Pattern Anal Applic
DOI 10.1007/s10044-006-0038-6

THEORETICAL ADVANCES

Automatic name extraction from degraded document images

Laurence Likforman-Sulem · Pascal Vaillant · Aliette de Bodard de la Jacopière

Received: 16 August 2005 / Accepted: 3 June 2006

Abstract The problem addressed in this paper is the automatic extraction of names from a document image. Our approach relies on the combination of two complementary analyses. First, the image-based analysis exploits visual clues to select the regions of interest in the document. Second, the textual-based analysis searches for name patterns and low-level word textual features. Both analyses are then combined at the word level through a neural network fusion scheme. Reported results on degraded documents such as facsimile and photocopied technical journals demonstrate the interest of the combined approach.

Keywords Document image analysis · Name extraction · Neural Networks · Facsimile processing · Journal title pages · Document understanding · Visual clues

L. Likforman-Sulem (✉) · P. Vaillant ·
A. de Bodard de la Jacopière
Ecole Nationale Supérieure des Télécommunications/TSI and CNRS-LTCI, 46 rue Barrault, 75013 Paris, France
e-mail: laurence.likforman@enst.fr

Present Address:
P. Vaillant
Université des Antilles-Guyane, Institut d'Enseignement Supérieur de Guyane, Campus de Saint-Denis, Avenue d'Estrées, B.P. 792, 97337 Cayenne cedex, Guyane française
e-mail: pascal.vaillant@guyane.univ-ag.fr

A. de Bodard de la Jacopière
e-mail: aliette.de-bodard@polytechnique.org

1 Introduction

Companies today receive an increasing amount of messages. These messages include voice mails, e-mails, forms, letters, faxes and invoices. The electronic form of e-mails makes them suitable for indexing and filtering, producing summaries or automatic answers [1, 2]. Forms may also be filled electronically but many of them are still paper-based. Most messages, however, are not electronic and need manual processing such as scanning, indexing or routing. Whenever an electronic form is available, it is generally in a raw form such as a speech signal or an image. Speech and document recognition techniques allow deeper analysis of this raw material. There are already speech systems which are able to recognize names through a vocal dialog [3]. Form reading is also widely in use [4, 5]: the specific layout of the form directs the items to recognize. Postal codes, reference and telephone numbers can be localized from handwritten mails [6]. For faxes, the recognition of names, such as the sender or the recipient of an incoming message, is a key task. These names may be related respectively to the directory of the company/organization or to the client database of the addressee. Recognizing automatically the sender name can thus speed up the indexing process and the retrieval of archived messages concerning that person. Similarily, the recognition of the author name(s) of a letter or of an article is necessary for indexing and retrieving it.

In this paper, we describe a method for the extraction of names in degraded documents such as faxes and title pages from photocopied technical journals. For this purpose, we do not make any geometric assumption about the position of searched names as the layout varies greatly from page to page. Rather, we postulate

that names can be extracted through both visual clues and reading. This is necessary since a full textual analysis cannot take into account the physical proximity of words in the OCR stream. Moreover, the textual stream is corrupted in the case of degraded documents (error rates may be greater than 30%). A full image analysis would require building layout models adapted to a finite set of documents. The resulting system would not be able to process a large class of documents.

The objective of the proposed method is to use both analyses, textual and image-based, in order to complement each other. Three stages are required for the name extraction. The first one is the selection of image areas of interest. It consists in pointing out layout blocks which are near anchor components (such as headers, addresses,...). Visual clues enhance the detection of these anchors. The second stage is the analysis of the OCR stream, searching for name patterns and word characteristics. Dictionary lookups and local grammar rules are used in this stage. The third stage is the combination of the previous analyses. We use two neural network-based classification schemes, linear and non-linear. They both produce a score for each word of the document. The higher the score, the more likely it is that the word belongs to a searched name string.

The paper is organized as follows. Section 2 justifies the proposed approach. Section 3 describes the image analysis: layout extraction, writing discrimination and the selection of anchor components. The textual analysis of the OCR stream is developed in Section 4 and the combination of both analyses in Sect. 5. Experimental results on faxes and photocopied technical journals are given in Section 6. We draw some conclusions in Sect. 7.

2 Overview of the approach

The proposed approach aims at extracting names (sender names, author names) in unconstrained degraded documents. By unconstrained, we mean that name location can occur anywhere in the image. Names can first be searched at locations derived from an analysis of the logical structure of the document. Analysing business letters in [7], the OfficeMaid system is trained from empty documents to build reference layout models and logical object locations (date, address, ...). In the invoice reading system of ([8, 9]) the issuing company is first recognized in order to find invoice areas where to extract the fields of interest. The system described in [10] incrementally learns journal model classes from block positions and attributes logical labels through a graph-matching scheme. Block positions are also used in [11] to build the geometric tree which describes document classes and layout models. The rule-based system of [12] exploits block position, block relations and font attributes to derive logical labels in medical journals. A survey of document recognition methods exploiting OCRed text can be found in [13].

Word capitalization is used in [14] to detect names in newspaper images, in conjunction with the length of the words, their position and grammatical function within the sentences. Name extraction can also be derived from the location of keywords. In [15], one assumes that the name of the sender for instance, is to be found near the keyword ''from'' (or an equivalent word). As such keywords are also words of the language, physical blocks that are likely to be headers are selected on geometric features. The FaxAssist routing system from BCL Technology [16] uses four OCR engines and a database (legal names) of recipient names in order to identify a fax recipient. Two of the OCR engines solely concentrate on reading the handwriting, as handwritten items may be included in the document. The OCR stream is parsed to retrieve names. Names located after a specific keyword (''to'' or ''Attention'') have their match value increased. The fax routing system of [17] extracts on each word a large amount of features (2,128 features). Then a boosting training algorithm builds a classifier by selecting the right features. A database of recipients (alias database) is also used. Authors address the problem of efficient matching between fax words and names of the recipient database. A branch-and-bound scheme is used to lower the complexity of the match. The scheme is based on the computation of an order-invariant edit distance followed by a sort. The true edit distance is then computed on a restricted set of words.

In order to label author name blocks, the PixED system [18] first selects text lines before the keyword ''Abstract'', then uses textual patterns to evaluate how likely it is that a text line belongs to an author name block. However, in the documents we process, such keywords may be not present: the senders may only be found through their signature, or there may be no abstract preceding the article. Moreover, as we process degraded documents, the existing keywords may not be properly recognized. As a consequence, we have considered two complementary analyses. The image analysis selects regions of interest through visual clues and reading. The textual analysis scans the OCR stream retrieving words and name patterns of interest. Our method can therefore be compared to the one

proposed in [19] which exploits both image and textual analyses. However, this approach needs both an accurate segmentation (ground truth segmentation is used) and a very high recognition quality for the OCR stream, as the imbedded rules strongly rely on error-free text patterns. The method we describe uses a set of low-level word features which are hardly corrupted by OCR recognition errors.

