Lexicon-free Writer Dependent Approach for Off-line Handwritten Word Recognition

Luciana Ribeiro Veloso, João Marques de Carvalho, Cinthia O. A. Freitas e Luiz S. Oliveira

***Resumo*— Neste artigo discute-se uma abordagem de reconhecimento de palavras manuscritas dependente do escritor para léxico livre. Diferentemente das abordagens mais corriqueiras, com escritor independente e léxico dependente, em que o mesmo modelo de aprendizagem de máquina é utilizado para todos os escritores, na abordagem dependente do escritor cada escritor tem o seu próprio modelo. A desvantagem de ter vários modelos é compensada por uma melhora desempenho e a não vinculação a um léxico específico. Os resultados experimentais obtidos com os quatro escritores distintos, cada um com um banco de dados de aproximadamente 4.000 palavras, produziu taxas de reconhecimento de 83% à 93%.**

***Palavras-Chave*— Reconhecimento de manuscritos, Reconhecimento de palavras, Dependente do escritor.**

***Abstract*— In this paper, we discuss a writer-dependent lexicon-free approach for handwritten word recognition. Differently from the more common writer-independent lexicon-dependent approach, where the same machine learning model is used for all writers, in the writer-dependent approach each writer has his/her own model. The drawback of having several different models is compensated by a better performance and by not being tied to a specific lexicon. Experimental results obtained with four different writers, each with a database of approximately 4,000 words produced recognition rates from 83% to 93%.**

***Keywords*— Handwriting Recognition, Word Recognition, Writer-Dependent.**

I. INTRODUCTION

Due to the amazing technological evolution that we have testified during the last few decades a common question one may ask is: will handwriting be threatened with extinction? [1] [2]. What we have observed in practice, however, is that handwriting persists in the age of the digital computer, because of the convenience of paper and pen as compared to keyboards for numerous day-to-day situations. For example, students in a classroom usually store language, equations, and graphs with a pen on a paper. Professors also use pen and paper to prepare their lecture notes. This typical paradigm has led to the concept of pen computing, where the keyboard is an expensive and non-ergonomic component to be replaced by a pentip position sensitive surface superimposed on a graphic display that generates electronic ink. This led us to two different approaches: off-line and on-line handwriting recognition. In the former data are converted to digital format by scanning the writing on paper, while in the latter data are acquired by writing with a special pen on an electronic surface. On-line systems for handwriting recognition are available in hand-held computers such as personal digital assistant and tablet PCs. The performance of on-line recognition systems in general is higher, since dynamic characteristics are available during feature extraction and also because the recognition engine is usually writer-dependent and the writing style is not unconstrained. This kind of strategy maximizes recognition accuracy by requiring that end users first provide some form of structured input in order to tailor the recognition engine to the user.

Off-line systems, on the other hand, are typically writer-independent and lexicon dependent, which provides the benefit that end users can simply step up to the application and start writing without any awareness of or special interaction with the underlying recognizer, although being restricted to a specific lexicon. To improve the performance of off-line handwriting recognition systems, some domain of applications, such as, writer identification [3] and symbol recognition [4] [5] have been using writer-dependent engines. In the specific case of handwriting recognition, the writer-dependent strategy has been investigated to recognize texts written by a single person [6] [7]. Since data for training is usually an issue, several works rely on adaptation techniques to improve the performance of the models over specific subsets of the data they are trained to recognize [8] [9]. This is exactly what we investigate in this work. First we selected four different writers and collected approximately 4,000 words for each one of them, to be used as training and test databases. The decision of testing with four writers, aims to exclude the possibility that the performance obtained is due to some particular writer feature that may influence the system. If the system presents a similar performance for four distinct writers, that is a strong indication that its behavior can be generalized to any writer. No lexicon restriction was imposed on the collected words. The sampled documents were basically lecture notes of different courses and texts in general written by professors and undergraduate students. Thereafter, these words were segmented into characters which were used to train writer-dependent HMM for characters. Here we have an important task that is convert the lexicon-free approach into a writer-dependent engine. Our system is not tied to a specific vocabulary because the HMM character models are trained in the writing style of the writer and not in a specific lexicon features. Therefore, the system should be able to recognize any words written by that writer. Comprehensive experimental results demonstrate that this kind of strategy works well with the amount of samples per writer that we used for training.

