Noname manuscript No.
(will be inserted by the editor)

A Categorization System for Handwritten Documents

Thierry Paquet, Laurent Heutte, Guillaume Koch, Clément Chatelain

Received: date / Accepted: date

Abstract This paper presents a complete system able to categorize handwritten documents, i.e. to classify documents according to their topic. The categorization approach is based on the detection of some discriminative keywords prior to the use of the well known *tf-idf* representation for document categorization. Two keyword extraction strategies are explored. The first one proceeds to the recognition of the whole document. However, the performance of this strategy strongly decreases when the lexicon size increases. The second strategy only extracts the discriminative keywords in the handwritten documents. This information extraction strategy relies on the integration of a rejection model (or anti-lexicon model) in the recognition system. Experiments have been carried out on an unconstrained handwritten document database coming from an industrial application concerning the processing of incoming mails. Results show that the discriminative keyword extraction system leads to better recall/precision tradeoffs than the full recognition strategy. The keyword extraction strategy also outperforms the full recognition strategy for the categorization task.

Keywords Handwritten Document Categorization · Handwriting Recognition · Keyword Extraction · Shallow language model

C. Chatelain
INSA-Rouen
LITIS EA 4108
Tel.: +33-23-2955210
Fax.: +33-23-2955022
E-mail: clement.chatelain@insa-rouen.fr

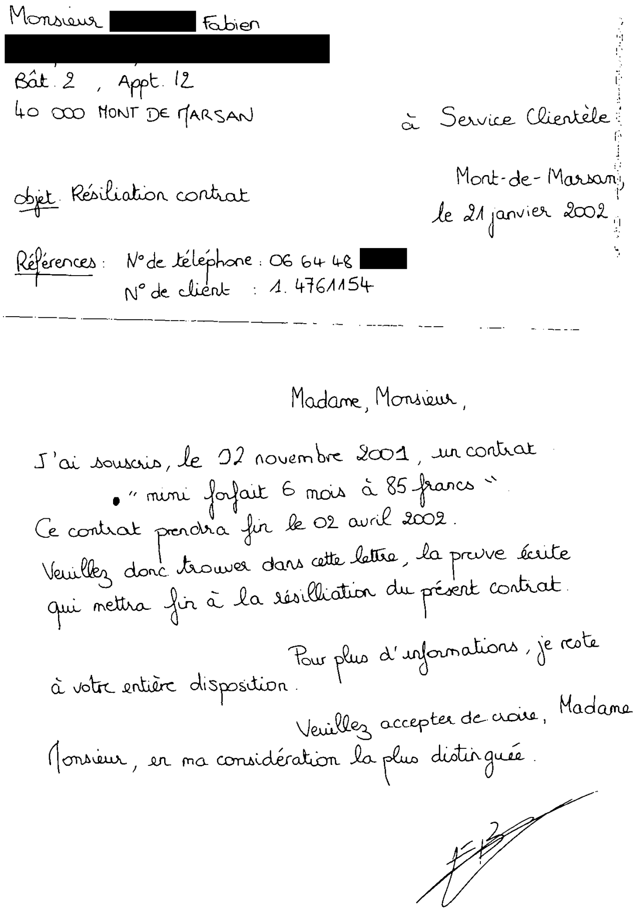
1 Introduction

These last years, the number of paper documents generated by administrative and economic activities has exploded. To facilitate the storage, processing and transferring of these documents, Electronic Document Management (EDM) systems have been developed, where paper documents are scanned, stored and transferred electronically. In this context, the automatic reading of the document image content has seen a fast expansion. We have thereby observed the development of applications for processing targeted, specific problems, such as the automatic reading of forms, postal addresses or bank checks [Plamondon 00, Koerich 05, Lorette 07]. Besides these specific applications, the automatic processing of handwritten documents remains a difficult and open problem: there is no system able to recognize an entire page of unconstrained cursive handwriting without using prior knowledge. This can be mainly explained by the huge variability in the writing style. In the literature, some recent works [Zimmermann 06, Vinciarelli 04, Bertolami 08] have addressed the processing of lightly constrained handwritten documents such as free mails. Among these projects, some address the full document recognition, whereas others are more oriented towards the rejection of misrecognized hypotheses or out of vocabulary words [Zimmermann 04, Koerich 05]. Some other projects e.g. [Cao 07] aim at indexing handwritten documents by their textual content for retrieval purposes. One alternative, called keyword spotting, has been proposed in order to provide indexation facilities of a collection of handwritten documents [Manmatha 97, Rath 07, Adamek 07]. In this case, word images are clustered using some appropriate features and elastic matching, thus avoiding the difficult task of recognition. Although interesting [1], these studies are not suitable for omni-writer mail documents since i) they are based on a word image matching process, assuming word images boundaries are known; and ii) these collections exhibit some stability in the writing styles of the various writers encountered in the collection.

To the best of our knowledge, only one specific study presented in [Vinciarelli 05] has been devoted to the categorization of handwritten documents. This pioneer study has been carried out using a mono-writer corpus built specifically. The proposed approach uses the word outputs of a mono-writer recognition system to feed a word vector representation optimized for the categorization task at end. The categorization stage is performed using a classifier such as SVM, or KNN.

In this paper, we address the categorization task of omni-writer handwritten documents such as incoming mail documents. Thousands of such documents are received day by day in customer services of companies for various claims (address change, change of contract, contract cancellation, etc.). One example of such a document is illustrated on Figure 1 (two other documents can be found in appendix 6. Today, paper documents are scanned and then digitally sent to some remote service in charge of mail topic identification. The mail is finally sent to the appropriate department of the company. Automating this process requires the machine to read omni-writer handwritten documents for detecting its topic. This second task is known as categorization in the field of Information Retrieval [Aas 99, Baeza-Yates 99] and for document images [Doermann 98]. This task aims at classifying documents according to their subject matter. It is based on the detection of some specific keywords that are selected for their discriminative power among the various classes of documents. As opposed to keyword spotting where keywords are determined in an unsupervised manner, here keywords must be determined in a supervised manner considering the category (label) associated to each document. While the major difficulty in the categorization of electronic documents lies on the selection of these relevant keywords, the categorization task of handwritten documents also requires the system to detect these keywords in the document image, whatever the handwriting style. This is an additional difficulty that has not received very much attention until now except in [Vinciarelli 05] with a limitation to a single writer.

This paper has two major contributions. First it addresses the question of omni-writer handwritten document categorization. The second contribution lies in the use of a shallow handwritten keywords extraction system on real-world documents. The three main processing stages of the proposed system are: layout analysis, keyword extraction and categorization (see figure 2).



Monsieur [redacted] Fabien
[redacted]
Bât. 2 , Appt. 12
40 000 MONT DE MARSAN

à Service Clientèle

Mont-de-Marsan,
le 21 janvier 2002

objet. Résiliation contrat

Références : N° de téléphone : 06 64 48 [redacted]
N° de client : 1. 4761154

Madame, Monsieur,

J'ai souscris, le 12 novembre 2001, un contrat
• " mini forfait 6 mois à 85 francs "
Ce contrat prendra fin le 02 avril 2002.
Veuillez donc trouver dans cette lettre, la preuve écrite qui mettra fin à la résilliation du présent contrat.

Pour plus d'informations, je reste à votre entière disposition.

Veuillez accepter de croire, Madame Monsieur, en ma considération la plus distinguée.

Fig. 1 An incoming mail document (for confidentiality reasons, personal information has been blurred).

This article is organized as follows. Section 2 is dedicated to the definition of the document categorization task. Section 3 presents the recognition system of omni-writer handwritten words, based on a lexicon directed analytical approach with an explicit segmentation. The keyword extraction task is studied in section 4, where two statistical models of handwritten lines are proposed: the first one is based on a full recognition (FR) strategy, whereas the second one is based on a shallow language model dedicated to keyword extraction (KE). Section 5 is devoted to the experimental results for the incoming handwritten mail categorization task. Conclusion and future works are drawn in section 6.

[1] see for example the historical documents such as the Georges Washington Collection at *http://memory.loc.gov/ammem/gwhtml/*

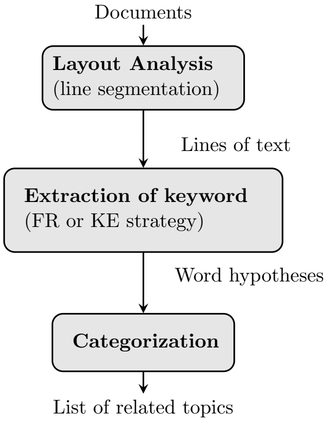


Fig. 2 Flowchart of the proposed system.

2 Categorization of electronic transcriptions of documents

This section is dedicated to the definition of the categorization task of textual documents, disregarding the writing recognition system performance and the difficulty in extracting textual information in the image of a handwritten document. In the first part, we review the concepts used in the field of electronic document analysis to describe and categorize these documents. In the second part, the retained approach is evaluated and compared to the literature on the *Reuters 21578* reference corpus, then on the specific categorization task of incoming mail that we consider throughout this study.

2.1 Categorization of electronic documents

A document categorization system aims at detecting the topic addressed in the document through the examination of its textual information. It is a supervised classification task where each class is considered as a topic. To achieve such a task, one must first define a text characterization space (feature space) before defining a decision rule (of the classification system). This task has been widely studied in the literature for processing electronic documents. We can refer to [Sebastiani 02, Baeza-Yates 99] for a review of the possible techniques. One of the most effective approaches to characterize electronic documents is based on a vector model of documents known as "bag of words". This description is generally coupled with learning techniques such as neural networks. Like all pattern recognition systems, a document categorization system consists in three main sequential steps following the traditional diagram on Fig. 3. In the following paragraphs, we describe each step.

