A NEW STRATEGY FOR IMPROVING FEATURE SETS IN A DISCRETE HMM-BASED HANDWRITING RECOGNITION SYSTEM

F. GRANDIDIER AND R. SABOURIN

CENPARMI, Concordia University,
1455 de Maisonneuve Blvd West, Montréal H3G 1M8, Canada
LIVIA, Ecole de Technologie Supérieure,
1100 rue Notre Dame Ouest, Montréal H3C 1K3, Canada
E-mail: frede@cenparmi.concordia.ca, sabourin@gpa.etsmtl.ca

C.Y. SUEN

CENPARMI, Concordia University,
1455 de Maisonneuve Blvd West, Montréal H3G 1M8, Canada

M. GILLOUX

RMO, Service de Recherche Technique de La Poste,
10, rue de l'Ile Mabon, BP86334, 44263 Nantes Cedex 2, France

In this paper we introduce a new strategy for improving a discrete HMM-based handwriting recognition system, by integrating several information sources from specialized feature sets. For a given system, the basic idea is to keep the most discriminative features, and to replace the others with new ones obtained from new feature spaces. After evaluating the individual discriminative power of each single feature, the set is divided into two subsets: one containing the discriminative features, and the second the others. Considering feature classes in the non-discriminative feature subset allows the specialization of new feature sets on specific problems. The application of this strategy to an existing system showed an improvement of 16% in the recognition rate when a lexicon of 1000 city names was used.

1 Introduction

The domain of handwriting recognition belongs to the field of 2D pattern recognition, challenged by high intra- and inter-classes variability. Thus, during the development of a handwriting recognition system, the extraction of features from images is very important [1], as is the integration of this information into the system.

Researchers have been working in the field of handwriting recognition for a few decades already, so we can find in the literature many techniques for feature extraction, especially designed for the recognition of characters [1] or for more general 2D patterns [2]. After this step of extraction, the recognition system must carry out its task based solely upon the features. Thus we can consider that the feature space is the perception that the recognition system has of a shape.

The information embedded in the features must be integrated into the recognition system. To incorporate multiple sources of information, several options

In: L.R.B. Schomaker and L.G. Vuurpijl (Eds.), Proceedings of the Seventh International Workshop on Frontiers in Handwriting Recognition, September 11-13 2000, Amsterdam, ISBN 90-76942-01-3, Nijmegen: International Unipen Foundation, pp 113-122

are possible. First, from each feature space a classifier can be built, then we will try to combine their scores in an optimal way [3, 4]. Another strategy is to combine the feature sets by constructing their Cartesian product, and build a new feature set [5-7]. The drawback of this approach lies in the exponential increase in the number of parameters.

This paper introduces a new strategy for improving the performance of a recognition system by integrating several information sources. In the next section we present the formalism of this new strategy. The description of its implementation on an existing system [5], and the results obtained, are discussed in Section 3. Finally some conclusions and plans for future work are presented.

2 A new strategy for improving a feature set

The basic idea of this new strategy comes from some observations obtained from the evaluation of the SRTP handwriting recognition system [6]. We concluded that word length has a strong influence on recognition performance, and thus it is easier to recognize long words than short ones. In the case of long words the system has more features and contextual information to perform the recognition. Moreover, the presence of the most discriminative features is more probable in long sequences of observation. This leads us to conclude that the individual discriminative power of each single feature is important and should be improved in order to better recognize short words as well as long words.

In summary, the basic idea of the new strategy is to improve a given feature set by keeping features showing good discriminative power, and replacing the poor ones with new features.

2.1 Evaluating the discriminative power of features

The first step is to evaluate the discriminative power of each single feature. The conditional perplexity introduced in [7] was chosen for this purpose. This indicator is based on the statistical notions of entropy and perplexity from information theory. They were introduced to the field of speech recognition by Bahl [8] to evaluate the difficulty of a specific recognition task. In this case, the entropy $\boldsymbol{H}$ is given by:

$$\boldsymbol{H} = -\sum_{w} \boldsymbol{p}(\boldsymbol{w}) \cdot \log \boldsymbol{p}(\boldsymbol{w}) \tag{1}$$

where $\boldsymbol{p(w)}$ is the *a priori* probability of word $\boldsymbol{w}$, and the sum is calculated over all the words from the vocabulary of the application.