Luciana Ribeiro Veloso, Federal University of Paraíba - UFPB, Brazil, luciana.veloso@ct.ufpb.br. João Marques de Cavalho, Federal University of campina grande - UFCG, Brazil, carvalho@dee.ufcg.edu.br. Cinthia O. A. Freitas, Pontifical Catholic University of Paraná - PUCPR, Brazil,cinthia@ppgia.pucpr.br. Luiz S. Oliveira, Federal University of Paraná - UFPR, Brazil,lesoliveira@inf.ufpr.br.

The recognition rates achieved range from 83% to 93% and they compare favorably to the writer-independent approach. The paper is organized as follows: Section 2 presents the advantages and problems related to writer-dependent approach. Section 3 presents the system and its stages as well as the database. In Section 4, the experimental results are presented and discussed. Finally, in Section 5 our conclusions and a plan for future works are presented.

II. Writer-Dependent Approach

Although writer-independent systems offer the user the convenience of 'off the shelf' operation, they cannot match the accuracy of systems that are trained to a specific user writing style. This is especially true regarding cursive handwriting where the inherent variations in writing style are considerable. For most people the natural style of writing is a mix of cursive and discrete writing. Another limitation is that writer-independent approaches are always lexicon-dependent, i.e., they can only recognize words from a specific and usually not very large vocabulary, like the month names [10] or the legal amounts [11]. Due to this characteristic, this class of systems usually looks at the handwritten words as single patterns, for which global features are extracted. The drawback of the writer-dependent approach is the requirement to train the system by providing it with samples of the writer´s own handwriting. The inconvenience of the initial training, however, is compensated by an increased accuracy, especially for those writers that make an extended use of such a system. A writer-dependent recognition system normally makes only a fraction of errors that a similar writer-independent system would make. Another significant gain with a writer-dependent system is that it can be trained to recognize any word written by its specific user, not being tied to any specific vocabulary. For that, words are usually segmented into characters or pieces of characters for which features are extracted. The characters, once identified, are then assembled to form the recognized word. This is the approach that we investigate in this work.

III. System Overview

The handwriting recognition system is composed of the following stages or modules: pre-processing, word segmentation into characters, multi-images generator, feature extraction (global, perceptual and directional characteristics), vector quantization, and HMM training [12] [13]. As we can notice, this architecture can be applied for both writer-dependent and writer-independent approaches. The difference in this case lies in the training procedure, since the writer-dependent approach generates several specific HMM character models for each writer. As mentioned before, the system applies a lexicon-free approach to a writer-dependent engine.

A. Pre-processing

The pre-processing stage includes the following operations: binarization [14], words baseline and skew correction and smoothing. The main goal of this stage is to reduce the writer intra-class variability, as illustrated in Figure 1. The best results were obtained using a method presented by Côté [15]. For proper multi-images generation it is necessary that the words are normalized for skew and slant, as these parameters affect directly the segmentation process. Since we are dealing with a writer-dependent approach, the first thought was to avoid skew correction, since it is a writer specific feature. However, after some experiments we observed that skew correction is necessary even for the writer-dependent approach.

B. Segmentation

Different segmentation methods and approaches were implemented and verified during this work [16] [17]. The best results were obtained combining projection histograms and local minimum methods, as shown in Figure 2. As we can observe from Figure 2, our method is based on character over-segmentation. Hence, the character segments (pseudo-characters) produced by segmentation should be somehow combined to form characters recognizable by the classifier. In this manner, the next step in the segmentation process consists in combining the pseudo-characters to validate the segmentation hypotheses.