2.1.1 Preprocessing

The first step consists in eliminating all undesired characters. We have chosen to accept only alphabetical characters and to eliminate all the others. Following this, empty words are filtered. This is performed using a list of words considered as the most frequent empty words (571 empty words for English and 463 for French). Finally, stemming (suffix elimination) is traditionally carried out using Porter's algorithm [Porter 80]. The French version of the algorithm differs from the English version only by the set of rules used. We chose to implement the Tf.Idf feature, which is easy to compute and provides very good results in practice [Aas 99, Baeza-Yates 99, Salton 88]. The weight w_{ij} assigned to the term t_i of the document d_j is defined by the following expression:

$$w_{i,j} = \mathrm{tf}_{i,j} \times \mathrm{idf}_i = \frac{\mathrm{freq}_{i,j}}{\max_k \mathrm{freq}_{k,j}} \times \log \frac{N}{n_i}$$

Where:

- N is the number of documents in the database
- n_i is the number of documents in which the term t_i appears
- $\mathrm{freq}_{i,j}$ is the number of occurrences of the term t_i in the document d_j

Each document is represented by a vector of weights $w_{i,j}$. The dictionary to refer to for the construction of the feature vector is determined during the learning stage of the system by performing feature selection.

2.1.2 Feature selection

It is generally necessary to reduce the size of the feature space because preprocessing produces a high dimensional description that consists in several thousands of words. Three common approaches are reported in the Information Retrieval domain [Yang 97]. They are based on frequency threshold, information gain and the χ^2 measure. For comparison purpose with other studies [Joachims 98] we choose to use the χ^2 measure.

Let the categories of the K documents be denoted $c_1, c_2, \ldots, c_K$; the probability $P(c_i)$ of category c_i is estimated as the ratio of documents in the database that belong to the class c_i and $P(t)$ is estimated as the ratio of documents that contain term t. It follows that $P(c_i, t)$ can be computed by the fraction of documents of the class c_i containing the term t. Similarly $P(c_i, \bar{t})$ is the fraction of documents of the class c_i that do not

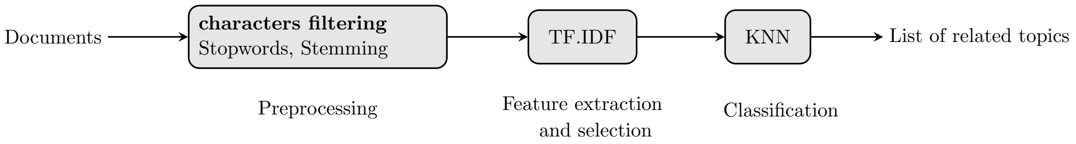


Fig. 3 Categorization of electronic documents.

contain the term t. The χ^2 measure is the correlation between a term t and a category c, computed as follows:

$$\chi^2(t,c_i) = \frac{N \times \left[P(t,c_i)P(\bar{t},\bar{c}_i) - P(\bar{t},c_i)P(t,\bar{c}_i)\right]^2}{P(c_i)P(\bar{c}_i)P(t)P(\bar{t})}$$

Usually, two scores are used for feature selection based on this measure, they are:

$$\chi^2_{\mathrm{mean}}(t) = \sum_{i=1}^{K} P(c_i)\chi^2(t,c_i)$$

$$\chi^2_{\mathrm{max}}(t) = \max_{i=1}^{K}\left(\chi^2(t,c_i)\right)$$

The final list of retained terms is composed of those with the N best scores.

2.1.3 Classification

The classification of a document in a topic can be performed using different classification methods. K nearest neighbors, neural networks, SVM, are some of the most popular approaches [Sebastiani 02, Joachims 98, Vinciarelli 05]. In this study, we retained a K nearest neighbor classifier for its simplicity and performance. We used the classical "cosine" similarity measure that is the most popular metric defined by the following relation:

$$\mathrm{sim}(q,d_j) = \frac{\overrightarrow{d_j}\overrightarrow{q}}{|\overline{d}_j| \times |\overline{q}|} = \frac{\sum_i w_{i,j} \times w_{i,q}}{\sum_i w_{i,j}^2 \times \sum_i w_{i,q}^2}$$

Where d_j and q stand for the vector representations of respectively the document d on the learning database, and the query document q to be categorized.

2.2 Evaluation

The *Reuters 21 578* corpus [Lewis 92] is used to validate the methodology by comparing the performance with those reported in the literature. Then it was possible to evaluate the incoming mail document categorization task using the electronic transcriptions of each document with the same system. This experimentation allows the determination of the optimal performance that we expect to achieve on the handwritten documents.

2.2.1 Reuters 21578 corpus

This widely used corpus is a set of 21 578 articles annotated according to their topic, among nearly 120 topics. The topic distribution is unbalanced: some topics are represented by over 3 700 articles, while some others are represented by less than 50. A protocol (modApte) describes how to split the database into a learning set and an evaluation set. This gives 7 063 documents in the learning set and 2 742 in the evaluation set. After having carried out the preprocessing of the learning database, a vocabulary of 15 453 terms is obtained.

The various parameters of the system are first optimized using the learning set. They are reported in Table 1 and compared with the values reported in [Joachims 98]. We observe a slight difference between the parameter values, which is certainly due to the character-filtering step.

Parameters	Joachims98	Our approach
KPPV	K=30	K=30
# of terms before selection	9947	6347
Measure of selection	χ^2	χ^2
# of terms after selection	1000	1000
Minimum # of doc. per class	3	2

Table 1 Characteristics of Joachims's categorization system and our categorization system

Table 2 presents the results obtained with our categorization system as well as those obtained by [Joachims 98]. For each of the ten most frequent topics, the Break-Even-Point (BEP) is reported. This value is obtained when recall equals precision. Let us recall that recall is similar to a detection rate, and precision is similar to a pertinence rate. More formally, one can define recall and precision as:

$$\mathrm{recall} = \frac{tp}{tp + fn} \quad \text{and} \quad \mathrm{precision} = \frac{tp}{tp + fp}$$

where tp, fp and fn stand respectively for *true positive*, *false positive* and *false negative* rates. The micro average measure corresponds to the computation of the global BEP. This unique measure allows summarizing the system performance by one single measure. The results obtained are very similar to those presented in [Joachims 98], and thus validate our categorization approach.

topic	# of samples per topic	Joachims98 (BEP)	Our approach (BEP)
earn	1044	97.3	96.9
acq	643	92.0	92.2
money-fx	141	78.2	78.7
grain	134	82.2	84.6
crude	161	85.7	83.5
trade	113	77.4	78.2
interest	100	74.0	74.4
ship	85	79.2	82.8
wheat	66	76.6	62.1
corn	48	77.9	73.7
Micro average		82.3	81.9

Table 2 Break Even Point (BEP) of Joachims's categorization system and our system on the Reuters 21578 corpus.

2.2.2 Incoming mail corpus

Performance of the categorization system is now determined on the incoming mail corpus. We use the ground truth of a handwritten mail corpus made of the electronic transcription of each handwritten document. These mails are classified into 43 topics: "standard cancellation" (A500), "changing of bank address" (A020), etc. The topic A500 ("standard cancellation") contains over 300 documents whereas topic A020 ("changing of bank address") contains only 30 documents. Note that the annotation of topics comes from a real-world database of handwritten mails addressed to a french company. The learning set contains two thirds of the documents of each category and the evaluation set contains the remaining third.

The optimal parameters of the system are as follows. The dictionary is composed of nearly 7000 terms. The lexicon is reduced to 980 words by removing the words that appear in less than five documents and using the χ^2 measure for term selection. This leads to retain 450 discriminative terms after stemming. Classification is carried out using a 5-nearest-neighbor classifier. Table 3 reports the results obtained on the 5 most frequent classes. The micro-average is computed using all the categories, which allows a global evaluation of the system. It appears that some topics are more difficult to model than others. The "information requests" (A240 and A255) are not as well classified as "cancellations" (A500 and A502). This can be due to more variability in this class. In fact, the two "cancellation" topics are very well defined whereas the "information requests" are more heterogeneous. With an equivalent amount of samples, it is not surprising that topic A502 is better recognized than topic A240. Finally, these results highlight the optimal categorization performance that can be expected on this particular corpus assuming that perfect recognition of the informative handwritten keywords can be achieved. The following section will now consider the adaptation of this categorization system for handwritten documents.

topic	BEP	#
A500 (cancellation)	86.6	206
A255 (info account/service)	62.7	26
A020 (change bank address)	71.1	23
A030 (loading post address)	87.9	21
A240 (claim / info fact)	25.8	16
A502 (cancellation with portability)	43.5	12
micro-average	76.6	

Table 3 Categorization results obtained by the annotation on the basis of incoming handwritten mail.

3 Recognition of handwritten words

Despite the success of some very specific industrial applications such as the reading of postal addresses or bank checks, off-line handwriting recognition remains an open problem.

From a methodological point of view, one can distinguish two major approaches in the literature [Plamondon00, Wang00, Koerich03, Lorette07]. The lexicon directed approaches, where the recognition process takes its decision at the lexical level only by discriminating the words that belong to the lexicon. The lexicon free approaches, where the decision comes at the character level. In this case the lexicon is used in the post-processing phase to correct the character recognition errors. Beyond the lexicon aspects, we can distinguish two main categories. The first one refers to the holistic approaches that consider the word as an indivisible entity. Words are recognized using global features extracted on the whole shape of the word. This kind of approach depends on a static, and often small lexicon. Note that word spotting approaches generally fall into this category. The second category of approaches refers to the analytic methods, where words are recognized through their constitutive characters [Kim97, ElYacoubi99, Wang00, Vinciarelli00, Rodriguez08, Graves09]. Within this framework it is possible to model any word and thus any lexicon during the recognition phase. Among the analytic approaches, we can then distinguish the implicit and explicit segmentation approaches. Explicit segmentation methods introduce a segmentation stage which proposes several character segmentation hypotheses. These hypotheses are then validated by the recognition stage. Inversely, implicit segmentation approaches do not introduce any complex (adhoc) segmentation stage and they let the recognition process find the best segmentation into characters. Most of the recent approaches fall into this last category of methods by relying on the Hidden Markov Models [Grosicki09], including Vinciarelli's work for noisy text categorization. This success is mainly due to the relative ease of implementation of the approach, as opposed to the explicit segmentation, even if one major well-known drawback of Hidden Markov Models is their low capacity to discriminate between classes. In fact, Hidden Markov Models are generative models which are trained class by class by maximizing the likelihood of each training dataset (one per class). Some recent approaches have therefore propose the use of recurrent networks [Graves09] that use discriminative training to overcome this drawback, and this has proven to be efficient.