In [7] the authors define the conditional entropy of a feature $\boldsymbol{f_j}$ by:

$$\boldsymbol{H}(\boldsymbol{f}_j) = -\sum_{i=1}^{N_C} \boldsymbol{p}(\boldsymbol{c}_i \mid \boldsymbol{f}_j) \cdot \log \boldsymbol{p}(\boldsymbol{c}_i \mid \boldsymbol{f}_j) \tag{2}$$

where c_i are the classes considered in the modeling and N_c the number of those classes. $H(f_j)$ quantifies the capability of feature f_j to discriminate between the classes c_i. This function reaches a maximum value corresponding to $\log N_c$ when:

$$p(c_i \mid f_j) = \frac{1}{N_c} \quad \forall c_i \tag{3}$$

In this particular case, no information is embedded in feature f_j to discriminate between the N_c classes, and it can be considered as useless. The minimum value, which is **0**, is obtained when there exists one class c_i such that:

$$p(c_i \mid f_j) = 1 \quad \text{and} \quad p(c_k \mid f_j) = 0 \qquad \forall k \neq i \tag{4}$$

The conditional perplexity $PP(f_j)$ of a feature f_j is obtained from the relation:

$$PP(f_j) = 2^{H(f_j)} \tag{5}$$

This function varies between **1** and N_c; thus it can be directly compared to the number of classes c_i involved. This is the advantage of using the perplexity instead of the entropy.

The conditional perplexity quantifies the capability of each single feature to discriminate between all classes, without the help of recognition results. To quantify the discriminative power of a feature set, the global entropy H must be calculated:

$$H = \sum_{j=1}^{N_f} p(f_j) \cdot H(f_j) \tag{6}$$

where N_f is the number of features and $p(f_j)$ the *a priori* probability of the feature f_j. The global perplexity PP of a feature set is related to the global entropy H as in (5).

For a given feature set E_i, we are now able to rank it according to conditional perplexity values. The greater the value, the less discriminative the feature. The notation of the ordering set is $(E_i, \succ)$. By determining a perplexity threshold τ, the feature set may be divided into 2 subsets: D_i, composed of the discriminative features, and $\overline{D}_i$, the non-discriminative ones.

2.2 The descent of a perceptual level

The second step of our novel strategy is to replace the features judged non-discriminative by some others. For this a new feature space will be used to obtain a new characterization of the information present in the handwriting segment previously labeled with the non-discriminative features. This phase is called the descent of a perceptual level. As mentioned in the introduction, the features can be considered as the perception of the shape by the recognition system; then the shift of feature space can be considered as a change of perceptual level.

We considered two ways of carrying out this step, depending on features contained in $\overline{D_i}$. We must introduce here the notion of class of features: it is a subset of features sharing some of the same basic properties. Considering the first feature set used in [5], which is based on ascenders, descenders and loops, a class of features can be defined as the subset of all features related to ascenders. It may also be characterized from an associated objective value, like the conditional perplexity [7]. In this case, a class of features is defined by a range of values from the perplexity domain.

If there is only one class of features in subset $\overline{D_i}$ (or none), then the descent of a perceptual level will be carried out with only one new feature set. If several feature classes are considered, then one new feature set will be used for each class, as in Figure 1. Each feature set $\boldsymbol{E_i}$ at the second level will be specialized considering specific properties of feature class $\boldsymbol{C_{E_i}}$.