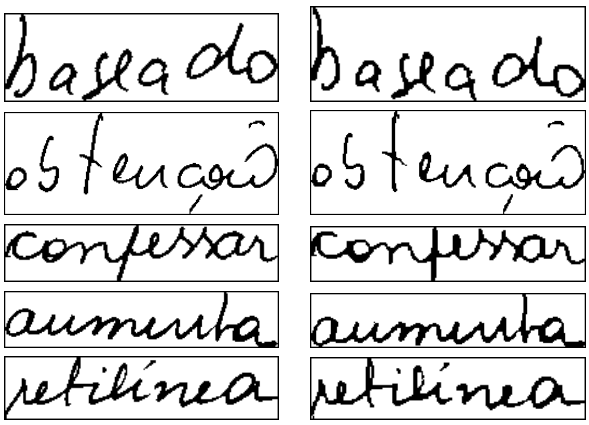

Fig. 1. Pre-processing results.

Fig. 2. Segmentation points created using projection histogram and local minimum methods.

This process considers up to five consecutive segments to create a character. Based on observations, we have concluded that a character is always segmented in less than five pieces. Figure 3 shows an example of the concatenation process where the word "vamos"is segmented into seven pseudo-characters, from which four different hypotheses are generated. In the training, the correct segmentation hypotheses were manually labeled.

C. Feature Extraction

In this work, global, perceptual and directional features are extracted and represented by different numerical values. The features extracted from each character segment form a vector

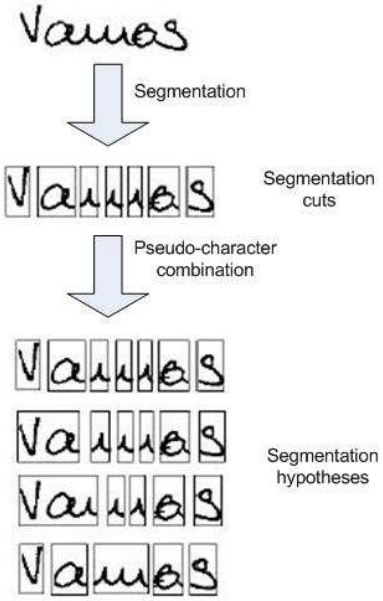

Fig. 3. Example of the segmentation hypotheses generation.

of different dimension. Perceptual features are considered high-level features due to the important role they play in the human reading process, which uses word features like ascenders, descenders, and loops. The components of the feature vector are as follows:

- Perceptual (10 features): ascender and descender positions, ascender and descender heights, closed loop size and location (upper, body or lower zone) [11];
- Global (06 features): metrics based on circularity and squared proportionality, center of gravity position computed for each generated sub-image, normalized central moments;
- Directional (10 features): based on concavity analysis [18], where for each white image pixel (background pixel) it is tested which of the four main directions (NEW) leads to a black (contour) pixel [**?**]. In this representation we used 10 different labels.

D. Vector Quantization (VQ)

After feature vectors extraction we applied a vector quantization process based on Linde-Buzo-Gray (LBG) algorithm [19]. During training, VQ is used for construction of the dictionaries associated with each writer features vector. At the recognition stage, VQ is employed to generate the symbols used by the HMMs.

E. Hidden Markov Models

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 the works by Rabiner and Juang [13]. The recognition process consists of determining the word maximizing the a posteriori probability that a word w has generated an unknown observation sequence O,

$$P(\hat{w} \mid O) = \max_{w} P(w/O) \tag{1}$$

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

$$P(w \mid O) = \frac{P(O \mid w).P(w)}{P(O)} \tag{2}$$

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

$$P(w, O) = P(O \mid w).P(w) \tag{3}$$

where $P(w)$ is the a priori probability of the word w and is related to the language of the considered task. The estimation of $P(O \mid w)$ requires a probabilistic model that accounts for the shape variations O of the word w. Our choice of the HMM is due to 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 character models are based on a left-to-right discrete topology (Bakis Topology) with 6, 11 or 16 states depending on the number of symbols considered by each image [13]. The best topology considered 16 states, as shown in Figure 4.