In this work, the handwritten word recognizer uses a lexicon driven analytical approach with explicit segmentation derived from [Koch04] with discriminative training of character models. Considering the state of the art in handwriting recognition the proposed approach combines the strength of discriminative training with a limitation due to the segmentation stage. We briefly present the word recognition system, and refer to the aforementionned paper for more implementation details.

A first preprocessing step is carried out on the binary word images and allows the reduction of writing variability using slant and skew corrections inspired by [Kimura94]. The segmentation step splits the images into informative zones known as graphemes. The graphemes, or groups of graphemes, are then submitted to a character recognizer. Finally, the word hypotheses are built by the exploration of the segmentation lattice.

The segmentation stage generates hypotheses of segmentation points through the analysis of the word contour: each local minimum and maximum of the upper contour of the word is considered as a potential segmentation point [Nosary 02].

For each word, a segmentation lattice is built (see Figure 4), containing elementary graphemes at level 1, and all possible aggregations of n adjacent elementary graphemes at level n. The segmentation hypotheses of the first level are likely to be oversegmented, whereas hypotheses of the last levels are likely to be undersegmented. The segmentation statistics demonstrate that the distribution of the number of graphemes is strongly unbalanced depending on the character classes. For example, character $'c'$ is frequently segmented into only one grapheme, whereas character $'m'$ is frequently segmented into 5 graphemes. We also observed that the maximum number of levels needed was 7 to prevent from under segmentation. In order to benefit from this *a priori* knowledge, we have chosen to model the segmentation process by a duration statistical model, presented in the following paragraph.

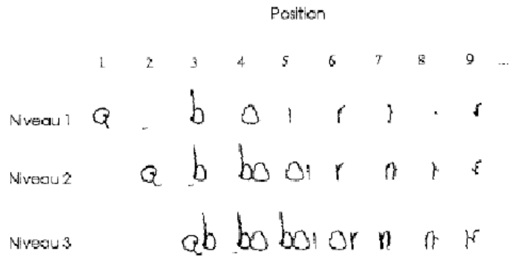


Fig. 4 Illustration of the segmentation lattice, where level n contains all the possible aggregations of n adjacent elementary graphemes.

In order to find the best segmentation path in the lattice, each aggregation hypothesis is submitted to a character recognizer which aims at providing the *a posteriori* probability of the character classes $\{a, b, \ldots, z\}$. To estimate these probabilities, several information sources are combined according to the diagram given on Figure 5.

Two classifiers are built to exploit complementary information on each grapheme, at each level. On one hand structural/statistical information such as curvatures, occlusions, horizontal and vertical strokes are coded into a 117 feature set according to [Heutte 98]. On the other hand, directional information on the contours is coded into a 128 feature set according to [Kimura 94]. These two information sources are exploited by two multi-layer perceptron (MLP) classifiers [Bishop 95], called MLP-117 and MLP-128. MLP-117 and MLP-128 produce *a posteriori* probability estimates[2] of the character classes $P(L_i/O_1)$ and $P(L_i/O_2)$, where O_1 and O_2 stand for the two feature vectors. We refer to [Koch04] for more details concerning the production of the *a posteriori* probability $P(L_i/O)$.

[2] See [Richard 91] for the proof that a MLP generates approximations of *a posteriori* probability outputs. In practice, the MLP outcomes are normalized using a softmax function.

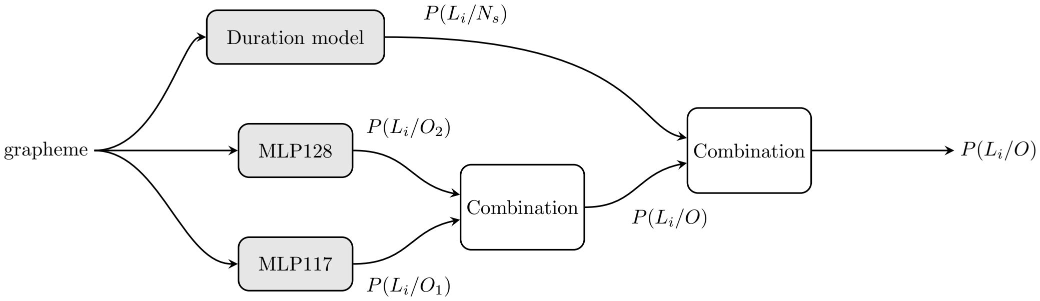


Fig. 5 Architecture of the information combination in the character model.

A statistical duration model then combines the segmentation information with the character hypothesis. The character distribution over the different segmentation levels is exploited during the recognition step. Let N_s be the number of segments of a character, the term $P(L_i/N_s)$ is estimated by counting the number of character samples on the learning database that occur on a particular number of segments N_s.

The final word recognition stage is performed through the exploration of the recognition lattice using dynamic programming. The introduction of lexical constraints at this stage reduces the number of solutions during the exploration. The lexicon is modeled by an automaton λ with N character states, such as the one shown in Figure 6.a . The complexity of this algorithm is of order $(\text{max-levels} \times N)^2 \times T = (7 \times N)^2 \times T$, where T is the length of the lattice. As the complexity is a function of N^2, a reduction of the number of states will have a large influence on the computation time. This can be done by adopting a tree-structured representation of the lexicon. For example, the model presented in Figure 6.a can be reduced to the one in Figure 6.b. In this example, the number of states can be reduced from 41 to 27. The complexity is then reduced by a factor of 2.3 whereas the number of states is only reduced by a factor of 1.5.

Evaluation

The word recognition engine is evaluated using a learning database of 4600 words and a test database of 500 words, all coming from real incoming handwritten mail documents. Table 4 gives the word recognition rates for MLP-128, MLP-117 and the average combination of the two MLPs. Results are presented with and without considering the character duration model, and for different lexicon sizes N by randomly selecting $N-1$ words among a 1400 word lexicon (complete lexicon of the word database). We can observe that whatever the size of the lexicon, the combination of the two MLPs improves the word recognition performance significantly. These results also bring out the relevance of the character duration model. Figures 7 and 8 show some examples of properly recognized words and misrecognized words from different writers. Our results appear to be fairly lower than state-of-the-art approaches such as [Kim97, Koerich05, Zimmermann06], but our real-world database exhibits multiple significant degradations due to: (i) low resolution (200 dpi) (ii) industrial digitizing stage (iii) strongly heterogeneous writing styles (see examples of Figure 7). Let us also notice that our system was probably trained using less data than in some other studies.

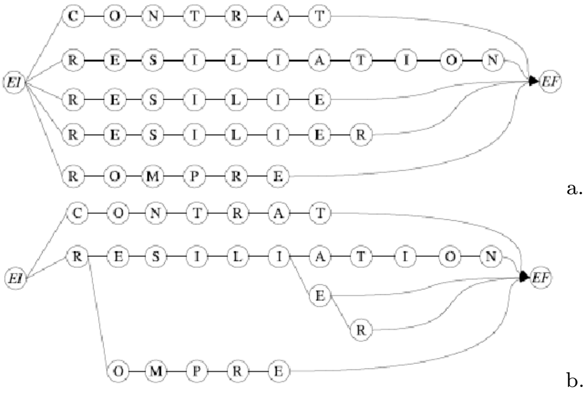


Fig. 6 a) Flat lexicon and b) its tree-structured representation.

4 Keyword extraction in handwritten documents

As presented in section 2, the categorization of handwritten documents is based on a word vector model of discriminative keywords. We must therefore highlight that the main objective of the handwritten word recognition system is to detect and recognize these relevant keywords. As opposed to the full recognition of handwritten content, some studies focus on the detection of

Lexicon 100/1000 words

	Classifier $P(L_i/N_s)$	Top 1	Top 2	Top 5	Top 10
MLP-128	without duration	71.8/42.4	82.2/57.8	92.2/69.2	96.6/78.0
	with duration	80.2/58.0	88.0/68.2	94.2/80.6	97.8/87.2
MLP-117	without duration	79.6/49.0	88.4/64.0	94.0/78.2	96.0/85.2
	with duration	81.6/57.0	90.0/69.2	93.6/80.4	96.8/86.8
Average	without duration	85.0/58.6	92.2/72.2	96.0/84.8	98.2/91.2
	with duration	85.6/65.6	92.8/76.8	97.0/86.8	99.0/91.6

Table 4 Word recognition performance for different configurations of the character recognition engine, for a lexicon size of 100/1000 words.

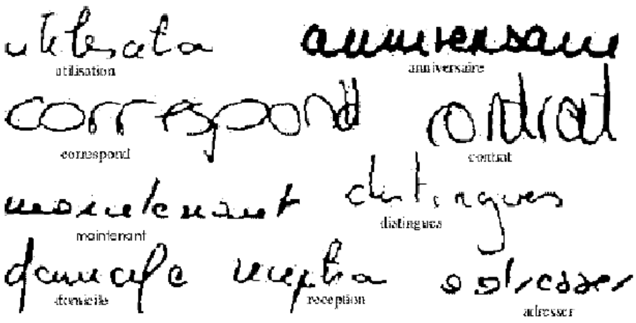


Fig. 7 Examples of correctly recognized words.