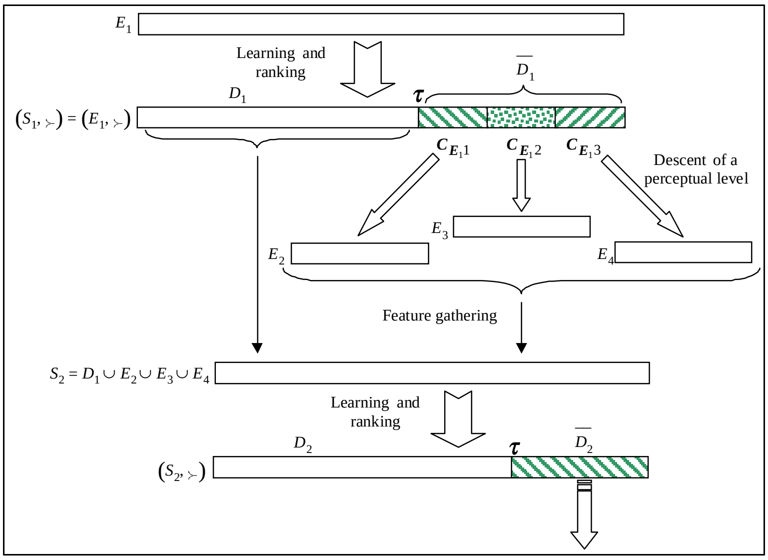

Figure 1: Synopsis of the new strategy for the improvement of feature sets

During the recognition phase, the system will extract from all graphemes a feature from the set of the first perceptual level, $\boldsymbol{E_1}$. For each grapheme, if the extracted feature belongs to the non-discriminative subset $\overline{D_1}$, then another extraction process is performed according to the specific feature class.

The next step of the process is to gather the discriminative features of the upper level $\boldsymbol{D_i}$ with the new ones, and build the new feature set $\boldsymbol{S_{i+1}}$. Then, a new training of the system is carried out. The global discriminative power of the improved feature set may be evaluated and compared with the value obtained at the previous level. The conditional perplexity of each single feature can also be calculated, and the feature set ranked. The entire process can then be re-iterated if necessary.

In this process some parameters must be fixed. First, the perplexity threshold value τ must be chosen in order to keep only the best features in D_i, and the number of perceptual levels to be considered must be determined in advance.

3 Applying the new strategy to an existing recognition system

To evaluate the pertinence of the strategy presented in the previous section, we tried to improve the performance of the handwriting recognition system described in [5]. This is a discrete HMM-based off-line handwriting recognition system, using an analytic approach with explicit segmentation. After some pre-processing and segmentation steps, each grapheme is represented by two symbols, each from a different set of features. The first set E_1 (27 symbols) is based on global features: ascenders, descenders and loops (see Table 2), and is more dedicated to cursive handwriting. The second feature set E_2 (14 symbols) is based on an analysis of the horizontal and vertical contour transition histograms, and better characterizes handprinting. Three databases were used in this experiment: 12023 city names for learning, 3475 for validation, and 4674 for testing. The performance of this system, using only E_1, only E_2 or the combination $E_1 \times E_2$, are shown in Table 1.

Table 1: Evaluation of the standard system [5]

Feature set used	Number of features	Global perplexity	Lexicon size 10	100	1000
E_1	27	38.29	96.64%	88.29%	71.58%
E_2	14	35.23	97.50%	91.16%	78.01%
$E_1 \times E_2$	378	20.72	98.69%	95.42%	86.82%

We chose to apply the process described in Section 2 to improve feature set E_1, for two reasons. First, its global discriminative power is worse than that of the second feature set. Secondly, we noticed after training that more than 50% of the graphemes are characterized by the same feature "-", corresponding to the absence of ascenders, descenders and loops.

3.1 Evaluating the discriminative power of the features

The first step is to evaluate the individual discriminative power of each feature in E_1. Practically, the evaluation of the entropies and perplexities needs the estimation of probabilities $p(c_i \mid f_j)$ and $p(f_j)$. In Markovian modeling, the number of classes c_i depends on the model architecture. In our system, where observations are emitted along transitions, each feature related to a grapheme can be generated by either a letter, or the first, second or third part of a letter. Consequently there are 4 distinct classes for each character model. The use of tied state concept (from HMM theory) to share the transition associated with the third part allows us to reduce the number of these classes (3×68 models + 5 = 209 classes). Since the exact labeling of each training sample is available, the backtracking procedure of the Viterbi algorithm is used to recover the best path in the word model, and to label each feature/grapheme couple with its class $\boldsymbol{c}_i$ in an automatic way, as explained in [7]. The probabilities are then computed from the frequencies of occurrence. The conditional perplexity values of $\boldsymbol{E}_1$ are partially shown in Table 2. Here features are represented by an arbitrary code letter, and defined by the matrix of basic properties below. For detected loops, the size is evaluated only for those located in the median zone. Moreover, the relative position of these loops, with respect to ascenders or descenders, is taken into account. They are sorted in order of increasing perplexity.