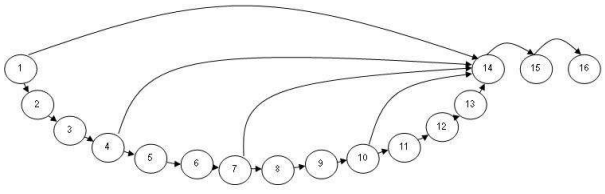

Fig. 4. HMM topology.

The character models training are based on the Baum-Welch Algorithm and the Cross-Validation process. 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 iteration of the Baum-Welch Algorithm on the training data, the likelihood of the validation data is computed using the Forward Algorithm. Finally, the matching scores between each model and an unknown observation sequence are carried out using the Forward Algorithm.

F. Database

The database has a total of 15,707 images from four writers, which were split into Training, Validation and Testing sets, as presented in Table I. The sex and the schooling of writers are represented by Female or Male (F or M) and Undergraduate or Graduate (U or G), respectively. The set of words collected in the data base was different for each writers as reflected in the average length of the words (number of characters) for each writer (Table I). Figure 6 presents some examples of long and short words in the database.

The most important aspect of the database is the number of word images for each writer. In our case this number is

TABLE I
NUMBER AND AVERAGE LENGTH (NUMBER OF CHARACTERS) OF WORDS FOR EACH WRITER.

Writer	Training	Validation	Testing	Avg. Length (std) (%)
W1: F, U	2383	517	1350	6 (2.4)
W2: F, U	2009	489	1380	7 (2.3)
W3: M, G	1998	459	1057	7 (2.7)
W4: F, U	2383	518	1164	7 (2.9)

enough to support the character segmention hypothesis and HMMs training. We are using a lexicon-free approach into a writer-dependent system. For that, a generic system is adapted to each writer using the individual writing style represented by his/her database. The number of each character in the training set as well as the number of each one in the validation set must be representative, so that the system can learn each individual writing style. Figure 5 presents the occurrence of each character in the training set for writer 1.

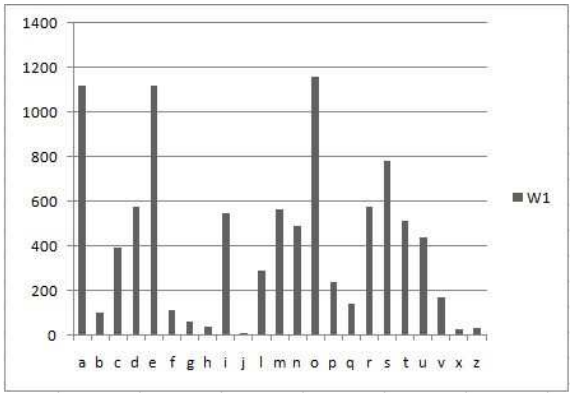

Fig. 5. Number of characters in the training set of the writer 1.

The isolated vowels ("a", "o", "e") are not being considered in this analysis. Only words written in lower case style were used in the experiments. The database contains no occurrence of the letters "k", "y" and "w" because were not presents in the documents sample.

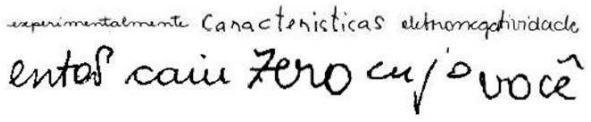

Fig. 6. Long and short words in the database.

IV. EXPERIMENTAL RESULTS AND DISCUSSIONS

Table II reports the experimental results for each writer. The best results were obtained when only the characters combinations allowed by the language vocabulary were tested. This does not harm our lexicon-free claim, as the system is still not tied to a specific set of words. The only requirement is to have the recognized word checked against a language dictionary. Therefore, given a sequence of m images, our goal was to find which valid word with m characters maximizes P(w|O), where w is the word to be recognized and O is the observation sequence. What we could notice is that the writer-dependent approach solved confusions that the writer-independent strategy could not handle properly. On the other hand, the experiments also show that different writers with different writing styles share some other confusions.

