the field. Therefore, the perception of the form is linked with the best relation between the form and our brain, as shown in Fig. 1, where some letters "A" have an easier perception, i.e. the letters are more interpretable or readable.

The human being has a tendency to interpret a visual stimulus as a complete scene. This tendency is known in Gestalt theory as closure concept.[10]

Fig. 1. Letter "A" in different styles: printed and handwriting.

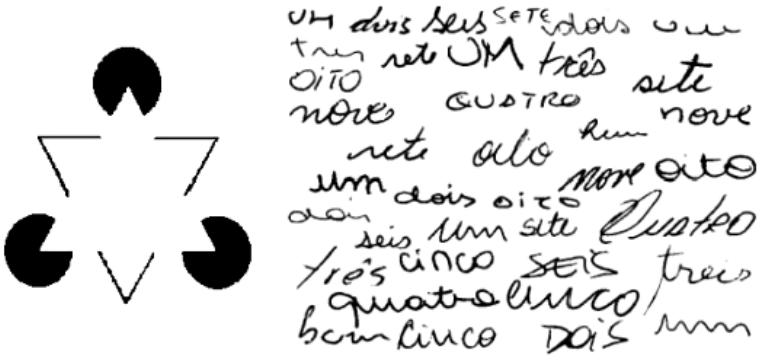


Fig. 2. Closure principle: (a) Kanizsa Triangle and (b) handwritten words.

Figure 2 illustrates this human capacity using the Kanizsa Triangle and handwritten words. When you look at the Triangle you see a complete figure, and you can read the words and letters even if they are poorly written and not missing.

For example, the first letter in the alphabet, letter "a", can appear in different writing styles, as depicted in Fig. 1. We can say, in terms of perceptual features, that the letter "A" is so different from "a", how "p" is different from "m". Moreover, "p", "b" and "d" are most similar among themselves than "b" and "B", "d" and "D". For the human reader, the difference is not perceived each instant during the writing or reading process. For an adult person, an "A" is an "A", independently of how it is written. Remember that the Gestalt principle of similarity refers to the perceptual tendency to group together objects that are similar to one another in texture, shape, and so on.[10] Since only particular properties promote grouping, these properties may constitute the basic elements of perception.[10] We call these visual elements *features*. In this work we consider that the feature set incorporates the visual perceptual concepts because the perceptual features are easily perceived by the human eye, and other authors, such as Madhvanath and Govindaraju,[5,6] have mentioned their use to human reading. The perceptual features are the most popular primitives for global recognition of handwritten words.

Other important discussion is about the linguistic background of the studied lexicons: Portuguese and French. Both are Indo-European languages. This family of language originated in the forests north of the Black Sea (in what is now Ukraine) during the Neolithic period (about 7000 BC). These people began to migrate between 3500 BC and 2500 BC, spreading west to Europe, south to the Mediterranean, north to Scandinavia, and east to India. The Indo-European family

is divided into 12 branches. Our interest is the Latin Branch. Also called Italic or Romance languages, as presented by Lehmann.[4]

Italian and Portuguese are the closest modern major languages to Latin. Arabic and Basque have influenced Spanish language. French has moved farthest from Latin in pronunciation, only its spelling gives a clue to its origins. It is because the Portuguese, French, Spanish and Italian show consistent similarities[4] when we compare, for example, the following words:

- "dear" = English, Portuguese: *caro*, French: *cher*, Spanish: *caro* and Italian: *caro*.

If we examine the words for "eight" as following:

- Portuguese: *oito*, French: *huit*, Spanish: *ocho* and Italian: *otto*.

It would seem difficult to derive them from a common source. Basing our analysis on observation of changes in language, we reconstruct as the earlier forms of this word: *okto*. This reconstruction by the comparative method can be corroborated by noting the Latin: *octo*, moreover, we can account for the changes in these four languages and also see their essential regularity, as presented by Lehmann.[4]

Concluding, this work takes into account the similarities between the studied lexicons related to the linguistic background, such as, the same origin: Indo-European languages. The feature set extracted from the word images is presented in Sec. 4.

3. Handwritten Portuguese Word Recognition

There exist several international databases of handwritten bank checks: CENPARMI, CEDAR, NIST, SRTP.[3] However, these databases do not deal with the Portuguese language. Owing to the difficulties of obtaining databases with real document checks through national bank institutions, the creation of a bank check laboratory database was chosen.[1] The database is called PUCPR-Cheques. The acquisition of the images was offline, 300 dpi, 256 gray levels. The images were binarized through the OTSU Method.[7] The database is omni-scriptor (one writer by check) and in unconstrained writing style. Our database also involves the presence of different grammatically correct words, which correspond to the same word class (1 - "um" and "hum"). This possibility exists in the Portuguese language and is not found in the French or English languages.

Our laboratory database has the following properties: minimum value of R$ 0,01 ("um centavo") and maximum value of R$ 999.999,99 ("novecentos e noventa e nove mil, novecentos e noventa e nove reais e noventa e nove centavos"); existence of the words: "real", "reais", "centavo" and "centavos".