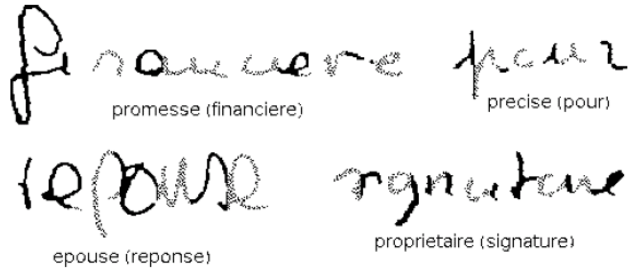


Fig. 8 Examples of mis-recognized words (correct labels within brackets).

keywords that are useful in indexation or categorization tasks. The basic idea lies in the fact that a major part of the information contained in a document is useless to capture its overall meaning, e.g., empty or stop words. This strategy known as keyword spotting has been first proposed for printed documents. It became popular in the handwriting recognition community for querying databases of digitized historical documents, for instance the Georges Washington's manuscripts [Rath07]. Two different approaches can be distinguished depending on the nature of the documents considered. On the one hand, template-based methods try to match image queries with pre-labeled segmented word image templates [Gatos05,Terasawa09, Belongie02. This kind of approach is restricted to querying mono-writer document databases. On the other hand, recognition-based approaches allow to work on more heterogeneous data (from different writers for instance). The recognition process involves a classification stage either as a holistic process [VanDerZant08,Rodriguez09] or as an analytical process involving character models [Rodriguez08,Koerich04]. A post-processing stage working on the recognition scores is generally needed to reject false hypotheses. Obviously, this second approach is also subject to limitations : it is assumed that word boundaries are known (line segmentation issues are avoided) and rejection is often carried out using a simple threshold on normalized scores.

In this article, we introduce a general line model for the extraction of keywords. This analytical model take account of keywords and out of keywords vocabulary. It also introduces an space model between words that allow line segmentation into words. This general model can be parametrized by any keyword lexicon and does not require any specific training when using a new keyword lexicon. This stochastic line model allows keyword detection, line segmentation and out of vocabulary word detection in a combined manner by using a dynamic programming decoding algorithm of each line of text. Two statistical models of handwritten lines are proposed in order to proceed to keyword spotting. Both of them rely on dynamic programming and integrate an inter-word space model within the line. They differ in the lexicon they use. The first one corresponds to the full recognition (FR) of the documents using a large lexicon (several thousands of words). The second one is based on a shallow language model dedicated to keyword extraction (KE). It is composed of a lexicon of relevant keywords and a stochastic bi-gram model of characters that accounts for irrelevant words. The two recognition strategies (FR and KE) are evaluated for their capacity to extract the relevant keywords in the handwritten documents as defined in section 2.

4.1 Full Recognition Model (FR)

Following the notations of section 3, we consider that each text line is composed of an observation lattice. The recognition of a text line consists in finding the best path in this lattice using dynamic programming and verifying the constraints of the considered model of the line. In this FR strategy, we consider that a line of text only contains words from the lexicon that are separated by an inter-word space. The line model is depicted in Figure 9 where state BL refers to the beginning of the line, state EL refers to the end of the line and state IW refers to the inter-word space state.

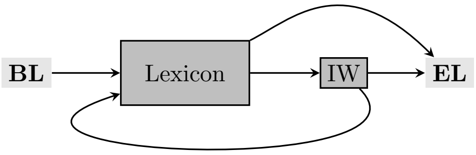


Fig. 9 Line model of the Full Recognition (FR) strategy.

We must highlight that the observation lattice contains the hypotheses of the 7 levels of segmentation as described in section 3, but also the observations characterizing spaces between connected components. The joint probability of the observation lattice and the word sequence can be decomposed according to the following relation:

$$P(O,Q^*) = \prod_{i=1}^{N} P(O^*_{M_i}/M_i)P(o^*_{M_iM_{i+1}}/IW) \times \prod_{i=2}^{N} P(M_i/M_{i-1})$$

Where:

- $o^*_{M_iM_{i+1}}$ is the observation attached to space between word i and word j in the optimal path
- $O^*_{M_i}$ is the observation associated to word i in the optimal path
- N is the number of words in the optimal match
- Q^* is the optimal state sequence
- M_i is the model of the i^{th} word
- IW is the inter word space

This expression can further be decomposed by rewriting the word likelihood of the optimal path. It finally yields:

$$P(O,Q^*) = P\left(o^*_{M_iM_{i+1}}/IW\right) \prod_{i=2}^{N_m} P\left(M_i/M_{i-1}\right) \times \prod_{i=1}^{N_m}\left(\prod_{j=1}^{length(M_i)} P\left(o^*_{i,j}/q^*_{i,j}\right) P\left(o^*_{i,j,j-1}/EL\right) P\left(q^*_{i,j}/q^*_{i,j-1}\right)\right)$$

Where

- $q^*_{i,j}$ is the j^{th} character in the word i
- $o^*_{i,j}$ is the observation associated to the j^{th} character of word i
- $o^*_{i,j,j-1}$ is the observation corresponding to the space between character j and character $j-1$.
- EL is the state representing an inter-character space in a word

We must notice that in the lexicon directed strategy, the character transition probability is equal to 1 if the transition belongs to the automaton (the transition belongs to a word of the lexicon), and to 0 otherwise. Furthermore, if we do not use a language model, the equation simply reduces to:

$$P(O,Q^*) = P\left(o^*_{M_iM_{i+1}}/IW\right) \prod_{i=2}^{N_m} P\left(M_i/M_{i-1}\right) \times \prod_{i=1}^{N_m}\left(\prod_{j=1}^{length(M_i)} P\left(o^*_{i,j}/q^*_{i,j}\right) P\left(o^*_{i,j,j-1}/EL\right)\right)$$

This probability is computed using dynamic programming in the same way as for the recognition of isolated words.

4.2 Keyword Extraction Model (KE)

The main objective of this second model is to limit the size of the vocabulary to the relevant keywords only. We expect to benefit from the reduced size of the keyword vocabulary so as to improve the recognition performance of the relevant information. We must however build a model of the irrelevant information so as to consider the set of all the other words that are irrelevant for the categorization task. This later model will act as a rejection model for the recognition system. It is a model of irrelevant words or *out of keyword vocabulary* words. Similar approaches have been proposed for speech processing [Yazgan 04]. We can consider that a line of text is a sequence of relevant and irrelevant words. These words are naturally separated by a space. Figure 10 illustrates the line model used by the KE strategy. This Figure highlights the competition of the relevant lexicon model developed in the FR strategy with the model of irrelevant words that we clarify now.

The irrelevant lexicon is composed of many words because it is potentially made of all the words of the language except the relevant keywords. We have chosen to use a character bi-gram stochastic model. More precisely, this model is composed of 28 states among

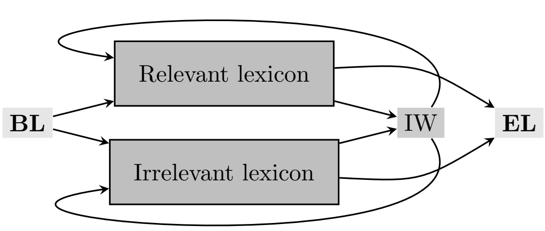


Fig. 10 Line model in the keyword recognition strategy (KE)

which 26 states correspond to characters. In addition, an initial and a final Non-Lexicon states are considered, modelling the beginning and the end of out of vocabulary words. Probabilities of bi-grams of characters can either be all equal, or determined on a set of examples. We can now clarify the expression of the joint probability of the observation lattice and the best word sequence conforming to this KE model. At first, let us consider the observation sequence that corresponds to the i^{th} word in the observation sequence. Its likelihood is expressed in two different ways depending on whether word i is part of the relevant lexicon or not.

$$P(O^*_{M_i}/M_i) = \max \begin{cases} P(O^*_{M_i}/M_i) \in \text{Relevant lexicon} \\ P(O^*_{M_i}/M_i) \in \text{Irrelevant lexicon} \end{cases}$$

Let us define:

$$\beta = P(M_i \in \text{Irrelevant lexicon})$$

Then the joint probability of the best observation sequence on a whole line is written as follows:

$$P(O, Q^*) = \prod_{i=1}^{N_m} \max \left\{ \begin{array}{r} (1-\beta)\times\ P(O^*_{M_i}/M_i \in \text{Relevant lexicon}) \\ \beta\times\ P(O^*_{M_i}/M_i \in \text{Irrelevant lexicon}) \end{array} \right\}$$
$$\times P(o^*_{M_i M_{i+1}}/IW)$$

Once again, this quantity can be computed using dynamic programming on each of the observation lattices associated to each line of text. When $\beta = 0$ one can notice that the model implements the Full Recognition strategy. When $\beta = 1$, the model implements a lexicon free recognition strategy. The whole KE model can be viewed as a model that puts in parallel models of the relevant keyword lexicon with the out of vocabulary word model which simply acts as a rejection model. The implementation of these two keyword extraction strategies (FR and KE) is depicted in the next subsection. Experimental results are presented in section 5.

4.3 Description of the keyword extraction system

At first, layout analysis allows the segmentation of the document into lines of text. Once layout analysis has been carried out, additional preprocessing steps help the recognition process (slant correction, diacritic filtering). Derived from these pre-processing steps, each line of text is represented by an observation lattice. All these pre-processing steps have been described in detail in section 3 and they are directly applied to the set of detected lines. In the following paragraph we give some details concerning layout analysis and the detection of possible word separators within lines of text.

4.3.1 Layout analysis

The line segmentation process is an important and difficult task, mainly due to variable skew angle along the document or even along a text line, and adjacent text lines. The recent handwriting segmentation contest [Gatos2009] has shown that many strategies perform well, such as run length analysis [Shi09], function minimization exploiting the distance between the separators and the local foreground density [Stafylakis08] or connected component bounding box analysis [Yin08]. We have chosen a connected component based approach, which has shown to perform well on real-world, free layout documents [Chatelain 06]. We now briefly describe this approach.