Our system for name extraction is divided into four main components:

- Layout extraction and writing discrimination: the document image is preprocessed in order to discard graphical elements (logos, images, ruled lines) and small noise components. This filtering is based on the size of connected components. Layout blocks are obtained through a segmentation process. The resulting blocks are pseudo-words (PWs) for fax images and line blocks (PLs) for journal images. A PW can either be: an entire word, a part of a word, several words grouped if their spacing is narrow. Pseudo-words are then classified into two categories, printed or handwritten.
- Anchor detection and name block location: this stage is the core of the image analysis. It allows the extraction of anchors such as sender/recipient headers, addresses, near which the name blocks are to be found. To enhance anchor detection, we consider Anchor Pairs which are obtained by grouping anchor components through visual clues and reading. Visual clues help to focus on blocks with perceptual saliency. Reading is necessary to select the blocks of interest as they are many other blocks including key information such as the date, the subject of the message etc. Reading is performed from an off-the-shelf OCR software [20] which produces corrupted strings due to the low quality of the image processed. Potential name blocks are then retrieved near anchor positions.
- Word analysis and name detection: the textual analysis detects all names and extracts typographical attributes at word level. This low-level textual analysis scans the corrupted OCR output and uses two dictionaries: a language dictionary and a first name dictionary. This general-purpose system does not use any dictionary of family names by contrast with routing systems. Here, the set of users is open.
- Combination: at the end of the above stages, binary features are set for each word. These features are then combined to produce a scored list of physical blocks or strings according to the likelihood of being a sender/author name. The combination uses a neural network classification scheme.

Figure 1 highlights the system flow of the name extraction task.

3 Image analysis

A fax cover page generally consists of headers filled with information relevant to the sender and the recipient. The searched names can be handwritten when the user has filled in a preprinted form. Hence, we need to extract the layout structure of the document and separate the handwritten blocks from the printed ones. Although the recognition of handwritten items is beyond the scope of this paper, the knowledge of the nature of the handwriting can be used to direct the recognition to a specific handwriting recognizer. The fact that a block is handwritten is also a good guess that the information conveyed is important. As a commercial OCR cannot process handwritten items, a segmentation distinct from the one performed by the OCR is necessary.

Then the layout blocks which are likely to include the searched names are extracted in the neighborhood of anchor components. The choice of anchor components depends on the application targeted (fax or journal processing). Anchor detection is enhanced by grouping anchor components into Anchor Pairs. To avoid relying on predefined positions, Pair selection is performed by using Gestalt principles of grouping.

3.1 Extraction of layout objects and writing discrimination

The resolution of fax images is 200 dpi when they reach the fax server through a fax/modem card. Then images are reduced by half to speed the segmentation process into physical blocks. The document analysis stage begins with a classical connected components analysis. Most graphical elements such as boxes, logos, and very small components considered as noise, are discarded if their dimension (bounding box height and width) and their density are below threshold values. Typical threshold values for fax components are 3 pixels, 2 pixels and 0.1 respectively. To extract layout blocks, the Run Length Smoothing Algorithm [21] smears the cleaned image along both horizontal and vertical directions using the d_x and d_y parameters respectively. Typical values are $d_x = 100$, $d_y = 50$ pixels for fax images and $d_x = 200$, $d_y = 50$ pixels for journal images (originally at 300 dpi resolution and reduced by half). For journal images, the resulting blocks are pseudo text lines (called PLs). For fax images, the segmentation results in layout blocks

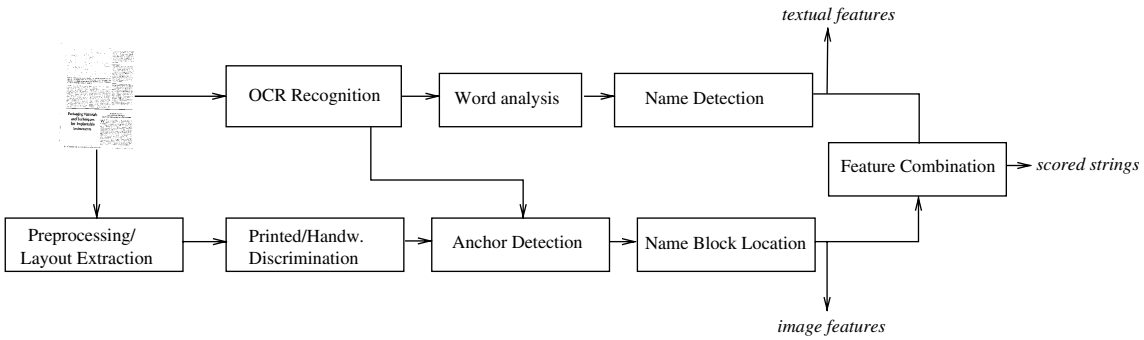

Fig. 1 Image and textual processing of document images

framing a subword, a word, or several words. These blocks are called pseudo-words (PWs) and they generally contain several connected components (CCs). In the case of handwriting, CCs may enclose one or several characters, a cursive word or part of a cursive word. The following writing discrimination concerns only PWs and CCs extracted from facsimile images which potentially include handwritten items.

On PWs we extract a set of features inspired from [22] and [23] but adapted to our short-sized layout components. These are:

height_diff: height difference (in pixels) between the highest and the smallest CC within the PW.
total_cc: the total number of CCs within the PW.
av_bottom_diff: average difference between the bottom positions of two consecutive CCs within a PW.

The first and last features are related to the regularity/irregularity of handwriting. The *height_diff* feature expresses the regularity/irregularity in the height of the connected components within a layout block. The last feature is the average difference in pixels between the bottom positions of two consecutive CCs. The second feature is related to writing connectiveness/separation: printed characters are mostly separated, unlike handwritten characters which may touch each other. This restricted set of features was selected from a larger set that also included the average and standard deviation of the density, the height of characters, the inter-character distance (the distance between two connected components belonging to the same word). The combination of the three above features was found the most discriminating. The choice of a multilayer perceptron (MLP) was determined by time considerations: as a facsimile front cover may contain hundreds of PWs, classification speed is important. The architecture, determined experimentally, is composed of three input cells, one hidden layer of 30 cells and two output cells. However, satisfactory convergence of the network could not be obtained due to the overlap of the two classes. To overcome this problem, we decided to privilege the correct learning of the handwritten class at the expense of the printed one, as printed information can be recovered later by the OCR. In the Bayesian framework, this could be done by including penalty costs in order to move the decision frontier so that it benefits one class. The global error rate is increased but the privileged class is better recognized. In the neural network context, a loss matrix and a weighted cost function could be considered [24, 25]. However, to cope with ambiguous data, points belonging to the printed class and lying in the main overlapping zone (corresponding to $total\_cc \leq 5$) were withdrawn from the training set. The test set was unchanged. The training set consisted of samples from 10 faxes and 960 training samples remained after overlap cleaning. Then the network converged after 200 epochs with the backpropagation algorithm [26]. We used a logistic sigmoid activation function and a learning rate of $\eta = 4.10^{-4}$

The classification rule, classifies as printed (resp. handwritten) a PW such as $O_p > \theta_p$ (respectively, $O_h > \theta_h$), taking into account the values of the two output cells (O_p and O_h for printed and handwritten respectively). The two thresholds values are : $\theta_p = 0.6$ for the printed class, $\theta_h = 0.25$ for the handwritten one ($\theta_h < \theta_p$) so as not to miss the handwritten class. Classification results on the test set (one third of the labeled examples of the database) are given in Table 1.

While the global error rate is rather high, we are able to obtain a low error rate for the (privileged) handwritten class. Figure 3 is an example of layout segmentation and writing discrimination. The original image is preprocessed then segmented into physical blocks (Fig. 3b). Rectangles in Fig. 3c are the pseudo-words classified as handwritten. Some short printed pseudo-words are misclassified as they contain few connected components.