The experiments held use three databases, called: Training (60%), Validation (20%) and Testing (20%). We have a total of 11,936 isolated words.

A "legal amount" corresponds to a numerical value, which obeys a known grammar. From the numerical value, it is possible to define five meta-classes of words, such as:

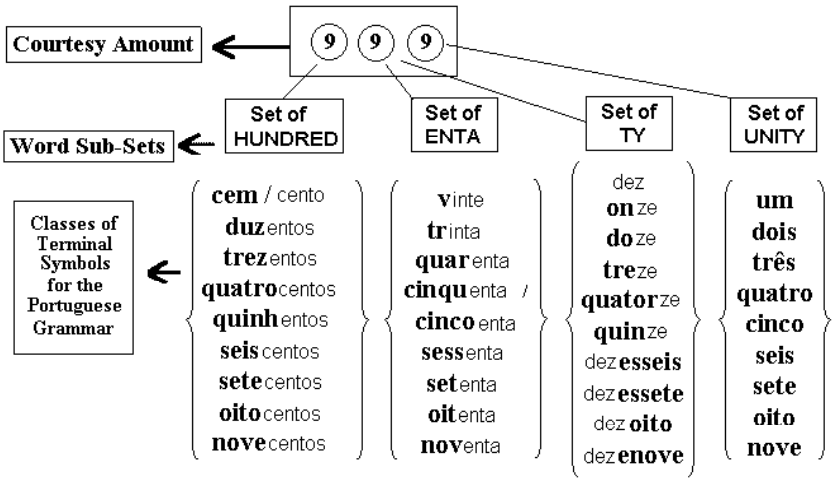


Fig. 3. Brazilian legal amounts: meta-classes of words.

- "entos"/hundred: words corresponding to numbers 100, 200, 300, 400, 500, 600, 700, 800, 900;
- "enta"/"ty": words corresponding to numbers 20, 30, 40, 50, 60 ,70, 80, 90;
- "dezena"/"teen": words corresponding to numbers 10, 11, 12, 13, 14, 15, 16, 17, 18, 19;
- "unidade"/unity: words corresponding to numbers 1, 2, 3, 4, 5, 6, 7, 8, 9, as shown in Fig. 3;
- key words (KW): "mil", "reais" or "real" and "centavos" or "centavo".

We can also observe from Fig. 3 the similarity of the suffixes and prefixes of the words in the lexicon, a fact which increases the complexity of the recognition problem. Figure 4 presents the Portuguese lexicon distribution of the training samples, following the same order within the meta-classes as shown in Fig. 3 (for example, in the unity meta-class, the number 1 corresponds to the word "um"; in the "dezena"/"teen" meta-class, the number 10 corresponds to the word "dez"; and so on, until finally, in the "entos"/hundred meta-class, the number 900 corresponds to the word "novecentos"). The meta-class "dezena"/"teen" represents only 3.16% of the training database for the Portuguese lexicon. As a result, we have only a few samples for the HMM training models, and this is the main explanation for the recognition rate observed in this meta-class in previous experiments.

Considering that the database is a laboratory database which was generated based on a numerical value for the courtesy amount, the words from meta-class "dezena"/"teen" are only written by the writer when the number at the middle position in the courtesy amount (Fig. 3) is the number "1". Otherwise, the words will belong to the "enta"/"ty" meta-class. This is an appropriate reason for collecting a few samples for this meta-class of words.

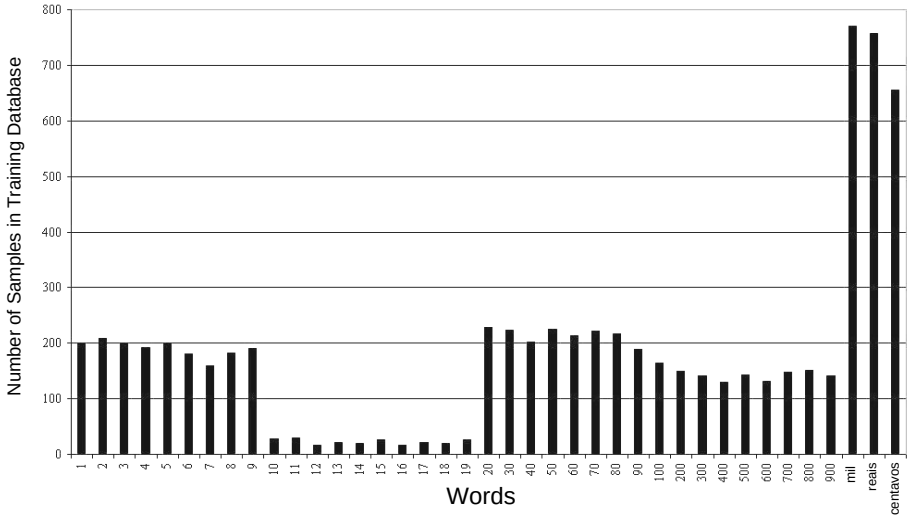


Fig. 4. Portuguese training database distribution.

Table 1. Portuguese writing styles.