Lines of text are formed by successive merging of connected components based on a distance criterion. It is implemented in three steps after the detection of connected components (Figure 11). The first step detects alignment seeds. Then each alignment seed is extended to its nearest connected component using the following pixel distance:

$$d(a, b) = (x_a - x_b)^2 \times \alpha(y_a - y_b)^2$$

Where a and b are connected components and x and y are their centers of gravity. The parameter α allows to weight the horizontal ($\alpha > 1$) or vertical direction ($\alpha < 1$). The value of α has been experimentally fixed to 20. To build these alignments, only the most representative components (of sufficient size) are considered.

Segmentation results are reported in Table 5. A line is considered as being well segmented if all its components are grouped together. Three types of errors are counted. Over-segmentation is counted if the line is segmented into several alignments. We can observe that nearly 80% of the lines are properly segmented. Concerning segmentation failure (over and under segmentation), only one connected component (therefore one or two words) is often involved. In this situation,

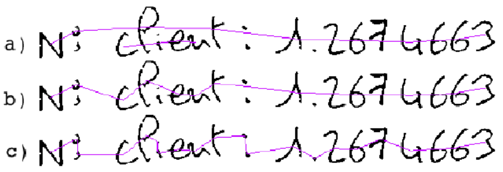


Fig. 11 The three steps of line segmentation. a) Initial grouping of the largest components; b) fusion of alignment; c) aggregation of small components.

the other words of that line can still be recognized. We estimate that nearly 90 to 95% of the document words can be processed following this line segmentation process. The performance of document segmentation into lines of text is far from being perfect but seems sufficient to apply the two keyword extraction strategies that we have presented above.

	Number of lines	%
Well segmented	262	79
Over segmented	29	9
Under segmented	0	0
Other error	42	12
Total	333	100

Table 5 Evaluation of the line segmentation process.

4.3.2 Estimating inter-word and inter-character space probabilities

Spaces between two consecutive components of a line are assigned to the corresponding class (IW or IC). The measure is carried out using minimal Euclidean distance [Seni 94]. To eliminate the variability between different writers, the distance is normalized in reference to the median value of the width of the elementary graphemes. Two normalized distributions are obtained. Finally, the probability of an Inter Word space having a distance d is given by the equation below:

$$P(IW/d) = \frac{\#IW(d)}{\#IW(d) + \#IC(d)}$$

Where $\#IW(d)$ (respectively $\#IC(d)$) matches the proportion of Inter-Word spaces (similarly Inter-Character spaces) that have a distance equal to d. We determine the same probability of inter-character spaces:

$$P(IC/d) = 1 - P(IW/d)$$

The distributions of these two probabilities are represented on Figure 12. These two a posteriori probabilities are considered as likelihood scores in the observation lattice. The integration of these observations is done with straightforward modifications of the observation lattice.

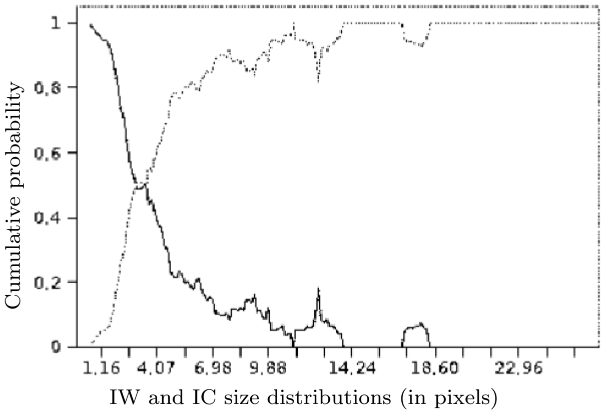


Fig. 12 Distribution of the inter-word space (IW) probability and the inter-character (IC) space versus the distance between components.

5 Experimental Results

In this section we present the keyword extraction system performance for both FR and KE strategies.

5.1 A document database for evaluating FR and KE strategies

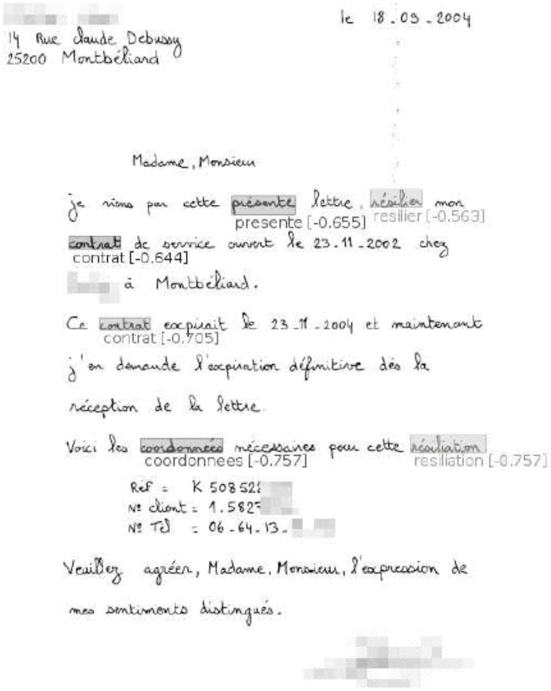
A database of Incoming Handwritten Mails has been built for this purpose. A set of 1100 real documents have been scanned with a resolution of 200dpi, where the words of text bodies have been manually labeled. Of course, the word database used for training and testing the word recognizer has been design using a different set of documents than the document database. The 1100 documents are made of over 46000 word samples that compose a lexicon of 3700 distinct words. As this is an industrial database coming from real customers, it can not be diffused and personnal information have been hidden for this article. Exemple of documents can be found in appendix 6. One can note the presence of noise due to imperfect numerization and binarization. This noise has been deleted using standard and simple operations such as filtering of too small connected component and mathematical morphology.

5.2 Performance evaluation of the keyword extraction system

Let us remind that the first strategy for keyword extraction consists in carrying out the full document recognition (Full Recognition) then to retrieve keywords on the basis of the recognition results. The capacity of the system to detect keywords is directly related to the performance of the text recognition system that works with a large lexicon. For this experimentation, the entire lexicon composed of the 3700 different entries of the test database is used. In order to depict the recall precision curve, a variable threshold is applied on the recognition scores of the word recognition hypotheses. The score of the words is calculated by averaging the output score of the neural network classifier, and is normalised according to the length of the word.

The second strategy operates with a lexicon of relevant keywords and uses the particular strategy developed in section 4.1. This lexicon has a reduced size which can vary from 46 words up to 980 words. The irrelevant lexicon is modeled in our experimentation using a uniform ergodic stochastic model where parameter β varies between 0.01 up to 0.99 so that the whole recall precision curve can be explored. In the various experimentations, the relevant lexicon used is defined as "KE n" where the value n is the number of words in the keyword lexicon. Figure 5.2 illustrates the results obtained by keyword extraction on an incoming mail.

Figure 14 presents the set of results for the two strategies examined. These results illustrate the superiority of the keyword extraction strategy as compared to the full recognition strategy and whatever the keyword lexicon. The FR strategy provides 22% recall for 50% precision while working with a 3700 word lexicon. On the other hand, with a 295 keyword lexicon (lexicon "KE 295") the performance is 36% recall with 70% precision. We can see that the performance is rather stable when increasing the lexicon size from 46 to 295 keywords (with the "KE 46", "KE 96", "KE 165" and "KE 295" lexicons). These various lexicon sizes correspond respectively to 10,25,50 and 100 radicals. These results clearly highlight the contribution of the relevant keyword extraction strategy that allows focusing the recognition system only on the keyword lexicon while modeling irrelevant information by an ergodic character model.


[redacted] le 18.09.2004
14 Rue Claude Debussy
25200 Montbéliard

Madame, Monsieur

je viens par cette présente lettre, résilier mon contrat de service ouvert le 23.11.2002 chez [redacted] à Montbéliard.

Ce contrat expirait le 23.11.2004 et maintenant j'en demande l'expiration définitive dès la réception de la lettre.

Voici les coordonnées nécessaires pour cette résiliation

Ref = K 50852[redacted]
N° client = 1.582[redacted]
N° Tél = 06.64.13.[redacted]

Veuillez agréer, Madame, Monsieur, l'expression de mes sentiments distingués.

[redacted]

Fig. 13 Example of the keyword extraction result on an incoming mail document.

5.3 Categorization of handwritten documents

In this section we analyze the interaction of the two systems (keyword extraction and document categorization) in order to optimize the overall performance of the handwritten document categorization. First we compare categorization performance for both FR and KE strategies. Then the role of the keyword lexicon is analyzed in depth regarding both the keyword recognition performance and the categorization performance.

Table 6 presents the categorization results obtained on the database of incoming handwritten mails following keyword extraction (KE 980). A global "BEP" of 62.3% is obtained despite relatively low performance of the keyword extraction system (27% recall and 57% precision). Compared to the ideal categorization system (TRANS) using the ground truthed document transcription, categorization performance only degrades by 14 points. The FR strategy gives an overall BEP of 46.4% which is very low compared to the 62.3% BEP obtained with the KE strategy.