Table 1 Classification results for handwritten/printed discrimination of pseudo-words (PWs)

	Correct classification rate (%)	Error rate (%)
Global	77.2	22.8
Handwritten class	97.4	2.6
Printed class	79.8	20.2

3.2 Anchor detection

Sender and recipient names in facsimile images are preceded by main sender/recipient headers (for instance ''From'', ''To''). The names may be in the immediate neighborhood of these main headers or in the neighborhood of secondary headers (for instance ''Name''). Similarily, in technical journals, author name(s) may be preceded by keywords such as ''By'' and ''And'' or may be in the neighborhood of address/affiliation blocks. Sample documents are shown in Fig. 2. The exact position of the searched names is hard to predict in both cases. It may appear in any position in the page, depending on the sending company, the journal editor, the location where a new article begins within the page....

Fig. 2 Sample documents: facsimile and technical journals

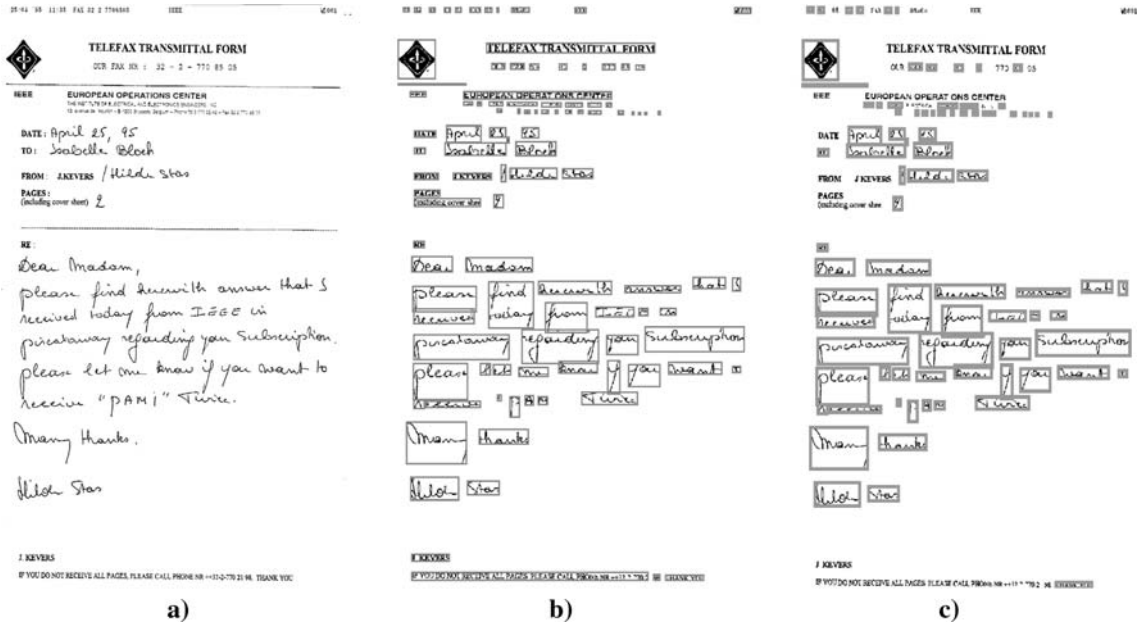

Fig. 3 **a** Original image. **b** Pseudo words. **c** Pseudo words classified as hand written

The Anchor detection stage consists in selecting candidate anchors and grouping them in a compound anchor block we call an *Anchor Pair*. The names we are looking for, can be found in the immediate neighborhood of the Anchor Pair. In facsimile images, the Anchor Pair includes the two main headers, one related to the sender, the other to the recipient of the message. For journal articles, Anchor Pairs are built from the layout blocks which include the keywords ''By'' and ''And'', or the keywords corresponding to address or affiliation items. To form a valid Pair, two layout blocks (PWs or PLs) have to share visual clues regardless of their absolute position in the image. The visual clues derive from the Gestalt laws of grouping [27, 28]. These laws explain the human structuration process by general principles such as proximity, similarity and direction continuity. Anchor Pairs are extracted through two of these principles: similarity and direction continuity. The components of a candidate Pair having the same horizontal axis must be bottom aligned, or if they have the same vertical axis, they must be both left (or right) aligned, or centered. These relations are respectively called H-Bottom, V-Left, V-Right, V-Center (Fig. 4).

Candidate anchors for a Pair are indentified by the strings they contain. A match must occur between a string included in a PW or PL and a string of the keyword dictionary. For fax images, we use a keyword dictionary split into three categories. These categories are: ''Recipient'', ''Sender'' and ''Name''. For instance category Recipient would include keywords such as: ''From'', ''De'', ''Attention'', ''Expéditeur''. For journal pages, the keyword dictionary is split into four categories : ''FirstAuthors'', ''LastAuthor'', ''Address'', ''Affiliation''. Category ''FirstAuthors'' includes keyword ''By'' while category ''LastAuthor'' includes strings ''And'' and ''&''. The comparison of block transcriptions with the keyword dictionary is implemented with a string matching algorithm. As OCR outputs are corrupted, we tolerate a distance of 1 or less to consider that the word belongs to the dictionary.

In fax images, each candidate anchor from the category ''Recipient'' is then associated with all candidate anchors from the category ''Sender'' to form Pairs. Pairs that do not satisfy the grouping criteria mentioned above are eliminated. A score is computed for each candidate Pair in order to select only the best hypotheses. The Pair score is a combination of two individual scores, one for each layout block of the Pair. In journal images, Pairs are built from either two anchors, one of class FirstAuthors, the other of class LastAuthor, or from two anchors of class Address or Affiliation (Figs. 6, 7).

An individual score is computed as follows: the further to the left of the layout block the keyword, the higher the score. For fax images the score decreases when the position of a layout block reaches the extreme upper and lower parts of the image while for journal images, the score decreases when the distance between the Pair anchors increases.

3.3 Name block extraction

From the Anchor Pair, we can draw sender and recipient regions, depending on whether the blocks of the Pair have the same horizontal or vertical axis (Fig. 5). The sender name block is then retrieved in the sender region The nearest block in the sender region, at the right of, or under the anchor block is the hypothesized sender name (SN) block. Blocks in the Sender Region and aligned with this hypothesized SN block are grouped with that last in a compound hypothesized SN block. If the sender region includes an anchor block of category ''Name'', it is this anchor block that is used to retrieve the sender name block. If no Anchor Pair is found, hypothesized sender name blocks are found from anchors of category ''Sender'' only.

For technical journals, the regions of interest derived from Pair block positions are shown in Fig. 7. Two regions are generally derived, except in the case where blocks of class FirstAuthors and LastAuthor are on the same textline. Then, the nearest block of each Anchor block, in the region of interest, is a hypothesized author name block.