Writing Style	Sample	(%)
Cursive		72%
Block print		13%
Disconnected		7%
Mixed		8%

Another important observation concerns the writing style distribution in the database. The cursive is the writing style most frequently encountered in the training database, as shown in Table 1. It is important to note that we specified our feature set for the cursive style knowing that another feature extraction approach is required for the block-print style.

4. Hypothesis: Perceptual Feature Similarity

The hypothesis analyzed in this paper is the following: Is it possible to improve handwritten Portuguese word recognition by using French words?

Taking the considerations presented in Sec. 3 into account, our objective is to improve the word recognition rate in cases where the training database contains only a few examples for obtaining the HMM models. This means improving the training of the models by adding words from the French database. We add French words because we are analyzing the perceptual feature similarity of the Portuguese and French lexicons. Even when objects do not appear to be identical, they may

still resemble one another regarding several aspects. For instance, you may be able to judge whether the baby in the carriage looks more like its mother or its father.[10]

Now, consider one class of complex forms that you are constantly judging, letters of the alphabet. Reading would be very difficult indeed if all letters of the alphabet looked very much alike. And in fact, some letters do bear a strong resemblance to one another. So, the perceptual similarity of letters was defined in Ref. 10 based on their tendency to be confused. In the feature approach, it is necessary first to define a list of features, that is, to decide what features should make up the list.

In our case, perceptual similarity presupposes that the ascenders, descenders and loops occur in the same grapheme of the words, and that the perceptual feature extraction procedure is stable, independent of the database. These similarities are presented in Table 2, which shows the positions of the perceptual features. Moreover, in support of our hypothesis, we observed the following:

- The languages (Portuguese and French) have the same origin: Indo-European/Latin Branch.
- The lexicons contain some identical words: "onze", "quatorze" and "quinze".
- The lexicons contain some identical morphemes: "quatro" and "quatre", "quarenta" and "quarente", "cinquenta" and "cinquante". A morpheme is a linguistic element, such as "qu" and "tr".
- The lexicons contain some words which are similar in terms of perceptual features (ascenders, decenders and loops): "três" and "trois", "doze" and "douze", "treze" and "treize", "trinta" and "trente", "sessenta" and "soixante".

During our studies we investigated the lexicon in the Italian and Spanish, observing the perceptual features. It was done evaluating whether the proposed method could be extended to other languages with the same origin, such as, Italian, and Spanish, as presented in Table 2.

Table 2. Perceptual feature similarity.

Number	Portuguese	French	Italian	Spanish
3	**t**r**ê**s	trois	**tre**	**tres**
4	**qu**a**t**r*o*	**qu**a**t**r*e*	*quattro*	*cuatro*
11	*on***z***e*	*on***z***e*	*undici*	*once*
12	**d**o**z***e*	**d**ou**z***e*	*dodici*	*doce*
13	**t**re**z***e*	**t**rei**z***e*	*tredici*	*trece*
14	**qu**a**t**o**rz***e*	**qu**a**t**o**rz***e*	*quattrodici*	*catorce*
15	**qu**in**z***e*	**qu**in**z***e*	*quindici*	*quince*
30	**t**rint*a*	**t**re**nte**	**trenta**	**treinta**
40	**qu**aren**t***a*	**qu**aran**t***e*	**quaranta**	*cuarenta*
50	cin**qu**en**t***a*	cinquante	**cinquanta**	**cincuenta**
60	*sessen***t***a*	*soixan***t***e*	**sessanta**	**sesenta**

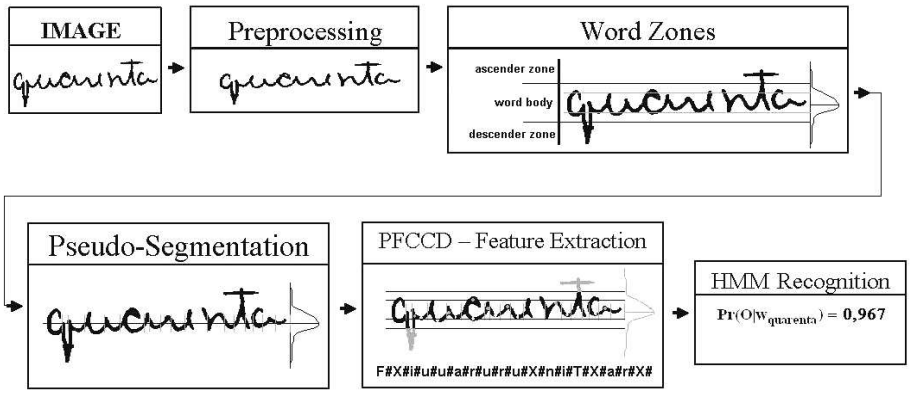


Fig. 5. System overview.

We concluded that the "ty" meta-class of words could not contribute to our study (see Table 2 — Italian and Spanish words in italic format) since the perceptual features extraction will result very different from Portuguese. Another important point is that this class of words represents our major problem. Our method can be extended to "enta" meta-class of words in additional words: "tre" and "tres" (see Table 2 — Italian and Spanish words in bold format).