Table 2: Conditional perplexity values of features from $\boldsymbol{E}_1$ (s: small, L: large, x: presence, r: right, l: left)

Feature	***l***	***g***	***G***	**…**	***S***	***D***	***d***	***b***	***T***	***y***	***f***	***B***	***t***	***H***	***h***	**-**
Ascender	s			…	L	L	s	s	L		L		s	L	s	
Descender		s	L	…						s	L	L	L			
Upper loop	x			…												
Median loop		L	L	…	L r	s l	s l	s r	s r							
Lower loop		X	x	…												
Occurrence (%)	0.01	0.01	0.1	…	0.9	0.8	1.3	1.1	0.5	0.3	0.4	1.3	0.4	11.8	5.8	51.9
Perplexity	1.9	3.5	4.7	…	18.5	20.2	20.9	23	23.6	28.5	30.3	32.7	41	44.3	49.6	61.1

The big difference in perplexity between the most discriminative features and the lesser ones confirms that some features are really discriminative and must be preserved. We notice also that the seven less discriminative features (conditional perplexity > 24) are the only ones with no loops. In fact, they belong to the same feature class defined by the basic property: features without loops.

3.2 The concavity feature space

As explained above, more than 50% of the graphemes are characterized by the feature “-” from $\boldsymbol{E}_1$. These graphemes must come from characters showing no ascender, no descender and no loop: *c, i, m, n, r, s, u, v, w, x*. A visual analysis of the data showed that parts of other characters, if the loops are broken, are also labeled with this feature. Due to the fact that for handprinted characters, upper and lower zones are usually non-existent, most of them are also labeled with “-”. The analysis of concavities seems to be appropriate to our problem, so we chose it as the new feature space for the second perceptual level.

We used the white pixel labeling technique to calculate the concavities [9]. By analyzing graphemes in their bounding boxes, only 39 different configurations are found. An extensive analysis is performed when close concavities are encountered: an exit is searched by following two consecutive directions; 4 configurations are added. The black pixel ratio is also taken into account. Finally, the size of the concavity vectors is 44, and all components are real values in the range 0 to 1.

As our system is based on a discrete representation of the information, a vector quantization algorithm must be used. We chose the LBG algorithm [10] for speed and simplicity of its use.

3.3 *Descent of a perceptual level based on one class of features*

In order to test the strategy described in Section 2, the feature set must be divided into two subsets, $\boldsymbol{D_1}$ and $\overline{\boldsymbol{D}}_1$. An analysis of the feature conditional perplexity values led us to choose 24 for the threshold $\boldsymbol{\tau}$, because there is a significant gap in the perplexities at this point, so this is a natural place to cut. In addition, with this threshold value, all the features contained in $\overline{\boldsymbol{D}}_1$ belong to the same class of feature: features without loops. We must notice also that the global frequency of occurrence of these features is really important (70%).

The concavity vector was extracted for all graphemes characterized at the first level by one feature in the non-discriminative subset $\overline{\boldsymbol{D}}_1$ (141907 graphemes). Then the vector quantization algorithm was applied twice, to obtain 64 and 128 centroids and build two new feature sets. Two experiments were performed; each feature set was used to replace the features in the non-discriminative subset $\overline{\boldsymbol{D}}_1$. Gathering features from the new set and from $\boldsymbol{D_1}$ allows the training and testing of the new system. The performances obtained are shown in Table 3.

Table 3: Performance obtained considering only one class of features

Feature set used	Number of features	Global perplexity	Lexicon size 10	100	1000
$\boldsymbol{E_1}$	27	38.29	96.64%	88.29%	71.58%
$\boldsymbol{S_2^1 = D_1 + 64}$	84	21.12	98.48%	94.61%	85.28%
$\boldsymbol{S_2^2 = D_1 + 128}$	148	18.08	98.72%	95.93%	86.05%
$\boldsymbol{E_1 \times E_2}$	378	20.72	98.69%	95.42%	86.82%

These results show that our new technique confers improvement to the discriminative power of the feature set, and significantly increases the recognition rate (14.4% improvement for lexicon 1000) without adding too many features (*i.e.* without adding too many system parameters). We conclude that the new strategy proposed to improve the performance of a recognition system is attractive.

3.4 *Descent of a perceptual level based on several classes of features*

In order to test the second strategy for perceptual level descent, an analysis of the features in the non-discriminative subsets $\overline{\boldsymbol{D}}_1$ was performed. We identified the following classes of features:

- "-": feature showing no ascender, no descender, and no loop,
- "***hH***": features showing small or large ascender only,
- "***tf***": features showing large descender and small or large ascender,
- "***By***": features showing small or large descender only.