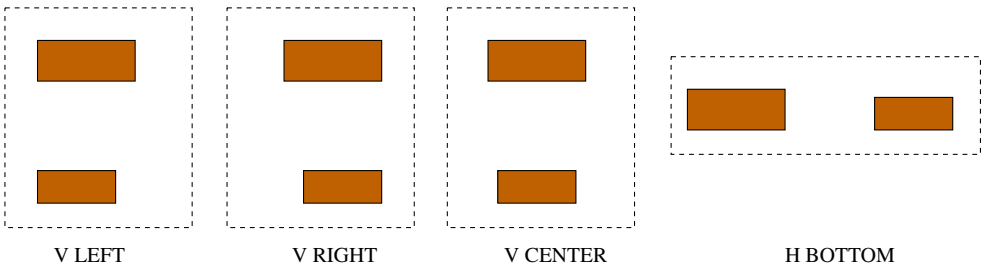

Fig. 4 Spatial relations between anchor pairs

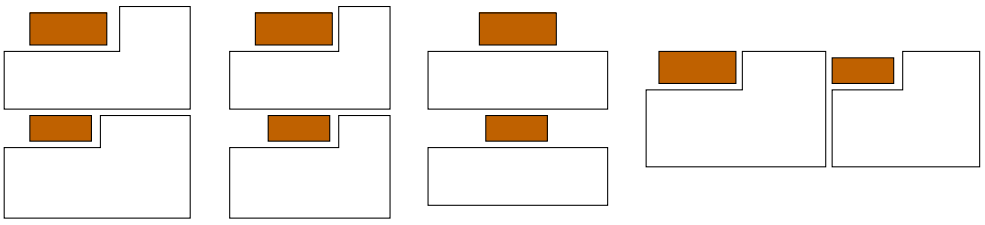

Fig. 5 Regions of interest derived from anchor pairs (facsimile images)

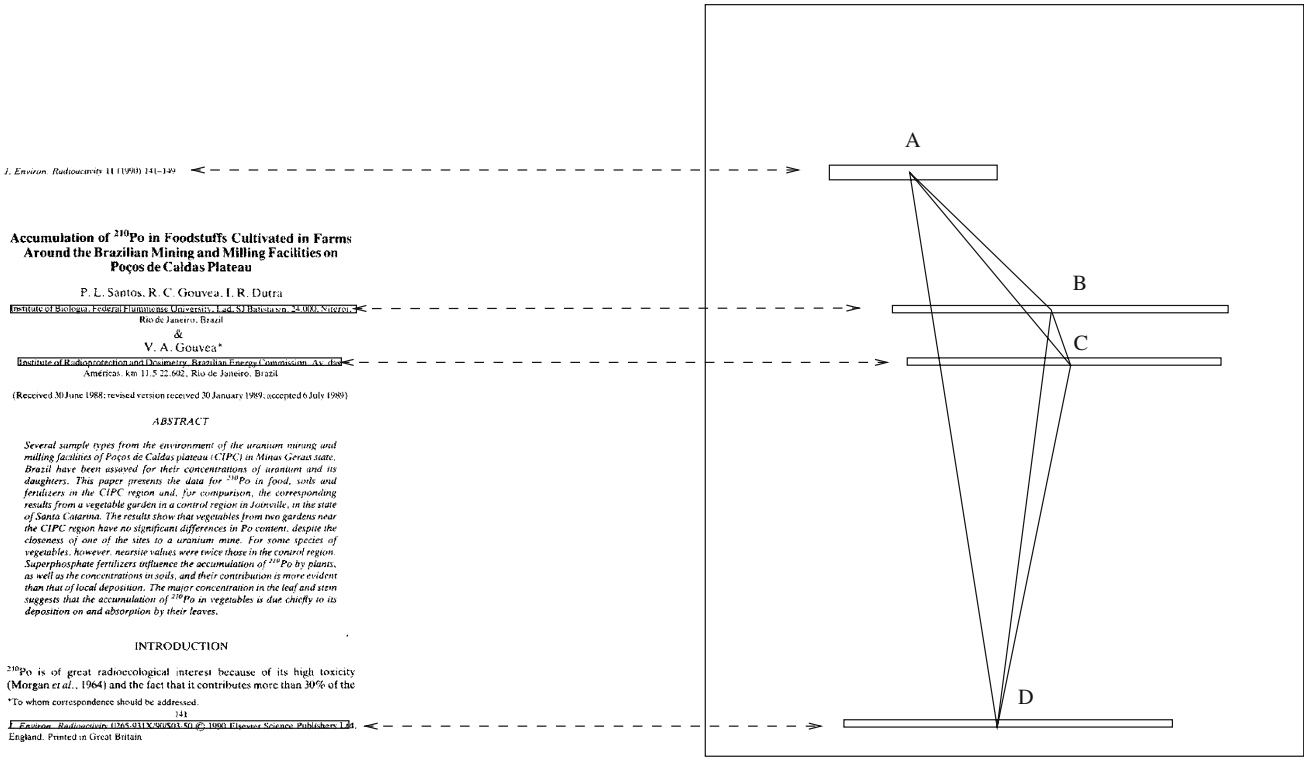

Fig. 6 Anchor components A, B, C, D are grouped into candidate pairs. The pair including PLs B and C is the highest scored pair

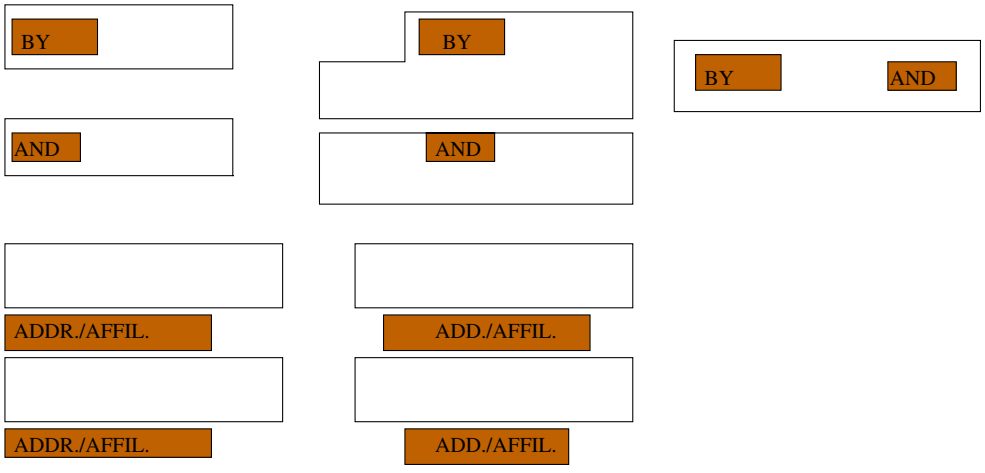

Fig. 7 Regions of interest derived from anchor pairs (technical journals)

4 Textual analysis

The textual analysis scans the word stream produced by an OCR (Optical Character Recognition) system. This stream provides strings of words and their positions but does not provide information on the logical function of words (simple word, institution, place, person names...), nor on their grammatical function.

Moreover the text strings provided by the OCR are corrupted and lacking punctuation. Deep parsing of these strings would lead to too many errors as sentences cannot be properly separated. The textual analysis proposed here is a lightweight parsing which aims at finding name patterns using only immediate word context, and at extracting simple textual features that can hardly be corrupted by image degradation.

4.1 Local grammars

Local grammar rules are used to detect the presence of simple name patterns within the OCR transcription of the document. For instance a first name or an initial followed by a capitalized word, or a common address form like ''Mrs'', followed by a capitalized word. The set of those patterns form a primitive local grammar aimed at spotting probable names.

If some of these patterns are language independent and OCR error resistant (for instance initial + capitalized word), others need to be searched in a dictionary (see Table 2). Here, words are searched in two different dictionaries: a list of usual first names, to check whether the word could possibly be the first name of a person, and a general wordlist, to check whether the word is in the language considered. The fact that the word is not included in the general dictionary has a positive correlation with the possibility of belonging to a name sequence. These lists contain 1,200 French first names [3] and 200,000 words from the French vocabulary [29]. The English general dictionary includes 60,000 words [30] and the first name dictionary was enlarged to 12,000 names [29] to process journal pages. All dictionaries are stored as binary balanced trees to speed up lookups. In order to avoid false alarms, we require an exact match to consider that the word belongs to the dictionary (first name, general). However, OCR segmentation errors can lead to several words being grouped together within one OCR string. In order to still detect words that have been concatenated in this way as being part of the dictionary, we detect separator symbols such as ''/'', '','', ''.'', etc., and form substrings that extend between two separator symbols. Those substrings are then matched against the dictionaries.