Thus, the idea is to work at a high level representation: the observation sequence extracted from the word images based on the feature set. It is important to note that we are merging representations of two different databases based on extracted features.

The baseline system can be categorized as a *Global Approach*, because it avoids the explicit segmentation of words into letters or pseudo-letters and uses Hidden Markov Models (HMM). In this HMM framework, the input data are the symbols of the alphabet based on graphemes as a visible part on the Markov modeling. The word recognition system does not include a post-processor based on Portuguese lexicon. Figure 5 presents the system overview.

The system gets as input a binary image. Then, a preprocessing step (composed of slant correction and smoothing) is applied. We do not correct the base line, because we take into consideration that the legal amount in checks has two printed guidelines as indicators in the standard check form. We implemented a simple but fast algorithm presented by Yacoubi *et al.*[15] that only uses the external contour of the words to estimate the average slant of the characters.

The smoothing of the word image is performed after the slant correction. The aim of this module is to regulate the continuous contour of the word, eliminating small noises in the image. The algorithm adopted in our case is the one described by Strathy.[12] Figure 5 shows the results obtained with the preprocessing stages.

Three zones are determined based on the horizontal transition histogram: ascender, body, and descender, as shown in Fig. 5. The body of the word is the area

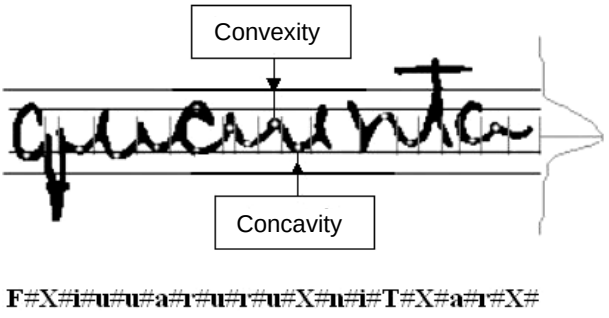


Fig. 6. Concavities and convexities.

located between ±70% from the maximum value of this histogram. This percentage value was obtained empirically by experimentation. The features are extracted from the word images and a pseudo-segmentation process is applied to obtain a sequence of corresponding observations, as depicted in Fig. 5. Between two black-white transitions over the maximum peak of the horizontal transition histogram, called the Median Line (LM), a segment is delimited and a corresponding symbol is designated to represent the extracted set of features, making up a grapheme. Only the transitions that are not found inside the loops of the body of the word are considered. If no features can be extracted in the analyzed segment, an empty symbol "X" is emitted.

In our study a grapheme is an entity that can correspond to a part of a letter, a letter, or connected letters. The features are extracted inside a pseudo-segment from the word producing a grapheme. The number of graphemes corresponds to the number of pseudo-segments observed over the LM.

Feature extraction is one of the most important steps in the success of a handwriting recognition system. It is from the features or from the characteristics of the word chosen to be extracted that it is possible to obtain the robustness of the system. The present work treats each word as a unit for its recognition.

We have used, a feature set called PFCCD (Perceptual Features and Concavities/Convexities Deficiences), which is defined as follows: ascenders, descenders and loops containing information about the feature type, the size and positional information. Concavities and convexities deficiences are extracted from word body area and labeled, as presented in Fig. 6. These deficiencies are obtained by labeling the background pixels of the input images.[8] This set of features was selected to complement the representation of the cursive word through its ligatures, as well as, we established a representation capable of describing graphemes made up of "C", "S", "E" and "Z" or, "u", "n", "r", and "i". The idea is to describe shapes not characters and to allow a high-level representation of different ligatures or strokes.

The symbol alphabet was defined based on the occurrence of the basic feature types, as well as, on the occurrence of the combination of these features in a same

Table 3. Feature set.

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

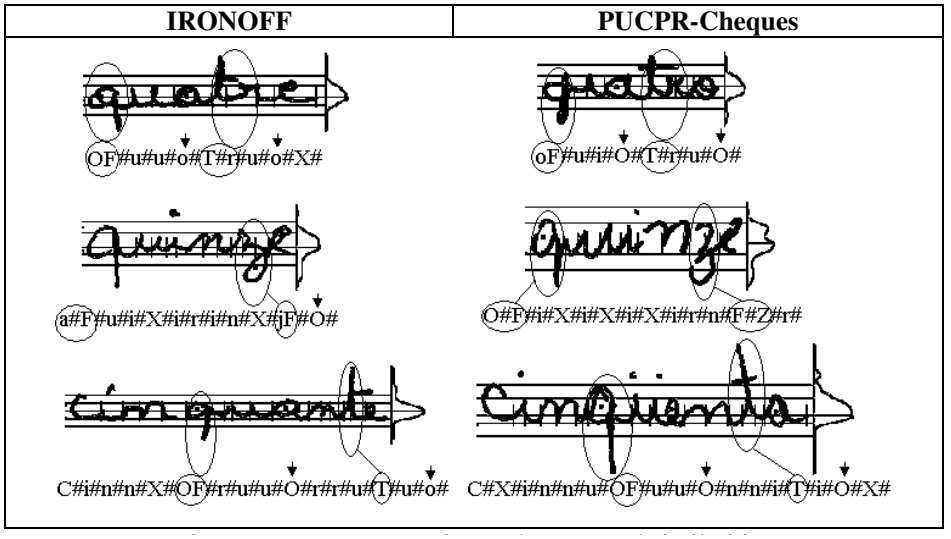


Fig. 7. Feature extraction and perceptual similarities.

segment based on Mutual Information criterion presented by Freitas *et al.*[2] Table 3 presents the feature set and its representation by symbols. The entire and definitive alphabet is composed of 29 different symbols.