Figure 15 allows a finer analysis of the results. Figure 13 clearly demonstrates that the KE strategy outperforms the FR strategy when considering the categorization task. The breakeven point only decreases by 14 points when using the KE strategy while the FR strategy loses 30 points. These results confirm the robustness of the categorization system with respect to keyword extraction errors. They also highlight the relevance of the proposed keyword extraction as opposed to the Full Recognition strategy. In the following sec-

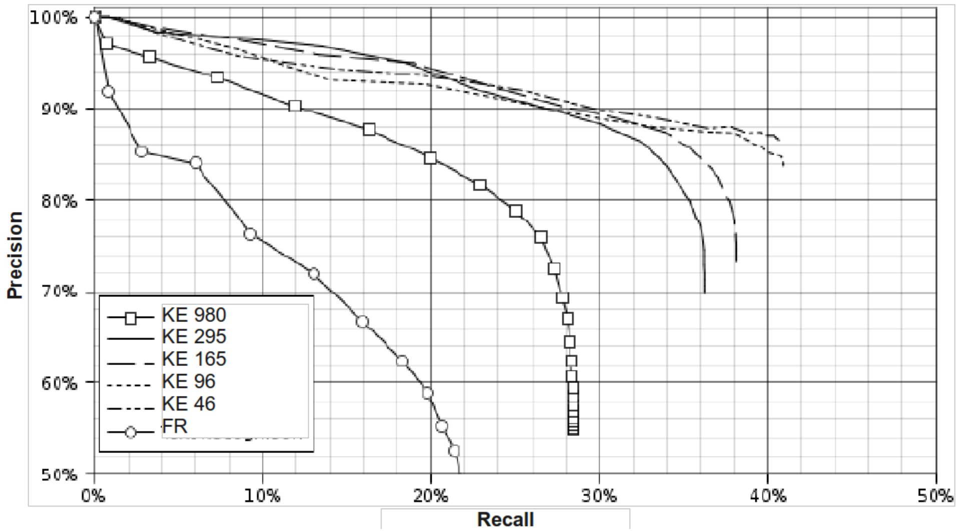


Fig. 14 Keyword extraction performance. "KE n"stands for keyword extraction with a lexicon of n keywords; "FR" refers to the Full Recognition strategy with a 3700-word lexicon.

topic	# of document	TRANS	KE 980	FR
A500 (cancellation)	206	86.6	73.5	68.1
A255 (info pass / serv)	26	62.7	34.6	27.6
A020 (change bank address)	23	71.1	39.1	32.3
A030 (change postal address)	21	87.9	48.7	42.8
A240 (claim / info fact)	16	25.8	0.0	0.0
A502 (cancel. with portability)	12	43.5	25.0	19.5
micro-average		76.6	62.3	46.4

Table 6 Categorization break even points on ground truthed transcription mails (TRANS) and on recognized mails using the keyword extraction strategy (KE 980), and the full recognition strategy (FR).

tion, we study the influence of the lexicon size on the categorization performance.

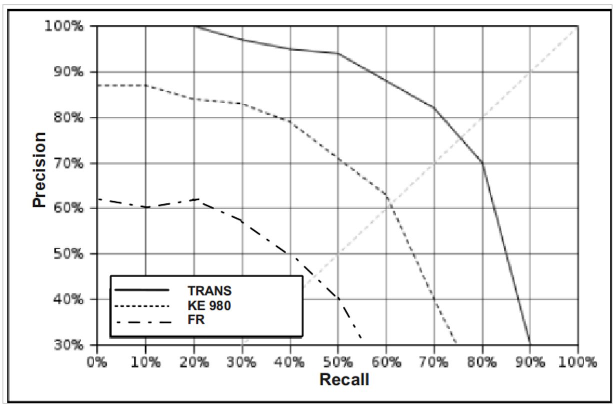


Fig. 15 Recall/Precision curves of the categorization task using ground truthed transcription (TRANS), Keyword Extraction strategy (KE 980) and Full Recognition strategy (FR).

5.4 Control of the lexicon size

As seen in section 5.2, the performance of the keyword extraction engine improves when the number of words in the lexicon decreases. Thus, a trade-off must be found between a reduction of categorization performance caused by an important lexicon reduction, and a keyword extraction performance gain. To find this trade-off, we can analyze the performance of the system on the ground truthed transcription database (TRANS) when the size of the lexicon decreases. To achieve this, we have used the categorization system configuration presented in section 2.2.2 (selection of 450 terms by the "Information Gain" measure, description of documents by the Td.Idf measure, classification carried out with 10 NN). The different lexicon sizes are determined on a learning database using the χ^2 measure. Figure 16 presents the categorization performance obtained for different lexicon sizes. We observe that a 75% reduction (from 450 to 100 terms (radicals)), results in a slight performance decrease. The lexicon is however reduced from 980 to 295 words (70% reduction). To observe a significant performance decrease, we must reach a much reduced lexicon (10 terms/46 words). To allow comparison, the performance obtained using the recognition system (KE 980) is recalled. We observe that it is also lower than the performance of different categorization systems constructed with the transcription of the database (TRANS), in a situation of perfect recognition.

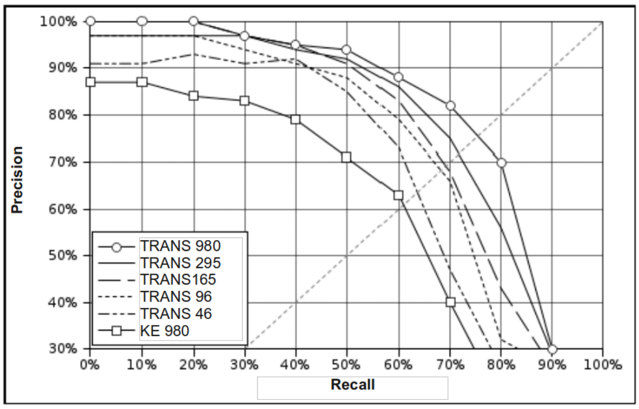


Fig. 16 Categorization performance on the annotated ground truthed corpus (TRANS) as a function of the keyword lexicon size and using KE strategy (KE).

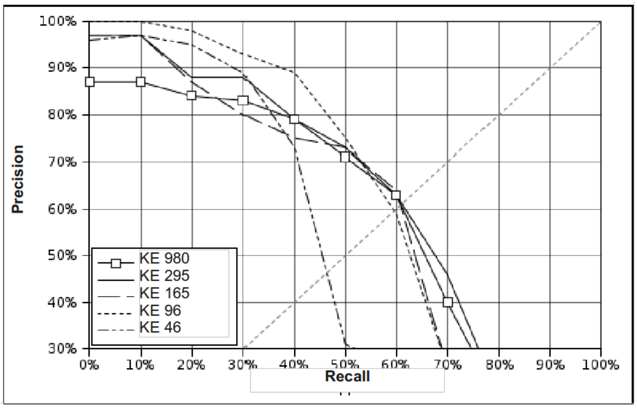


Fig. 17 Recall/precision performance in handwritten document categorization as a function of the keyword lexicon size using KE strategy.

Figure 17 presents the handwritten categorization performance using the KE strategy for the same lexicon sizes. Without modifying the Break Even Point, which remains at approximately 62% in all cases, we observe nonetheless that a reduction of the lexicon causes a strong improvement of categorization performance for all values on the curve inferior to the BEP. In fact, with 96 keywords, the performance reaches 90% precision for 40% recall, which is very close to the performance obtained on the ground truthed annotated database (TRANS 96) without recognition. We are in a situation where categorization performance are very similar to theoretical performance caused by a decrease of recognition errors due to the reduction of the lexicon size.

However, we observe a performance decrease when the lexicon used for categorization becomes too small (46 keywords). This seems natural because, as the lexicon is substantially reduced, a recognition error has more impact on the document description. We can conclude that the size of the lexicon is a key factor for handwritten document categorization. In our application, the best compromise seems to be a lexicon of 100 keywords. In fact, a more important reduction of the lexicon slightly improves the keyword extraction performance but it strongly reduces categorization performance.

Having only a single database of ground truthed annotated documents, it is rather difficult at this stage to bring a full answer concerning the optimal performance that can be expected in more general conditions and other categorization tasks. A similar study with documents from another area (other terms, other topics, etc.) should allow confirming these results. It seems, however, that below 30% recall for 70% precision in keyword extraction, the categorization performance drops quickly.

We have presented a complete system for handwritten document categorization based on a word extraction strategy rather than a full recognition approach. Apart from the experiments which shows that KE outperforms FR approach, we believe that the proposed modelisation also outperforms an FR strategy for the following reasons:

- In KE strategy, the OOV words are really modelled, whereas FR strategy does not.
- KE based on our line model is a dynamic approach able to take into account the lexicon size, the recognition, the rejection and the segmentation process, whereas in FR strategy the rejection is performed as a postprocessing stage using a threshold, which prevent from taking the best whole decision on the entire line of text.
- As the lexicon size is smaller, the KE strategy is faster than FR approach.

As a conclusion, let us emphasize that a perfect recognition would make our approach obsolete. Unfortunately recents systems are still far from having acceptable recognition results on weakly constrained handwritten documents.

6 Conclusions and future works

Building an automatic handwritten document categorization system calls upon techniques proposed in several areas of document analysis: automatic document layout analysis for detecting lines of text, handwriting recognition techniques for extracting keywords, and information retrieval for document categorization. For the first time, a complete categorization system of handwritten documents has been proposed with promising results on a real handwritten document database.

To overcome the limitations of a full recognition approach based on a large lexicon word recognition strategy, a new information extraction model has been designed, capable of locating and recognizing a restricted set of discriminative keywords. The information extraction method is based on an Out-Of-Vocabulary word model which is able to handle irrelevant information. We have shown in this study that it leads to better results than a more classical approach based on a full recognition strategy.

This first study raises a set of comments to be put in perspective for further studies concerning handwritten document processing. It is to be noticed that the keyword extraction strategy developed for document categorization could also be used for information retrieval. This study opens interesting prospects for future applications of the indexation of handwritten documents, regardless of the kind of documents to be processed: commercial documents, such as incoming mail or any other handwritten document with a certain interest like for example historical handwritten documents preserved numerically in digital libraries. No doubt that these topics will be addressed by many of the forthcoming researches led by the research community in the field of handwriting recognition.