4.2 Internal and external clues

Strings or string sequences possibly being names must satisfy internal or external clues. Internal clues (respectively, external clues) correspond to the possibility for a word to belong to a name sequence considering the word in isolation (resp. in its local context). These clues are the following:

Internal clues: is the word written in capital letters, does the word begin with a capital letter, is it in uppercase, is it an initial, does it belong to one of the dictionaries (first name, general)?
External clues: is the word near an identity marker (Mr, Mrs, Dr), is the word included in a predefined name pattern (ref. Section 4.1)?

The fact that the word is capitalized, or in uppercase, is a clue. But all the words at the beginning of a sentence are capitalized, and acronyms generally are in upper-case. Many isolated items found in fax headers or footers also are capitalized. The fact that the word belongs to the general word list of the language has a negative correlation with the fact of belonging to a name sequence, but common family names often correspond to words in the language, for instance 'Vaillant' in French, or 'Singer' in English. So information given by only one clue may be insufficient to determine whether a word belongs to a name sequence or not.

In conclusion, information gathered in the text only is useful for finding name strings. However, names abound in fax messages and journal documents, as they may be found also in addresses, bibliography and page body. Hence, the information collected from the text and the information collected from the image have to complement each other.

5 Combination of image and textual features

The two previous analyses are conducted in parallel. From these analyses, each word is given a set of binary values. Those values are related to internal and external clues, the presence or absence of the word in a sender name hypothesis block and the writing type

Table 2 Initial and name patterns

Pattern	Examples
Initial	C, C., Ph., J-P
(Initial|first name) + capitalized word	D. MARTIN, D Martin, Ellen Martin
(First name|initial) + (Initial|first name) + capitalized word	R D. Martin, R. D. Martin, Ellen D. Martin, R Ellen MARTIN
Capitalized word + (initial|first name)	Martin C., MARTIN C
Address form + capitalized word	Mrs Martin

(printed or handwritten) for facsimile images. The spatial coordinates of each word and the physical block it belongs to are also recorded. Our approach consists in using a machine learning scheme. Learned approaches for name-finding include probabilistic methods such as in [31] where an HMM is proposed to classify text strings from clean sentences. Due to the size of the data set and the fact that sentences are not precisely separated, we use a neural-network-based classification at word level. We first use a single layer network yielding a linear weighted function. This linear combination is performed on a first set of features extracted from the above analyses. Then for printed documents, we select a more robust set of features, derived from the previous one. Then, to improve the combination scheme we combined the features non-linearly using a multilayer neural network.

5.1 Linear combination for mixed documents

A weight function is a simple way of combining features. The first set of binary features $f1$, $f2$, ... $f5$ was empirically selected from the whole set of binary values produced by both analyses. This feature set is composed of:

- **f1** indicates whether the word is located in a potential searched name (sender) block
- **f2** indicates whether the word can be considered as a name. This is done according to some predefined patterns (ref. external clues)
- **f3** indicates whether the word is printed or handwritten. We assume that being handwritten argues in favor of the possibility of the word being a name (in some cases fax cover pages are printed and users fill in the fields by hand).
- **f4** indicates whether the word begins with a capital letter or if it is capitalized.
- **f5** indicates whether the word is not found in a general dictionary.

Features $f2$, $f4$ and $f5$ result from textual analysis, while $f1$ and $f3$ result from image analysis. A word in a PW classified as printed has feature f3 equal to 0. A PW classified as handwritten is considered as a word with features f2 and f4 equal to 0 but with f3 and f5 equal to 1.

The objective of the machine learning scheme is to detect the searched names by discriminating between two categories: searched name (SN) and simple word (SW). We perform learning on a data subset extracted from facsimile images. Each word of the data subset is assigned a label: 1 if it belongs to the sender's name class and 0 if not. Category SW contains general words but also names of people: those of recipients, those of street names. The training set is composed of all words included in five faxes (828 words). Category SW is over-represented while category SN is under represented as most words in a fax belong to class SW. Words from the SN class are duplicated in order to perform an equal number of word presentations for each class.

A single layer network with five input cells and two output cells is then trained with the Widrow-Hoff algorithm as sample data are not linearly separable. The network converged after 25 epochs (learning rate of $\eta = 0.01$, linear activation function, desired outputs: 7 for SN words—0 for SW words).

Each word is assigned a score which reflects the possibility of being a searched name. Here, the score is computed as a linear combination of the features. For each word w, the word score $score_f(w)$ is given by the value of the output cell dedicated to the SN class.

$$\begin{aligned} \text{score}_f(w) = {} & 3.84 \times f1 + 1.53 \times f2 + 0.04 \times f3 \\ & + 1.28 \times f4 + 0.54 \times f5 + 0.79 \end{aligned}$$

Words are considered and scored in isolation. However, when a local grammar rule has been triggered on a string sequence, the string sequence is considered as a whole and a global score is given as the addition of the individual words' scores of the sequence. The system outputs the strings that have the highest scores (top N scores). N is a system parameter ($N = 1, 2$ or 3) as well as the WAT: *word acceptance threshold* (the score must be higher than the WAT). Figure 8 shows the scores found for the physical blocks or strings extracted on two types of facsimile, one mixed, one printed. The score value of layout block 17 in the mixed facsimile is high, as features $f1$ and $f3$ are set to one. Other strings are extracted with lower scores as they include French first names (''Pierre'' and ''Marie''). In the printed facsimile, the string ''Jérôme TEISSON'' was actually identified as a name pattern ($f2 = 1$) within a sender name block ($f1 = 1$) and the resulting score also includes the addition of textual features ($f4$ and $f5$) of both strings forming the name. The same name in the signature (but presented as Initial+Family name) has a lower score value as feature $f1$ equals 0.

5.2 Linear combination for printed documents

In this section, a linear combination is applied but to a new feature set. We also restrict ourselves to printed documents (including no handwritten items) as the textual analysis is of significant use only for that type of documents. This feature set is also derived from the