Based on extracted features and visual perception, we intend to explore the shape information contained in this feature set and their consonant representation (see Table 2: t, q, z, d and Fig. 7: T, oF, OF, aF, Tr, jF, FZ; signed by circles). Another aspect is that, given that the first letter of the word is very important in the recognition process,[11] in our case the similarity of the lexicons is increased since the selected words (Portuguese and French) have the same first letter (see Table 2).

We also intend to explore the fact that the vowels (a, e, i, o — see Fig. 7, signed by arrows) exhibit the same behavior for the reader, as observed by Schomaker and Segers[11] and as expected for our system. It is because, in the letter "u" the reader has a different behavior to make the difference between "w" and "m" letters dependent on the writing style. This fact was studied in the work published by Schomaker and Segers,[11] where they observed that the reader needs to click more times on a location on the screen with the mouse pointer lighting up an area with a luminance curve which radially tapers off towards the gray background level.

5. Portuguese and French Databases

The experiments were carried out using the following databases:

- Portuguese lexicon: handwritten word database called PUCPR-Cheques, which is composed of 39 classes of words.
- French lexicon: handwritten word database from IRESTE/University of Nantes (France), called IRONOFF (IReste ON/OFF Dual Database), which contains 29 classes of words from Form B: B64 . . . B93 fields.[13,14]

Our database was collected from 2,016 writers (one writer by bank check), with no constraints regarding the writing style. The slant of the characters of the images derives from the different writing styles.

The IRONOFF database was selected because it is fully cursive. It was collected from about 700 writers, mainly of French nationality. The offline data were scanned at 300 dpi with 8 bits per pixel. In order to allow performance comparisons, the data have been divided into training database (two-thirds of total number of samples) and testing database (one-third). We used just the training database for the experiments, since the validation and testing contain only Portuguese words.

The experiments were carried out using three databases, which we called the Training, Validation and Test databases. Their composition is as follows: 60% for the Training + IRONOFF training database, 20% for the Validation database, 20% for the Test database.

The PUCPR-Cheques database contains 7,146 samples for training. When we merged the French words presented in Table 2, we obtained 2,619 new samples, for a total of 9,765 word images for the training of HMM models.

6. Word Recognition Method

The HMM theory has been successfully used to model writing variability. The theoretical formulation of HMM is beyond the scope of this paper. An excellent introduction to this subject can be found in Ref. 9.

Our interest in 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

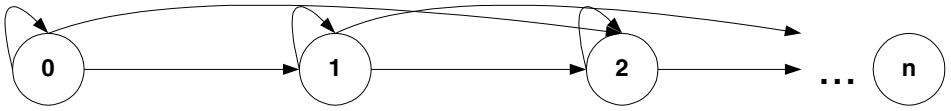


Fig. 8. HMM Bakis topology.

value for the model parameters. Markovian modeling assumes that a word image is represented by a sequence of observations. These observations should be statistically independent once the underlying hidden state sequence is known. The pseudo-segmentation and feature extraction processes are carried out to transform the image into an ordered sequence of symbols (Figs. 5 and 6). The usual solutions to overcome this problem is to first make structural assumptions and then use parameter estimation to improve the probability of generating the training data by the models. In our case, the assumptions to be made are related to the behavior of both stages. As our pseudo-segmentation process may produce a letter or a pseudo-letter (i.e. letter into two or more segments), the HMM will absorb the observation sequence during the training stage. Therefore, the HMM are capable to associate each state according to some probability density function (pdf). This property makes the HMM tolerant to spelling errors and writing style. The most important aspect is the word representation obtained by the feature extraction.

Our HMM word models are based on a Global Approach and a left-to-right discrete topology (*Bakis Topology*), where each state can skip at most two states, as shown in Fig. 8. Every HMM node has the same weight inside the topology. The lexicon size allows us to consider one model for each class. The word models are independent of the handwriting style or the orthography of a word, for example: 1 — "um" and "hum", 14 — "quatorze" and "catorze", 50 — "cinquenta" and "cincoenta". The variants are both considered correct. In the current system, a single model for the pairs "reais" and "real" and "centavos" and "centavo" is used, since the words "real" and "centavo" are found in the courtesy amounts "R$ 1,00 — um real" and "R$ XX,01 — um centavo" respectively, although they are not frequently found in financial applications.