For each class, several sets of features were built by increasing the number of centroids during vector quantization. Each new feature set was used to replace its respective class of features. Then one system training per new feature set was carried out. After each training, the global discriminative power of the new feature set was evaluated, as presented in Table 4. Only this indicator was used to estimate the improvement of the recognition system, because the computation time is really shorter than that needed for recognition rate evaluation, and we assume that there is a relation between these two functions [8].

Table 4: Discriminative power of the improved feature sets

Feature class	Number of samples	Frequency	Number of added features 4	8	16	32	64
"-"	77 898	51.9 %	34.07	30.28	27.42	24.41	22.30
"***hH***"	25 660	17.6 %	37.24	36.83	36.62	**35.81**	
"***tf***"	1 644	0.8 %	38.14	38.07	**38.05**	37.99	
"***By***"	2 144	1.6 %	38.12	38.06	**37.99**	37.86	

An analysis of Table 4 shows that in all cases the global perplexity value decreases with the injection of a new feature set, and particularly in the case of feature "-". This is due to its high frequency of occurrence, and also the important intra-class shape variations of related graphemes, as we previously observed. For these reasons we decided to give special care to this feature during the final system performance evaluation. The influence of the number of features replacing "-" was studied. For these experiments, each class of feature in $\overline{\boldsymbol{D}}_1$ was replaced by the number of features marked in bold in Table 4 (*i.e.* 64 new features), but class "-" was replaced by 8 to 128 features. From each case the system was trained and tested; the results are presented in Table 5.

Table 5: Performance obtained considering several feature classes

Feature set used	Number of features	Global perplexity	Lexicon size 10	100	1000
$\boldsymbol{E_1}$	27	38.29	96.64%	88.29%	71.58%
$\boldsymbol{S_2^3 = D_1 + 64 + 8}$	92	24.57	98.40%	93.67%	83.20%
$\boldsymbol{S_2^4 = D_1 + 64 + 16}$	100	22.26	98.63%	94.76%	85.17%
$\boldsymbol{S_2^5 = D_1 + 64 + 32}$	116	19.82	98.74%	95.76%	85.67%
$\boldsymbol{S_2^6 = D_1 + 64 + 64}$	148	18.08	98.78%	95.61%	86.56%
$\boldsymbol{S_2^7 = D_1 + 64 + 128}$	212	16.11	98.65%	95.38%	87.51%
$\boldsymbol{E_1 \times E_2}$	378	20.72	98.69%	95.42%	86.82%

The increase in features by replacing "-" brings improvement in recognition performance, and in feature set discriminative power. For lexicon size 1000, the recognition rate is increased by 16%. This performance is better than the standard system using the combination of the two feature sets $\boldsymbol{E_1 \times E_2}$. Moreover, this is done with 44% fewer features, and hence with fewer system parameters. The performance at lexicon size 1000 is pointed out because the vocabulary associated with the target application, *i.e.* mail sorting, usually exceed this number. If we compare the results in Table 5 with those in Table 3, we can conclude that the descent of a perceptual level with several feature classes is better. In the second case, each feature set used to substitute a feature class is specialized for the solution of a sub-problem identified by this feature class and the associated basic properties.

3.5 *Combining of the improved feature sets with* $\boldsymbol{E_2}$

In order to evaluate the global improvement of our system, we combined the best improved feature set obtained from each strategy of perceptual level descent (*i.e.* $\boldsymbol{S_2^2}$ and $\boldsymbol{S_2^7}$) with the handprinted feature set $\boldsymbol{E_2}$, based on the technique used previously in [5]. Two new systems were built and tested.

Table 6: Performance obtained by combining the improved feature sets with $\boldsymbol{E_2}$

Feature set used	Number of features	Global perplexity	Lexicon size 10	100	1000
$\boldsymbol{E_1} \times \boldsymbol{E_2}$	378	20.72	98.69%	95.42%	86.82%
$\boldsymbol{S_2}^2 \times \boldsymbol{E_2}$	2 072	10.44	98.91%	95.31%	87.44%
$\boldsymbol{S_2}^7 \times \boldsymbol{E_2}$	2 968	9.22	98.80%	95.81%	88.68%

From the results obtained (Table 6), we can observe a small improvement (1 or 2 %), and a drastic increase in the number of features. In this case, feature set $\boldsymbol{E_2}$ is distributed uniformly on $\boldsymbol{S_2^i}$, without any specialization on a specific sub-problem to be solved. We conclude that this combination strategy leads to saturation of the recognition rate, combined with a huge number of parameters to be evaluated. Moreover, by excessively increasing the number of features, the system could fall into an over-learning phase. For these reasons, the strategy for improving a recognition system presented in this paper is more attractive than the feature set combination tested here.