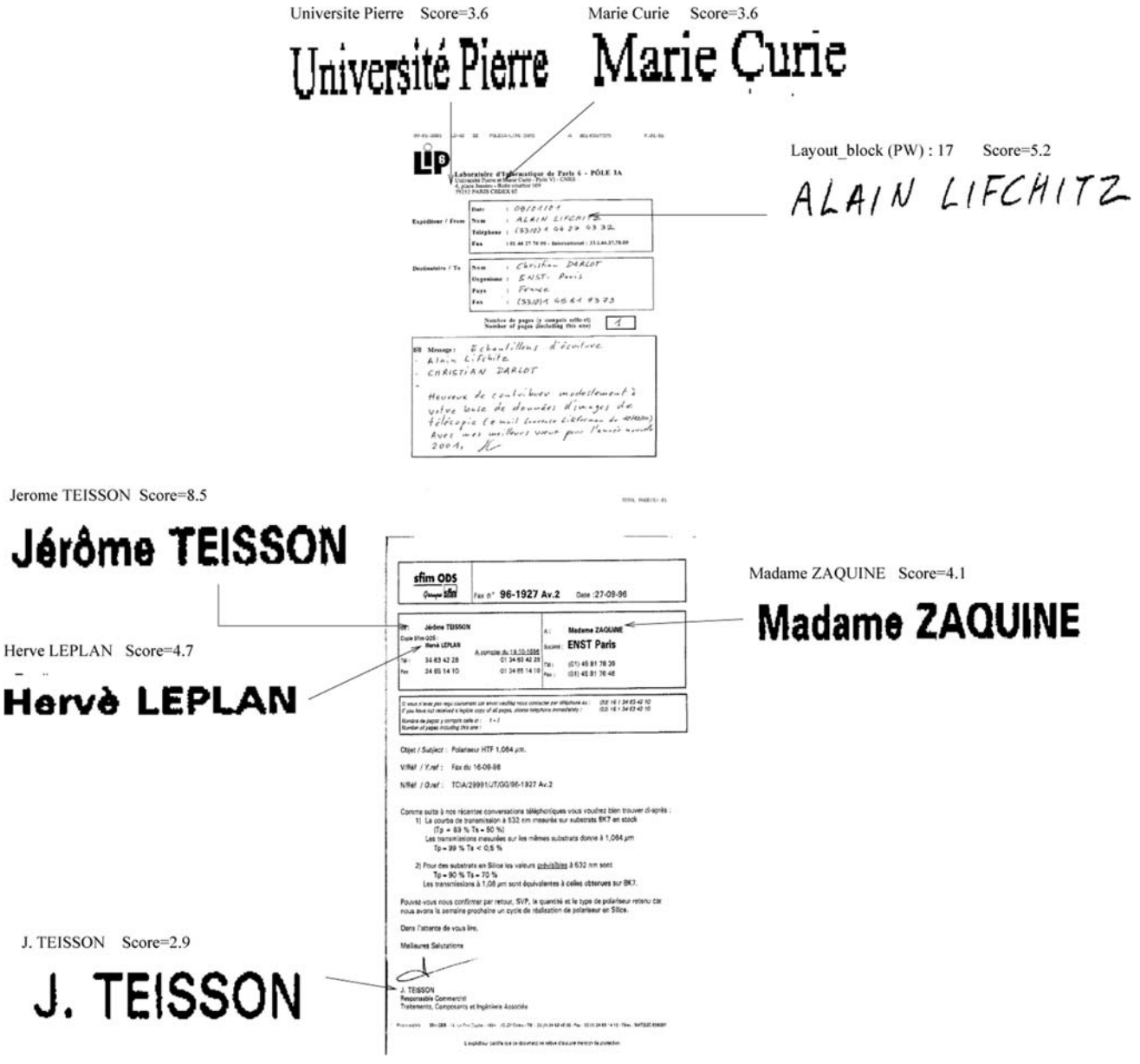

Fig. 8 Linear combination: sender name strings and corresponding scores extracted on two facsimile cover pages (printed and mixed documents)

entire set of binary values and and was found to be more discriminating. These features are:

- *g1* indicates whether the word is located in a potential searched name (SN) block
- *g2* indicates whether the word is found in a first name dictionary
- *g3* indicates whether the word is an initial
- *g4* indicates whether the word begins with a capital letter or if it is capitalized
- *g5* indicates whether the word is not found in a general dictionary.

The main difference between this feature set and the previous one, concerns features $f2$ and $f3$. Feature $f3$, which is no longer useful in the context of printed documents has been removed. Feature $f2$ has been replaced by features $g2$ and $g3$ which do not rely on any rule-based patterns and intermediate decisions (name sequence or not). The modified set is thus more adapted to corrupted printed text.

The linear classifier is retrained on a training set which includes 686 words from 4 printed faxes. The resulting linear function is for each word w:

$$\text{score}_g(w) = 3.62 \times g1 + 1.59 \times g2 + 3.52 \times g3 + 1.57 \times g4 + 0.11 \times g5 + 0.36$$

The network converged after 25 epochs (with a learning rate of $\eta = 0.01$, a linear activation function and desired outpouts: 7–0). We notice that feature $g1$ (equal to $f1$) has still an important weight: this stresses the importance of the image feature related to Pair extraction. The score function also shows that for a name, being capitalized (feature $g4$) is more important than not being included in the dictionary ($g5$).

5.3 Nonlinear combination

We improve in this section the classification scheme by using a non-linear combination on the previous feature set $g1$ to $g5$. Words from category SW include names whose features are close to those of class SN (street, recipient names, ...) so that a linear decision function hardly separates them.

The neural network is a MLP with one hidden layer of three cells and two output cells (one for each class). The network was trained by the back-propagation algorithm and converged after 200 epochs (with a learning rate of $\eta = 0.1$, a logistic sigmoid activation function and desired outputs 1–0). The two output values of the MLP clearly separate both categories (above and under 0.6 in our case). A score is then associated to each word by considering the output value of the MLP dedicated to the Searched Name category. The score represents the probability for this word to be a sender/author name.

Each word is classified individually but names are generally composed of several strings. One of these strings may be misclassified by the neural network. This is generally the case for initials that contain little information. Post-processing consists in iteratively grouping to a word classified as SN other words in its neighborhood. The SN component acts as an anchor for retrieving other name components. This introduces the concept of name group where words are gathered to form an *expression*. To be gathered as an expression two words w_1 and w_2 must be close enough in the horizontal and vertical direction. Typical threshold distance values for the horizontal and vertical distance are respectively 5 and 100 pixels. Expressions are then filtered : in journal pages, half the words of author name expressions should be capitalized. Once expressions are formed we assign them a score. The score is given accroding to one of the following rules: the maximum or the mean rule. With the maximum rule, the expression score is the highest of the word scores; with the mean rule, the expression score is the mean of the word scores. The maximum rule is used for facsimile and the mean rule for journal pages because expressions include more words in journal pages than in facsimile. Expressions over a threshold (the expression acceptance threshold (EAT): 0.7 with the maximum rule, 0.4 with the mean rule) are returned by the system. Figures 9 and 10 show expression strings extracted on facsimile and journal pages and their corresponding scores.

6 Experiments

The experiments reported here have been performed on two corpuses. The first corpus is made of 150 real-word facsimile images collected within the Majordome Project (2000–2003) [32]. The data were collected from about 50 different companies and institutions and the language is mainly French. The second corpus comes from the benchmark UW English Document Database [33]. This corpus is made of 299 images of scanned technical journals from various domains (medicine, physics, computer science, sociology...). Some images are scanned directly from the journal pages, but most of them are scanned from first generation or later-generation photocopies. We selected in this corpus the 154 images representing title pages.

We evaluate the different systems in this section. For facsimile images, we compare the two combination schemes, linear and non-linear. The linear function is first applied to printed and mixed documents using the first set of features ($f1$–$f5$) with one feature that is reserved for discriminating between printed and handwritten items ($f3$). Then the linear function is applied to printed documents with another feature set ($g1$–$g5$). The second combination scheme (the multi-layer perceptron) combines features in a non-linear fashion to detect the searched names.

The systems are evaluated on recall and precision rates [34]. Matches are set word by word, i.e. when a string belonging to the searched name has been correctly extracted by the system, we consider there is a match. To extract a complete name (family, first names, initials), several correct matches are required.