Model training is based on the *Baum–Welch Algorithm* and the *Cross-Validation procedure*.[9] The objective of the *Cross-Validation procedure* is to monitor the general outcome during the training process. This is achieved over two sets of data: training and validation. Following iteration of the *Baum–Welch Algorithm* on the training data, the likelihood of the validation data is computed using the *Forward Algorithm*.[9]

The recognition process consists of determining the word maximizing *a posteriori* probability that a word w has generated an unknown observation sequence O,

$$\Pr(\hat{w}\,|O) = \max_{w} \Pr(w|O)\,. \tag{1}$$

Applying Bayes' rule, we obtain the fundamental equation of pattern recognition,

$$\Pr(w|O) = \frac{\Pr(O|w) \,.\, \Pr(w)}{\Pr(O)} \,. \tag{2}$$

Since $\Pr(O)$ does not depend on w, recognition becomes equivalent to maximizing the joint probability,

$$\Pr(w, O) = \Pr(O|w) \,.\, \Pr(w) \,. \tag{3}$$

$\Pr(w)$ is *a priori* probability of the word w and is related to the language of the considered task. The estimation of $\Pr(O|w)$ requires a probabilistic model that accounts for the shape variations O of the word w. We assume that such a model consists of a global Markov model created by each word w.

During the experiments, the matching scores between each model and an unknown observation sequence are carried out using the *Forward Algorithm*. We also evaluate the use of the *a priori* probability of each word class $\Pr(w_i)$ in the training database during the recognition process. In this light,

$$k = \underset{i}{\operatorname{argmax}}[\Pr(O|w_i) \cdot \Pr(w_i)] \tag{4}$$

where k is the index i that maximize the function ($i = 1, \ldots, N_c$), N_c is the number of word classes in the lexicon (39 words), O is the word observation sequence and $\Pr(w_i)$ is *a priori* probability of each word class i in the training database.

7. Experimental Results

Our experiments take into account the following conditions:

- HMM model training: the Portuguese database is merged with the French database considering these words: "trois", "quatre", "onze", "douze", "treize", "quatorze", "quinze", "trente", "quarante", "cinquante", "soixante", as shown in Table 2 and Fig. 9. Thus, we are modifying 11 HMM models,
- HMM model validation and recognition experiments: Portuguese validation and test database.

The previous results reported in Ref. 2 (I and II) and the current results obtained (III and IV) are presented in Table 4. The achieved results in the recognition and error rate for each meta-class of words are shown in Table 5. The overall reduction in the error rate (-2%) is not compelling, however, we did obtain a reduction for all meta-classes. A significant reduction in the error rate is observed for the "dezena"/"teen" meta-class (-17%), which demonstrates the efficiency of the proposed methodology (see Table 5). However, the improvement in the recognition rate of this meta-class has only a minor effect upon the final result (see Fig. 9).

In order to better understand the results, we performed an error analysis on the validation set. We observed improvements with 22 words and a loss with 7 words,

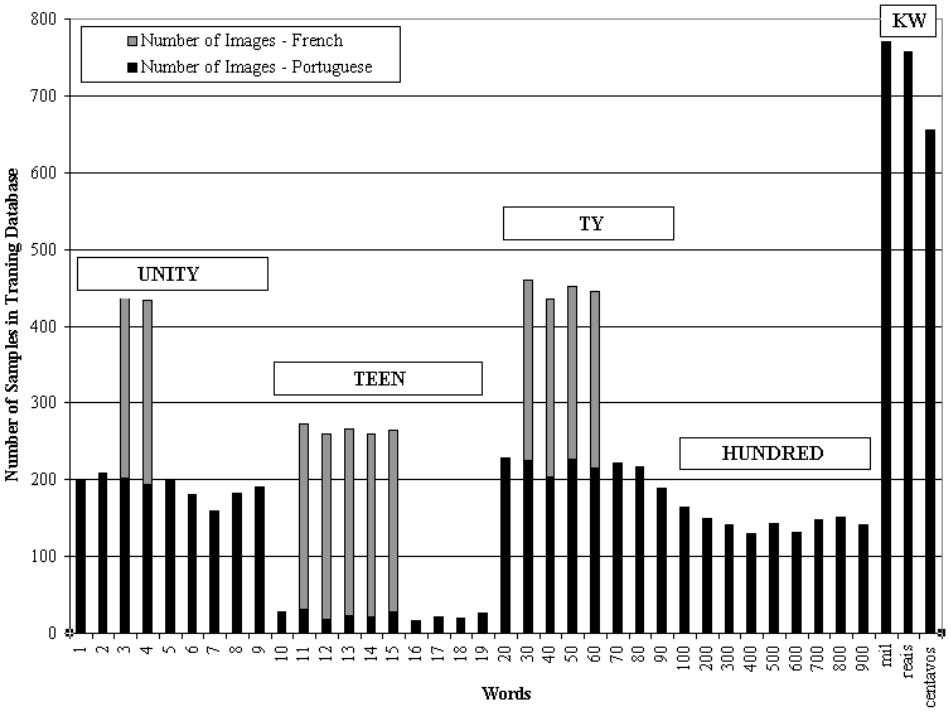


Fig. 9. Portuguese and French training databases.

Table 4. Recognition experiment (%).