TABLE II
EXPERIMENTAL RESULTS.

Writer	Recognition Rate (%)
W1	92.96
W2	92.85
W3	90.78
W4	83.31

After analyzing the confusion matrices we observed that characters "h", "g", e "x" are the ones that produce most of the confusion to writer W1. Character "g" is often confused with "q". Also, was observed a high similarity between letters in the following pairs: "i" and "e", "l" and "e", "m" and "n", "j" and "f", "j" and "z". For writer W2 most of the confusion appeared between characters "e" and "x" and a high similarity was observed among "b", "c", "f", "j", and "x". Writer W3 presented confusion problems with characters "h", "q", "v", e "x". For writer W4 the most significant confusions occurred with "b", "f", "h", "j", and "x". All these cases produced recognition rates lower than 60%. Writer W4 presented the worst performance, which can be explained by the use of the same pattern to represent different letters and vice-versa, as illustrated in Figure 7. We can observe in Figure 7 that character "x" and the syllable "se" are both written as very similar patterns. On the other hand, we can also observe the same character (letter "b") written in two different ways. Both cases make training and recognition difficult, therefore harming system performance. Table III presents the confusion matrix for writer W4 and points the most important confusions. Finally, we observed that character "x" has recognition problems regardless of the writer.

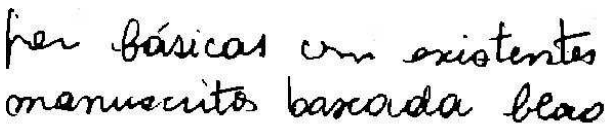

Fig. 7. Writing style of W4 (words: "ser", "básicas", "um", "existentes", "manuscritos", "baseada", "bloco").

V. CONCLUSION

In this paper we described a writer-dependent lexicon-free handwriting recognition system. We have collected and labeled more than 15,000 word images from four different writers, which were used for training and testing the system. As future work we plan to perform writer-dependent feature selection. We strongly believe that writer dependent approach can be further improved through specific feature selection.

Acknowledgements

The authors wish to thank CNPq (grant 303137/2007-0) and CAPES/PROCAD, which partially supported this work.

TABLE III

CONFUSION MATRIX FOR THE WRITER W4 (%).