For fax images, the weighted combination function can easily be modified so that we can observe the influence of the image-based and textual analysis separately. First, we only consider features $f2$, $f4$ and $f5$ and we re-train the weights of the linear classifier using

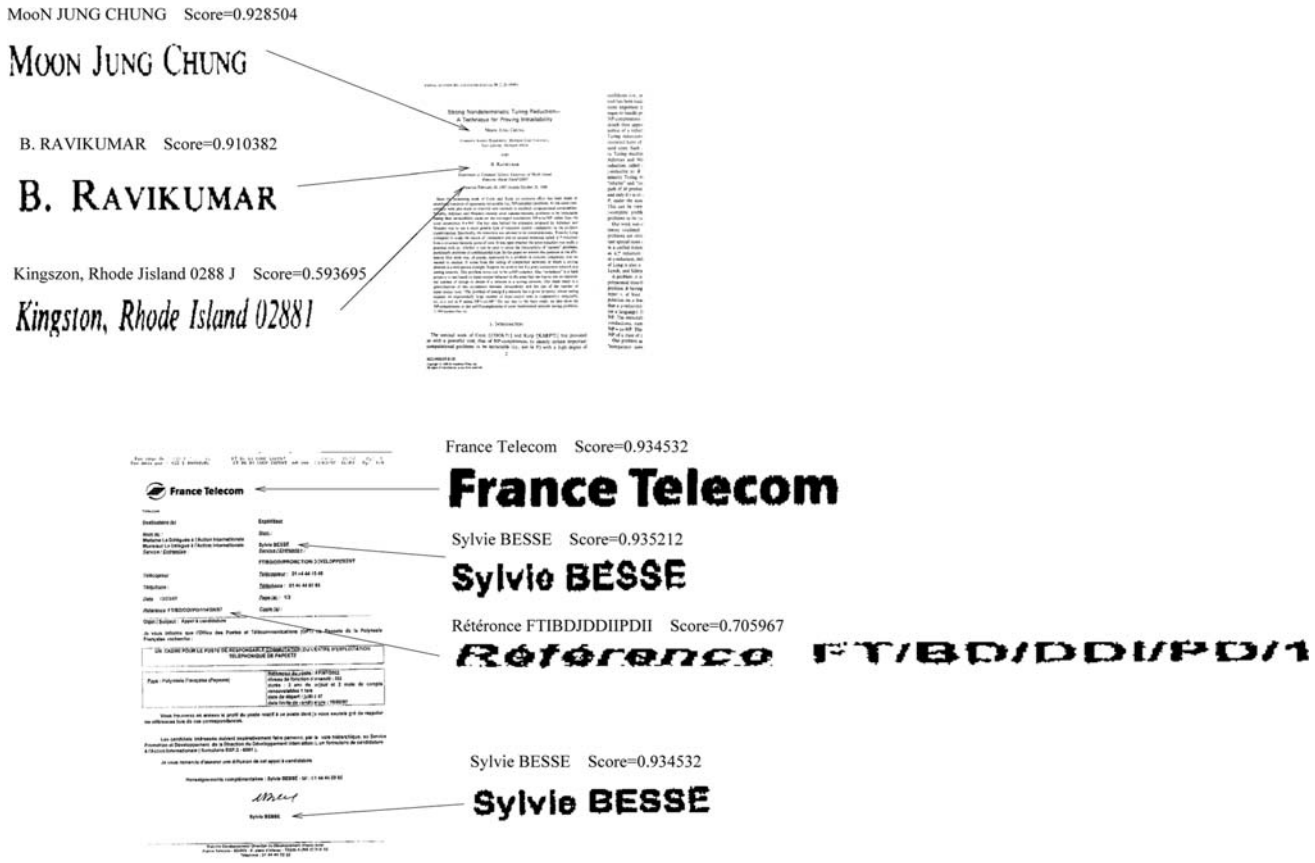

Fig. 9 The MLP combination: SN expressions and scores extracted on facsimile (max rule) and journal documents (mean rule)

only these features: we call this system Syst-T which corresponds to a system based on textual features only. Then, we re-train the linear classifier using only features f1 and f3: the system is called Syst-I, the image system, with no textual analysis. These two systems are compared in Table 3 with the system using both features (Syst-I-T): all five features ($f1$–$f5$) are used and combined according to the combination rule described in Sect. 5.1. Rates are obtained with $N = 3$ (top N scores) and word acceptance threshold WAT = 2 for System-I and System-T, and WAT = 3 for the other systems. System-I has the highest precision rate as it only extracts strings within Anchor Pairs and headers of class Sender. System-I has also the lowest recall rate as it does not extract name strings when Anchors Pairs do not exist, or when they are missed. Names within signatures also cannot be retrieved with System-I. System-T has a higher recall rate than System-I as it extracts name patterns in the OCR stream. Consequently, its precision rate is lower. However, System-T cannot process faxes including handwritten names and the given rates were evaluated on printed faxes only. The combination of image and textual features performed by System-I-T outperforms each system taken in isolation: image features of System-I-T allow the extraction of names, even when they are handwritten or when name patterns are not recognized (first names not in dictionary, or including OCR errors). Textual features of System-I-T allow the extraction of names without relying on anchors and headers which may be not detected. As expected, the I-T+ system, evaluated on printed material only, performs better. Missed names generally come from missed anchors due to OCR errors or first names not present in the dictionary. Short string headers (such as 'de' in French or 'to' in English) are more likely to be missed because of OCR recognition errors. The first name dictionary contains approximately 1,200 names but does not include all variants of them. This dictionary, which was extracted from a corpus of phonebooks in a French-speaking region, obviously fails to cover properly most non-French first names. More generally, precision is rather low for fax images. Firm names can be retrieved together with the sender name as these names do not appear in our dictionaries (first name, general). Words in capital letters in the neighborhood of word ''France'' may be retrieved. France is both a first name and a country name which appears in many addresses of this