Experiments	TOP1	TOP3	TOP5
I: PFCCD	67.66	86.65	92.21
II: PFCCD + Eq. (4)*	**70.61**	**88.08**	**92.84**
III: PFCCD + French	69.09	88.00	93.52
IV: PFFCD + French + Eq. (4)*	**71.96**	**89.45**	**94.17**

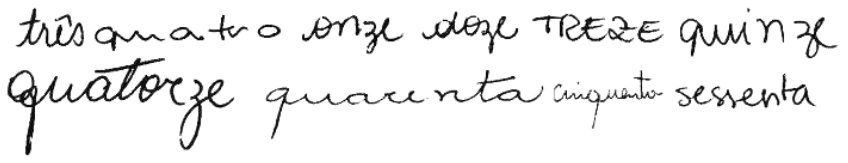
* *a priori* probability of each word class

while 10 words maintained their rates. It is important to remark that only 11 HMM models were modified, and some good results produced by our approach are shown in Fig. 10(a), demonstrating that we did, in fact, improve the cursive handwritten recognition rate. By visual analysis, it was possible to verify that 80% of the easily recognized images are cursive. Figure 10(b) shows some examples of misclassified images. These images have not perceptual features extracted from them (such as ascender and descender) prejudicing the representation level and the recognition.

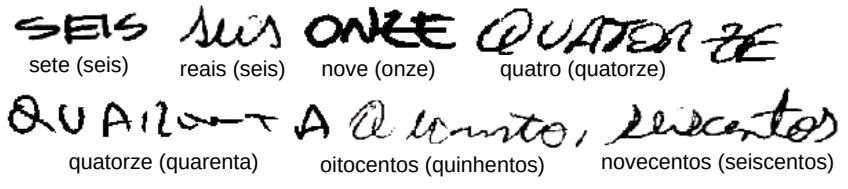
Based on these results, we can conclude that the French database selected (IRONOFF) was appropriate for our study. Taking the perceptual similarity between Portuguese and French lexicons we can enlarge the training database

Table 5. Improvement by meta-class.

Meta-Class	Recognition Rate (%) Experiment I	Recognition Rate (%) Experiment III	Error Rate Reduction (%)
Unity	70.53	71.40	− 0.87
Teen	52.00	69.33	**−17.33**
Ty	56.12	58.22	− 2.10
Hundred	64.05	64.27	− 0.22
Key Words	78.26	78.86	− 0.60
Average	67.66	69.09	− 1.43



(a)



(b)

Fig. 10. Examples of word images: (a) recognized and (b) not recognized (the correct word is the one in parenthesis).

for the meta-classes and the final results are promising. We can now compare "teen"/"dezena" meta-class on the same basis as the others. With the results obtained, we can say that maximum performance can be achieved based on the perceptual feature set from the Portuguese database for cursive writing.

8. Conclusion

In this paper, we have explored the perceptual feature similarity of two different lexicons: Portuguese and French. Normally, for recognition experiments, we do not use different databases collected in different situations, scanned through different machines, with writers of different nationalities. However, the experiments have shown the viability of our approach, which focuses on perceptual feature similarity. Moreover, these similarities were studied based on word representation through an observation sequence. We confirm that it is difficult, during these experiments, to apply some databases which are different in terms of morphological or perceptual features, e.g. English and French databases. Future work will provide a complementary feature set study and an evaluation of "artificial" words generated based on

warping algorithms as presented by Yaeger *et al.*[15] or by randomizing morphemes, such as prefixes and suffixes. These studies will contribute to investigate better the HMM training related to error tolerance.

Acknowledgments

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. C. O. A. Freitas, A. El Yacoubi, F. Bortolozzi and R. Sabourin, Isolated word recognition in Brazilian bank check legal amounts, *Fourth IAPR Int. Workshop Document Analysis Systems*, (DAS 2000), 10–13 December 2000, Rio de Janeiro, Brazil, pp. 276–290.
2. C. O. A. Freitas, F. Bortolozzi and R. Sabourin, Handwritten word recognition: an approach based on mutual information for feature set validation, *IEEE Sixth Int. Conf. Document Analysis and Recognition*, 2001, pp. 665–669.
3. J. J. Hull and R. K. Fenrich, Large database organization for document images, in *Fundamentals in Handwriting Recognition*, ed. S. Impedovo, Nato ASI Series, Vol. 124, 1993, pp. 397–414.
4. W. P. Lehmann, *Historical Linguistics: An Introduction*, University of Texas (Holt, Rinehart and Winston, 1962).
5. S. Madhavanath and V. Govindaraju, Perceptual features for off-line handwritten word recognition: a framework for prediction, representation and matching, *Advances in Pattern Recognition*, Sydney, Australia, 1998, pp. 524–531.
6. S. Madhavanath and V. Govindaraju, The role of holistic paradigms in handwritten word recognition, *IEEE Trans. Patt. Anal. Mach. Intell.* **23**(2) (2001) 149–164.
7. N. Otsu, A threshold selection method from gray-level histograms, *IEEE Trans. Syst. Man. Cybern.* **9**(1) (1979) 63–66.
8. J. R. Parker, *Algorithms for Image Processing and Computer Vision* (John Wiley, 1997).
9. L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition* (Prentice Hall, 1993).
10. R. Sekuler and R. Blake, *Perception*, 3rd edn. (McGraw-Hill, 1994).
11. L. Schomaker and E. Segers, A method for the determination of features used in human reading of cursive handwriting, *6th Int. Workshop on Frontiers in Handwriting Recognition*, 1998, IWFHR, Korea, pp. 157–168.
12. N. W. Strathy, A method for segmentation of touching handwritten numerals, Master's thesis, Concordia University, Montreal-Canada (1993).
13. C. Viard-Gaudin, The ironoff user manual, Technical Report, IRESTE, University of Nantes, France, 1999.
14. C. Viard-Gaudin, P. M. Lallican, S. Knerr and P. Binter, The ireste on/off (ironoff) dual handwriting database, *IEEE Int. Conf. Document Analysis and Recognition*, 1999, pp. 455–458.
15. A. Yacoubi, M. Gilloux, R. Sabourin and C. Y. Suen, Unconstrained handwritten word recognition using hidden markov models, *IEEE Trans. Patt. Anal. Mach. Intell.* **2**(8) (1999) 752–760.