	a	b	c	d	e	f	g	h	i	j	l	m	n	o	p	q	r	s	t	u	v	x	z
a	85,9	0,2	1,0	0,8	2,0	0,1	0,0	0,1	0,2	0,0	0,0	0,1	0,6	5,8	0,2	0,1	0,6	0,9	0,0	0,3	0,0	1,0	0,0
b	1,7	39,0	3,4	0,0	0,0	1,7	0,0	8,5	0,0	0,0	16,9	0,0	0,0	0,0	11,9	0,0	0,0	0,0	11,9	3,4	1,7	0,0	0,0
c	0,9	1,1	49,3	0,0	11,7	0,0	0,0	0,0	8,3	0,0	3,4	0,3	0,3	1,7	0,0	0,0	9,1	2,6	0,3	2,6	6,0	2,0	0,6
d	2,0	0,0	0,3	96,4	0,0	0,0	0,0	0,0	0,0	0,0	0,6	0,0	0,0	1,1	0,0	0,3	0,0	0,0	0,3	0,0	0,0	0,0	0,0
e	1,2	0,5	12,7	0,3	54,1	0,0	0,0	0,3	11,3	0,0	2,7	0,2	0,1	3,8	0,1	0,0	5,9	2,3	0,5	0,8	1,9	1,2	0,0
f	0,0	0,0	0,0	0,0	0,0	58,9	1,1	0,0	0,0	12,6	0,0	1,1	0,0	0,0	14,7	4,2	0,0	0,0	5,3	0,0	0,0	0,0	2,1
g	1,0	0,0	0,0	0,0	0,0	2,9	78,1	0,0	0,0	1,0	0,0	0,0	0,0	1,9	0,0	8,6	0,0	1,0	0,0	0,0	1,9	0,0	3,8
h	2,2	15,2	2,2	2,2	0,0	0,0	0,0	54,3	0,0	0,0	10,9	0,0	0,0	0,0	2,2	0,0	2,2	0,0	4,3	4,3	0,0	0,0	0,0
i	0,2	0,0	4,4	0,0	8,0	0,3	0,0	0,0	65,8	0,2	0,5	0,0	1,3	0,3	0,3	0,0	13,2	2,7	0,6	0,9	1,3	0,2	0,0
j	0,0	0,0	0,0	0,0	0,0	26,7	0,0	0,0	0,0	40,0	0,0	0,0	0,0	0,0	20,0	6,7	0,0	0,0	6,7	0,0	0,0	0,0	0,0
l	1,0	5,9	1,4	0,0	5,2	0,3	0,0	4,5	0,0	0,3	71,3	0,0	0,0	0,7	0,3	0,7	0,0	0,3	7,3	0,0	0,0	0,0	0,3
m	1,2	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,3	0,0	0,0	87,1	7,0	0,0	0,0	0,0	0,6	0,0	0,0	2,0	1,2	0,6	0,0
n	1,1	0,0	1,3	0,2	0,6	0,0	0,0	0,2	0,4	0,0	0,0	10,7	66,3	0,4	0,4	0,0	3,4	4,1	0,0	5,4	3,2	2,1	0,0
o	8,9	0,3	2,1	1,2	4,5	0,3	0,1	0,0	0,3	0,0	0,8	0,4	0,4	69,6	0,1	0,5	0,9	6,1	0,9	0,3	0,7	1,2	0,3
p	0,0	0,8	0,0	0,0	0,0	4,1	0,0	0,4	0,0	0,8	0,0	0,0	0,8	0,0	88,8	0,0	0,0	0,4	1,7	0,4	0,4	0,0	1,2
q	3,6	0,0	0,0	0,0	0,0	14,3	21,4	0,0	0,0	0,0	0,0	0,0	0,0	0,0	3,6	53,6	0,0	0,0	0,0	0,0	0,0	3,6	0,0
r	0,2	0,0	2,9	0,0	6,1	0,2	0,0	0,3	12,3	0,2	0,3	2,2	1,9	0,5	0,5	0,0	52,7	10,7	0,5	2,5	3,5	2,7	0,0
s	2,1	0,5	2,0	0,6	4,6	0,3	0,2	0,2	3,8	0,3	0,2	0,8	3,0	5,2	0,6	0,3	9,0	61,4	0,3	0,2	2,1	0,9	1,7
t	0,0	2,3	2,3	0,6	0,6	2,9	0,0	1,0	0,4	0,8	5,3	0,0	0,6	0,6	3,9	0,2	0,6	0,4	77,3	0,0	0,0	0,0	0,2
u	1,6	0,0	2,4	0,0	2,4	0,0	0,0	0,0	1,2	0,0	0,0	3,3	4,1	0,4	1,2	0,0	2,4	2,0	0,0	73,5	9,8	0,4	0,0
v	0,0	0,8	2,5	0,0	1,6	0,0	0,0	0,8	0,8	1,6	0,8	2,5	4,9	1,6	0,0	0,0	9,0	2,5	0,0	7,4	62,3	0,8	0,0
x	2,0	0,0	12,2	2,0	6,1	0,0	0,0	0,0	6,1	0,0	0,0	4,1	10,2	4,1	0,0	0,0	6,1	18,4	0,0	2,0	8,2	18,4	0,0
z	2,6	0,0	0,0	0,0	0,0	2,6	2,6	0,0	0,0	2,6	0,0	0,0	2,6	0,0	0,0	0,0	2,6	7,7	2,6	0,0	0,0	0,0	74,4