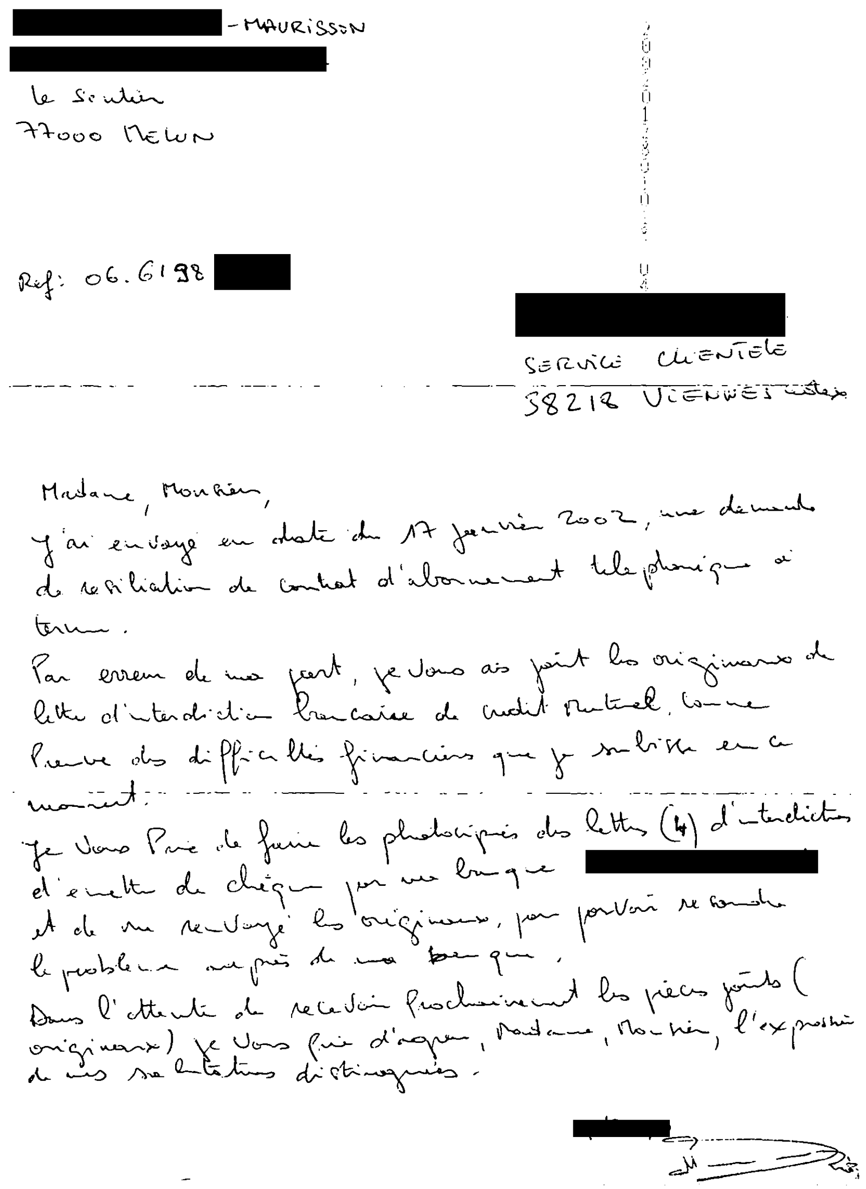
References

1. K. Aas, L. Eikvil. Text categorisation: A survey. Technical Report, Norwegian Computing Center, June 1999.
2. T. Adamek, N.E. O'Connor, N. Murphy, A.F. Smeaton, Word matching using single closed contours for indexing handwritten historical Documents, International Journal on Document Analysis and Recognition, Vol. 9, No. 2, pp. 153-165, 2007.
3. N. Arica, F.T. Yarman-Vural, "Optical character recognition for cursive handwriting", IEEE Trans. PAMI 2002, Vol. 24, No. 6, pp. 801-813, 2002.
4. R. Baeza-Yates and B. Ribeiro-Neto. Modern information retrieval, Addison-Wesley Longman Publishing Co., 1999.
5. S. Belongie, J. Malik, J. Puzicha, Shape matching and object recognition using shape contexts, IEEE Trans. on PAMI 24 (4) (2002) 509-522.
6. R. Bertolami and H. Bunke, Hidden Markov Model Based Ensemble Methods for Offline Handwritten Text Line Recognition, Pattern Recognition, vol. 41, pp3452-3460, 2008.
7. C.M. Bishop. Neural networks for pattern recognition, Oxford : Oxford University Press, 1995.
8. A. Brakensiek, J. Rottland, A. Kosmala, G. Rigoll, Offline Handwriting Recognition using various Hybrid Modeling Techniques and Character N-Grams, IWFHR'00, pp.343-252, Amsterdam, 2000.
9. H. Cao and V. Govindaraju, "Vector Model Based Indexing and Retrieval of Handwritten Medical Forms", Vol. 1, pp. 88-92, ICDAR 2007.
10. C. Chatelain, G. Koch, L. Heutte, and T. Paquet, Une méthode dirigée par la syntaxe pour l'extraction de champs numériques dans les courriers entrants, Traitement du Signal, vol. 23, iss. 2, pp. 179-198, 2006.
11. D. Doermann, The indexing and retrieval of document images: a survey. Computer Vision and Image Understanding, 70(3):287-298, 1998.
12. J. Edwards, Y. Whye, T. David, F. Roger, B. M. Maire, G. Vesom, Making latin manuscripts searchable using ghmms, In NIPS (2004) 385-392.
13. 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, vol. 21, no. 8, pages 752-760, 1999.
14. M.A. El-Yacoubi, M. Gilloux and J.M. Bertille. A Statistical Approach for Phrase Location and Recognition within a Text Line : An Application to Street Name Recognition. IEEE Trans. on PAMI, vol. 24, no. 2, pages 172-188, 2002.
15. B. Gatos, T. Konidaris, K. Ntzios, I. Pratikakis, S. J. Perantonis, A segmentation-free approach for keyword search in historical typewritten documents, ICDAR (2005) 54-58.
16. B. Gatos, N. Stamatopoulos and G. Louloudis, ICDAR2009 Handwriting Segmentation Contest, pp. 1393-1397, ICDAR 2009.
17. Alex Graves, Marcus Liwicki, S. Fernandez, Roman Bertolami, Horst Bunke, Jrgen Schmidhuber: A Novel Connectionist System for Unconstrained Handwriting Recognition. IEEE Trans. Pattern Anal. Mach. Intell. 31(5): 855-868 (2009)
18. R. Grishman, B. Sundheim: Message Understanding Conference - 6: A Brief History. In: Proceedings of the 16th International Conference on Computational Linguistics (COLING), I, Kopenhagen, 1996, 466-471.
19. L. Heutte, T. Paquet, J.V. Moreau, Y. Lecourtier and C. Olivier, A Structural/Statistical Feature Based Vector for Handwritten Character Recognition, Pattern Recognition Letters, vol. 19, no. 7, pp. 629-641, 1998.
20. S. Impedovo, P.S.P. Wang, H. Bunke, Automatic Bankcheck Processing, S. Impedovo, P.S.P. Wang, H. Bunke eds., Series in Machine Perception Artificial Intelligence, World Scientific, Vol. 28, 1997.
21. T. Joachims. Text categorization with support vector machines : learning with many relevant features. In Claire Nedellec and Celine Rouveirol, editors, Proceedings of ECML-98, pp. 137-142, 1998.
22. G. Kim and V. Govindaraju. A Lexicon Driven Approach to Handwritten Word Recognition for RealTime Applications. IEEE Trans. on PAMI, vol. 19, no. 4, pages 366-378, 1997.
23. G. Kim and V. Govindaraju. Handwritten Phrase Recognition as Applied to Street Name Images. Pattern Recognition, vol. 31, no. 1, pages 41-51, 1998.
24. F. Kimura, S. Tsuruoka, Y. Miyake and M. Shridhar. A Lexicon Directed Algorithm for Recognition of Unconstrained Handwritten Words. IEICE Trans. Inf. and Syst., vol. E77-D, no. 7, 1994.
25. S. Knerr, V. Asimov, O. Baret, N. Gorsky, D. Price and J.C. Simon. The A2iA intercheque system : Courtesy

amount and legal amount recognition for French checks. In Automatic Bankcheck Processing, pages 43-86. World Scientific, 1997.