16. L. Yaeger, R. Lyon and B. Webb, Effective training of a neural network character classifier for word recognition, in *Proc. Advances in Neural Information Processing Systems*, 1997, Seattle, USA, pp. 807–813.

Cinthia O. A. Freitas received the B.S. degree in civil engineering in 1985 from Universidade Federal do Paraná (UFPR-Curitiba-Brazil), a M.Sc. degree in electrical engineering and industrial informatics from Centro Federal de Educação Tecnológica do Paraná (CEFET/PR-Curitiba-Brazil) in 1990, and Ph.D. degree in applied computer science from Pontifícia Universidade Católica do Paraná (PUCPR-Curitiba-Brazil) in 2001. Since 1985 she is a Professor in Computer Science and Computer Engineering Departments at PUCPR. Currently, she is a Full Professor and researcher in the post-graduated program in applied computer science (PPGIA) at PUCPR.

Her research interests are handwriting recognition, symbol recognition, document image analysis and forensic science.

Flávio Bortolozzi obtained the B.S. degree in mathematics in 1977 from Pontifícia Universidade Católica do Paraná (PUCPR-Curitiba-Brazil), a B.S. degree in civil engineering in 1980 from PUCPR, and a Ph.D. degree in system engineering (computer vision) from the Université de Technologie de Compiègne, France, in 1990 where he worked on trinocular vision. From 1994 to 1999, he was the Head of the Department of Informatics, and Dean of the College of Exact Sciences and Technology at PUCPR. Currently, he is a Full Professor at the Computer Science Department and the Vice-Rector for Research and Postgraduate Programs at PUCPR. Since 1996, he is a member of CNPq — Conselho Nacional de Pesquisa Cientifica e Tecnológia in Brasil. He is the author (and co-author) of more than 200 scientific publications included in journals and conference proceedings. He was the Symposium Chair of BSDIA'97 (Brazilian Symposium on Document Image Analysis, Curitiba, Brazil). He was nominated as Conference Chair of the next ICDAR'07 (9th International Conference on Document Analysis and Recognition) that will be held in Curitiba, Brazil in 2007.

His research interests are in computer vision, handwriting recognition, document image analysis, educational multimedia, and hypermedia.

Robert Sabourin received the B.Ing., M.Sc.A. and Ph.D. degrees in electrical engineering from the École Polytechnique de Montréal in 1977, 1980 and 1991, respectively. In 1977, he joined the physics department of the Montreal University where he was responsible for the design, experimentation and development of scientific instrumentation for the Mont Mégantic Astronomical Observatory. His main contribution was in the design and implementation of a microprocessor-based fine tracking system combined with a low-light level CCD detector. In 1983, he joined the staff of the École de Technologie Supérieure, Université du Québec, in Montréal where he co-founded the Department of Automated Manufacturing Engineering where he is currently Full Professor and teaches Pattern Recognition, Evolutionary Algorithms, Neural Networks and Fuzzy Systems. In 1992, he joined also the Computer Science Department of the Pontifícia Universidade Católica do Paraná (PUCPR-Curitiba-Brazil) where he was co-responsible for the implementation in 1995 of a master's program and in 1998, a Ph.D. program in applied computer science (PPGIA). Since 1996, he is a senior member of the Centre for Pattern Recognition and Machine Intelligence (CENPARMI, Concordia University). He is the author (and co-author) of more than 150 scientific publications included in journals and conference proceedings. He was co-chair of the program committee of CIFED'98 (Conférence Internationale Francophone sur l'Écrit et le Document, Québec, Canada) and IWFHR'04 (9th International Workshop on Frontiers in Handwriting Recognition, Tokyo, Japan). He was nominated as Conference co-chair of the next ICDAR'07 (9th International Conference on Document Analysis and Recognition) that will be held in Curitiba, Brazil in 2007.

His research interests are in the areas of handwriting recognition, and signature verification for banking and postal applications.