4 Conclusions and future works

We have introduced a new strategy for improving of a feature set in a discrete HMM-based handwriting recognition system. The basic idea is to keep the most discriminative features, and to replace the others by new ones specialized for a specific sub-problem. The development of this technique needs an individual discriminative power indicator to rank the feature set and divided it in two subsets; we chose conditional perplexity. Then the descent of a perceptual level is carried out depending on the number of feature classes identified in the non-discriminative feature subset. For each class, a new feature set is built and specialized in order to solve the specific sub-problem characterized by the properties of the class of features. After gathering the features introduced at this level with the most discriminative of the upper level, system training and testing can be performed. This iterative process can be performed several times if necessary.

Some experiments were carried out in order to evaluate this new strategy. We conclude that this technique confers a significant improvement in recognition rate, while keeping a reasonable number of features. In our approach, the number of parameters grows with an additive factor, in opposition to the feature set combination used in [5], where a multiplicative factor can be observed. This reflects a strength of the proposed technique. Moreover, at a specific perceptual level, the integrated feature sets may be specialized on a sub-problem identified by the classes of features.

To be able to evaluate all the possibilities of our new strategy, we must develop other feature extraction techniques. This may be done according to the problems to be solved, *i.e.* the pattern recognition sub-problems identified at the first perceptual level by the feature classes. We want also to define some objective criteria, in order to choose the value of the threshold τ and to stop the iterative process automatically. Finally, the vector quantization technique should be improved to optimize the number of features in the different sets at each perceptual level, this will help the system avoid over-learning.

Acknowledgements: This work was supported by the Service de Recherche Technique de La Poste (SRTP) at Nantes, France, the Ecole de Technologie Supérieure, and the Center for Pattern Recognition and Machine Intelligence in Montréal, Canada.

References

1. O.D. Trier, A.K. Jain and T. Taxt, Feature Extraction Methods for Character Recognition - A Survey. *Pattern Recognition* **29** (1996) pp. 641-662.
2. S. Loncaric, A Survey of Shape Analysis Techniques. *Pattern Recognition* **31** (1998) pp. 983-1001.
3. P.D. Gader, M.A. Mohamed and J.M. Keller, Fusion of Handwritten Word Classifiers. *Pattern Recognition Letters* **17** (1996) pp. 577-584.
4. J. Kittler, M. Hatef, R.P.W. Duin and J. Matas, On Combining Classifiers. *IEEE Trans. on PAMI* **20** (1998) pp. 226-239.
5. A. El-Yacoubi, M. Gilloux, R. Sabourin and C.Y. Suen, An HMM-Based Approach for Off-Line Unconstrained Handwritten Word Modeling and Recognition. *IEEE Trans. on PAMI* **21** (1999) pp. 752-760.
6. F. Grandidier, R. Sabourin, A. El-Yacoubi, M. Gilloux, and C.Y. Suen, Influence of Word Length on Handwriting Recognition. *Proc. 5th Int'l Conf. on Document Analysis and Recognition* (Bangalore, India, Sep. 1999) pp. 777-780.
7. A. El-Yacoubi, M. Gilloux, R. Sabourin and C.Y. Suen, Objective Evaluation of the Discriminant Power of Features in an HMM-Based Word Recognition System. *Proc. 1st Brazilian Symposium: Advances in Document Image Analysis* (Curitiba, Brazil, Nov. 1997) pp. 60-73.
8. L.R. Bahl, F. Jelinek and R.L. Mercer, A Maximum Likelihood Approach to Continuous Speech Recognition. *IEEE Trans. on PAMI* **5** (1983) pp. 179-190.
9. L. Heutte, *Reconnaissance de Caractères Manuscrits : Application à la Lecture Automatique des Chèques et des Enveloppes Postales*. Thèse de doctorat, Université de Rouen, (1994) 240 pages.
10. Y. Linde, A. Buzo and R.M. Gray, An Algorithm for Vector Quantizer Design. *IEEE Trans. on Communication* **28** (1980) pp.84-95.