REFERENCES

[1] Plamondon, R. and Srihari, S. N. On-Line and Off-Line Handwriting Recognition: A Comprehensive Survery. *IEEE PAMI, 22(1):63-84, 2000.*

[2] G. Lorette. Handwriting recognition or reading? What is the situation at the dawn of the 3rd millenium? *IJDAR, 2(1):2-12, 1999.*

[3] L. Schomaker, M. Bulacu. Automatic Writer Identification Using Connected-Component Contours and Edge-Based Features of Uppercase Western Script. *IEEE PAMI, 26(6):787- 798, 2004.*

[4] J. LaViolla Jr, R. C. Zeleznik. On-Line and Off-Line Handwriting Recognition: A Comprehensive Survery. *IEEE PAMI 29(11):1917-1926, 2007.*

[5] B. Lazzerini, F. Marcelloni, L. Reyneri. Beatrix: A self-learning system for off-line recognition of handwritten texts. *Pattern Recognition Letters, 18:583-594, 1997.*

[6] U. Marti and H. Bunke,H. Handwritten sentence recognition. *15th Internat. Conf. on Pattern Recognition, 2000.*

[7] M. Liwicki, A. Schlapbach, H. Bunke. Writer-dependent Recognition of Handwritten White Board Notes in Smart Meeting Room Environments. *8th IAPR Workshop on Document Analysis Systems, p. 151-157, 2008.*

[8] A. Vinciarelli, S. Bengio. Writer adaptation techniques in HMM based Off-Line Cursive Script Recognition. *Pattern Recognition Letters, 23:905-916, 2002.*

[9] A. Nosary, L. Heutte, T. Paquet,. Unsupervised writer adaptation applied to handwritten text recognition. *Pattern Recognition, 37:385-388, 2003.*

[10] M. N. Kapp, C.O.A. Freitas, R. Sabourin. Methodology for the Design of NN-based Month-Word Recognizers Written on Brazilian Bank Checks. *International Journal of Image and Vision Computing, Image and Vision Computing, 25:40-49, 2007.*

[11] Freitas, F. Bortolozzi, R. Sabourin. Study of Perceptual Similarity between Different Lexicons. *International Journal of Pattern Recognition and Artificial Intelligence, 18(7):1321-1338, 2004.*

[12] M.-Y. Chen, A. Kundu, J. Zhou. Off-Line handwritten word recognition using Hidden Markov Model type stochastic network. *IEEE PAMI, 16(5):481-496, 1994.*

[13] Rabiner, L., Juang, B.H. Fundamentals of Speech Recognition. *Prentice Hall Inc., 1993.*

[14] N. Otsu. A threshold selection method from graylevel histograms. *IEEE Trans. Syst. Man. Cybern. 9(1):63-66, 1979.*

[15] M. Côté, E. Lecolinet, M. Cheriet, M., C. Y. Suen. Automatic reading of cursive scripts using a reading model and perceptual concepts. *International Journal on Document Analysis and Recognition, 1:3-17, 1998.*

[16] L. R. Veloso, R. P. Sousa, J. M. Carvalho. A New Method for Cursive Word Segmentation. *Proc. of IEEE International Conference on Image Processing - ICIP2000, September 2000.*

[17] L. R. Veloso, J. M. Carvalho. Segmentação de Escrita Cursiva por Histograma de Projeção Vertical. *CD-ROM of the XVIII Brazilian Telecommunications Symposium (SBrT2000), Gramado - RS, Brazil, September 2000.*

[18] J. R. Parker. Algorithms For Image Processing and ComputerVision. *Jonh Wiley Sons, 1997.*

[19] Y. Linde, A. Buzo, R. M. Gray. An Algorithm for Vector Quantizer Design. *IEEE Transactions on Communications, 28(1):84-95, 1980.*