26. G. Koch, T. Paquet, L. Heutte. Combination of contextual information for handwritten word recognition. 9th IAPR International Workshop on Frontiers in Handwriting Recognition, IWFHR'2004, pp. 468-473, 2004.
27. A.L. Koerich, R. Sabourin and C.Y. Suen. Large vocabulary off-line handwriting recognition : A survey. Pattern Analysis and Applications, vol. 6, pages 97-121, 2003.
28. A. L. Koerich, Rejection strategies for handwritten word recognition, IWFHR (2004) 479-484.
29. A.L. Koerich, R. Sabourin, C.Y.Suen, Recognition and Verification of Unconstrained Handwritten Words, IEEE PAMI, Vol. 27, no.10, pp. 1509-1522, 2005.
30. D. D. Lewis, An evaluation of phrasal and clustered representations on a text categorization task, Proceedings of the 15th annual international ACM SIGIR, pp. 37 - 50, 1992.
31. G. Lorette, T. Paquet, La reconnaissance de l'Ecriture manuscrite, Traite IC2, Les Documents Ecrits, chap. 2, ISBN: 2-7462-1143-2, 2007.
32. R. Manmatha and W.B. Croft. Word Spotting: Indexing Handwritten Archives, Intelligent multimedia information retrieval, pp. 43-64, 1997.
33. U.V. Marti, H. Bunke, Handwritten Sentence Recognition, volume 3, pages 3467-3470, ICPR 2000, Barcelona, 2000.
34. U.V. Marti, H. Bunke. Text Line Segmentation and Word Recognition in a System for General Writer Independent Handwriting Recognition. ICDAR, pages 159-163, 2001.
35. S. Marukatat, Une approche générique pour la reconnaissance de signaux écrits en-ligne, thèse de doctorat de l'université Pierre et Marie Curie, 2005.
36. A. Nosary. Automatique Recognition of Handwritten texts trough writer adaptation. Ph.D Thesis (in french), Universite de Rouen, 2002.
37. R. Plamondon, S. N. Srihari, On-Line and Off-Line Handwriting Recognition : A Comprehensive Suvey, IEEE-PAMI, Vol. 22, Numero 1, pp. 63-84, 2000.
38. M. F. Porter. An Algorithm for Suffix Stripping. Program, vol. 14, no. 3, pages 130-137, July 1980.
39. L. R. Rabiner. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proceedings of the IEEE, vol. 77, no. 2, pages 257-286, 1989.
40. T. Rath, R. Manmatha, Features for word spotting in historical manuscripts, ICDAR (2003) 218-222.
41. T. M. Rath and R. Manmatha, Word Spotting for historical documents, IJDAR, vol. 9, pp 139-152, 2007.
42. M.D. Richard and R.P. Lippmann. Neural network classifiers estimate Bayesian a posteriori probabilities. Neural Computation, vol. 3, pages 461-483, 1991.
43. J.A. Rodriguez-Serrano, F. Perronnin, Score Normalization for HMM-based Word Spotting Using a Universal Background Model, ICFHR 2008, 2008.
44. J.A. Rodriguez-Serrano, F. Perronnin, Handwritten word-spotting using hidden markov models and universal vocabularies, Pattern Recognition (2009) 2106-2116.
45. G. Salton and C. Buckley. Term-weighting approaches, In automatic text retrieval. Information Processing and Management, 24:513-523, 1988.
46. F. Sebastiani. Machine learning in automated text categorization. ACM Computing Surveys, 34(1):1-47, 2002.
47. G. Seni and E. Cohen. External Word Segmentation of Off-Line Handwritten Text Lines. Pattern Recognition, vol. 27, no. 1, pages 41-52, 1994.
48. Z. Shi, S. Setlur and V. Govindaraju, "A Steerable Directional Local Profile Technique for Extraction of Handwritten Arabic Text Lines", International Conference on Document Analysis and Recognition (ICDAR'09), Spain, July 2009.
49. T. Stafylakis, V. Papavassiliou, V. Katsouros and G. Carayannis, "Robust Text-line and Word Segmentation for Handwritten Documents Images", in Proc. Intl Conf. Acoustics, Speech and Signal Processing, pp. 3393-3396, 2008.
50. K. Terasawa, Y. Tanaka, Slit style hog feature for document image word spotting, ICDAR (2009) p116-120.
51. T. van der Zant, L. Schomaker, K. Haak, Handwritten-word spotting using biologically inspired features, IEEE Trans. on PAMI 30 (11) (2008) 1945-1957.
52. A. Vinciarelli, J. Luettin. Offline cursive script recognition based on continuous density HMM, IWFHR, pages 493-498, 2000.
53. A. Vinciarelli, S. Bengio and H. Bunke. Offline Recognition of Unconstrained Handwritten Texts Using HMMs and Statistical Language Models. IEEE Trans. on PAMI, vol. 26, no. 6, pages 709-720, 2004.
54. A. Vinciarelli, Noisy Text Categorisation, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 27, no. 12, pp. 1882-1295, December 2005.
55. W. Wang, A. Brakensiek, A. Kosmala, G. Rigoll, Hmm based high accuracy off-line cursive handwriting recognition by baseline detection error tolerant feature extraction approach, IWFHR VII, Amsterdam, pp. 209-218, 2000.
56. Y. Yang and J.O. Pedersen. A comparative study on feature selection in text categorization. In Douglas H. Fisher, editeur, Proceedings of ICML-97, 14th International Conference on Machine Learning, pp. 412-420, Nashville, 1997.
57. A. Yazgan, M. Saraclar, Hybrid language models for out of vocabulary word detection in large vocabulary conversational speech recognition, IEEE ICASP Processing, May 17-21, 2004.
58. F. Yin, C.-L. Liu, Handwritten text line segmentation by clustering with distance metric learning, Proc. 11th Int. Conf. on Frontiers in Handwriting Recognition, Montreal, Canada, 2008, pp. 229-234.
59. M. Zimmermann, R. Bertolami, H. Bunke, Rejection strategies for offline handwritten sentence recognition, Pattern Recognition, 2004. ICPR 2004, Vol. 2, pp. 550-553 Vol.2, 2004
60. M. Zimmermann, J.-C. Chappelier, and H. Bunke. Offline grammar-based recognition of handwritten sentences. IEEE Trans. Pattern Analysis and Machine Intelligence, 18(5):818-821, 2006.

A Appendix: Document samples



[redacted]-MAURISSON
le Sentier
77000 MELUN

Ref: 06.6198[redacted]

[redacted]
SERVICE CLIENTELE
38218 VIENNE

Madame, Monsieur,

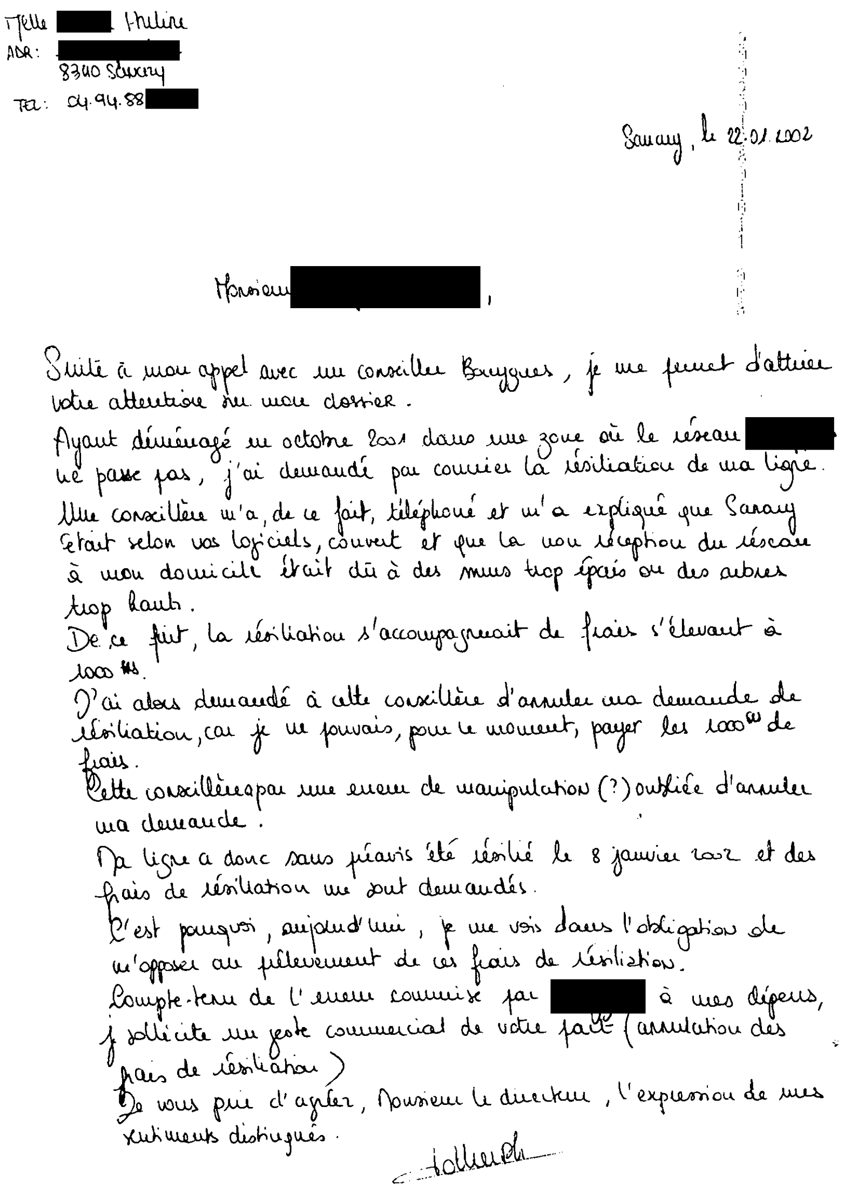
J'ai envoyé en date du 17 janvier 2002, une demande de resiliation de contrat d'abonnement telephonique à terme.

Par erreur de ma part, je vous ais joint les originaux de lettre d'interdiction bancaire de credit mutuel, comme preuve des difficultés financiers que je subisse en ce moment.

Je vous prie de faire les photocopies des lettres (4) d'interdiction [redacted] d'emettre de chèque par ma banque et de me renvoyé les originaux, pour pouvoir resoudre le problème auprès de ma banque.

Dans l'attente de recevoir prochainement les pièces jointes (originaux) je vous prie d'agréer, Madame, Monsieur, l'expression de mes salutations distinguées.

[redacted]



Mlle [redacted] Hélène
ADR: [redacted]
83110 Sanary
TEL: 04.94.88[redacted]

Sanary, le 22.01.2002

Monsieur [redacted],

Suite à mon appel avec un conseiller Bouygues, je me permet d'attirer votre attention sur mon dossier.

Ayant déménagé en octobre 2001 dans une zone où le réseau [redacted] ne passe pas, j'ai demandé par courrier la résiliation de ma ligne.

Une conseillère m'a, de ce fait, téléphoné et m'a expliqué que Sanary était selon vos logiciels, couvert et que la non réception du réseau à mon domicile était dû à des murs trop épais ou des arbres trop hauts.

De ce fait, la résiliation s'accompagnait de frais s'élevant à 1000 F.

J'ai alors demandé à cette conseillère d'annuler ma demande de résiliation, car je ne pouvais, pour le moment, payer les 1000F de frais.

Cette conseillère par une erreur de manipulation (?) oublié d'annuler ma demande.

Ma ligne a donc sans préavis été résilié le 8 janvier 2002 et des frais de résiliation me sont demandés.

C'est pourquoi, aujourd'hui, je me vois dans l'obligation de m'opposer au prélèvement de ces frais de résiliation.

Compte-tenu de l'erreur commise par [redacted] à mes dépens, je sollicite un geste commercial de votre part (annulation des frais de résiliation)

Je vous prie d'agréer, Monsieur le directeur, l'expression de mes sentiments distingués.