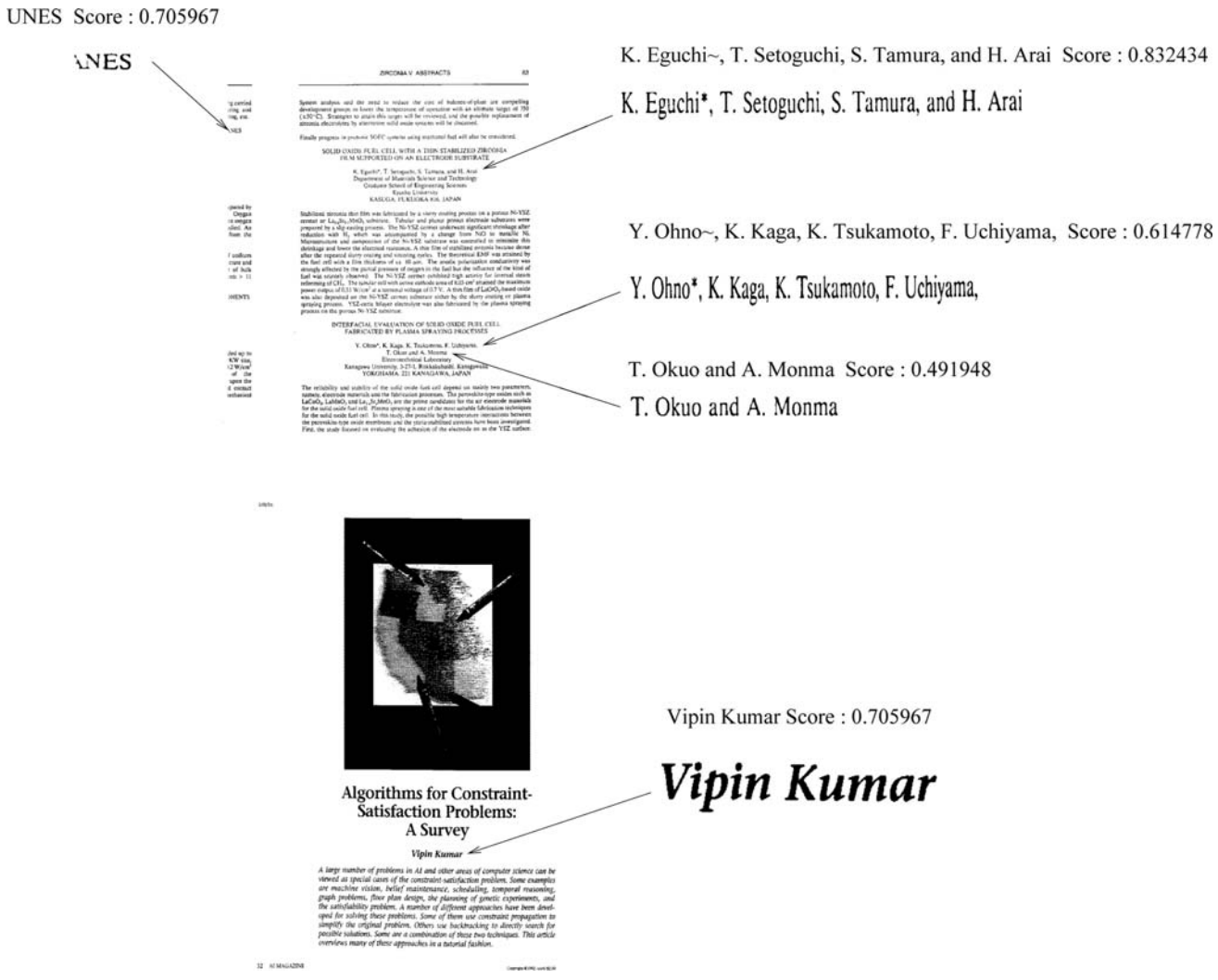

Fig. 10 MLP combination: SN expressions and scores extracted on journal pages

corpus. Finally recipient names may be also retrieved even if their score is lower than the one of the sender.

In Fig. 11a and b, the performance of each classifier is shown in the recall/precision space. The different operating points for linear systems are obtained by varying two system parameters: the WAT (word acceptance threshold) and the value N of the top N output scores. Points corresponding to the same WAT are joined through a line. Two lines are shown per system corresponding to WAT = 2 and 3. On each line, points correspond to two values of N: $N = 1$ and $N = 3$. The intermediate point ($N = 2$) has been omitted for the sake of readability. Figure 11c compares feature set $f1$–$f5$ with set $g1$–$g5$ on printed documents. When systems are tuned in order to reach precision rates within range [40 60], feature set f yields higher recall. But a system using feature set g can reach higher precision rates, over 70%. Similarly, Fig. 11d shows that higher precision rates are reached by the linear classifier and higher recall rates by the non-linear classifier. For the non-linear classifier, the operating points are obtained by varying the EAT (expression acceptance threshold) from 0.7 to 0.9. However around a middle precision value (40%), both the linear system (feature set f) and the non-linear system behave similarly (Fig. 11e)

Table 3 Linear system comparison (facsimile images): Syst-T (textual features only), Syst-I (image features only), Syst-I-T (both features), Syst-I-T+ (both features, printed material)

System	Recall (%)	Precision (%)
Syst-I	42.05	50
Syst-T	53.35	34.87
Syst-I-T	57.98	45.18
Syst-I-T+	64.66	45.86

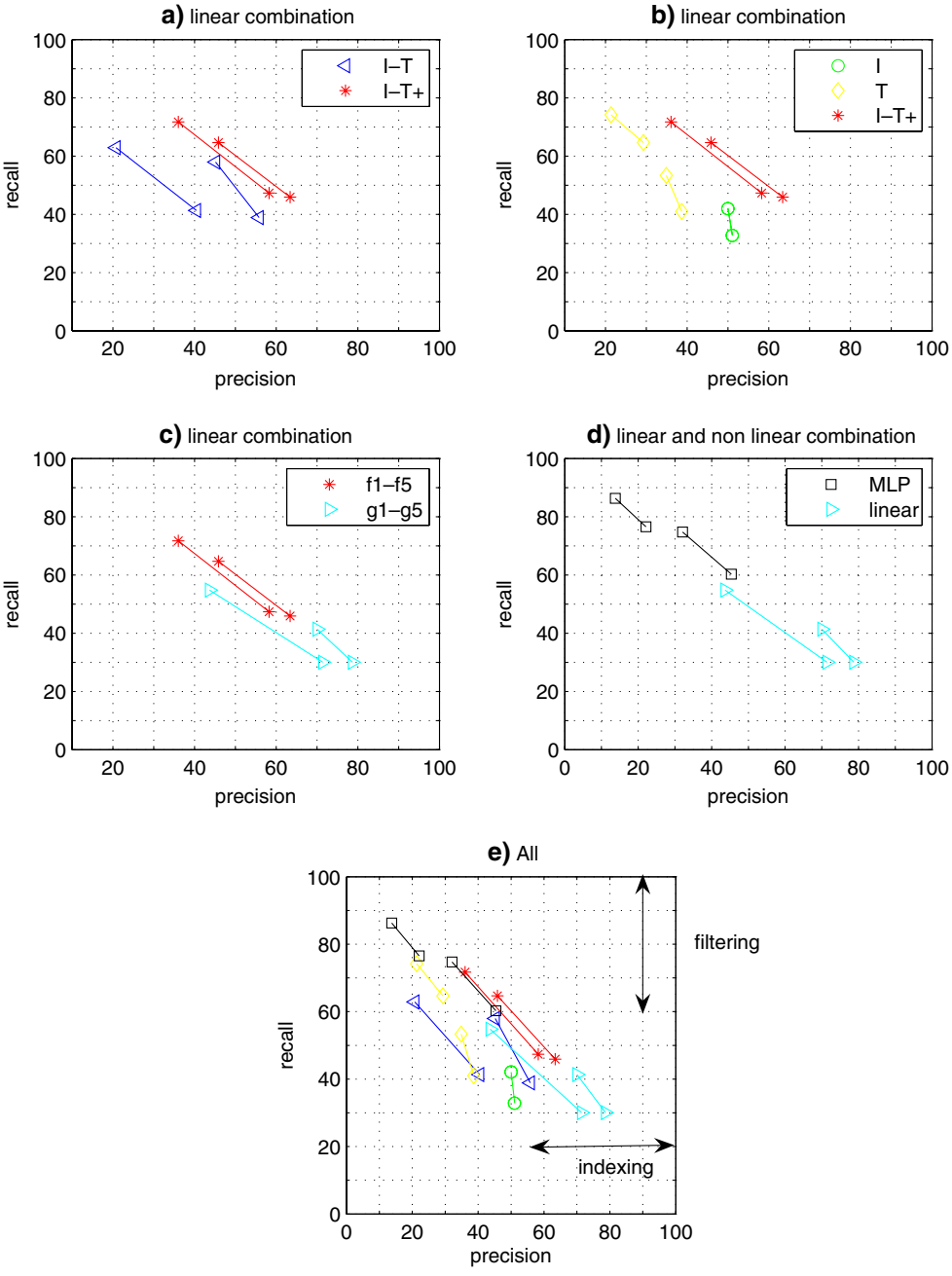

Fig. 11 Recall and precision rates for the different systems (facsimile images). For each linear system, operating points are obtained by changing word acceptance threshold (WAT = 2, 3) and the top N value (N = 1, 3). For the non-linear system, the EAT varies from 0.7 to 0.9. **a** Linear combination using features $f1$–$f5$ for printed only and mixed documents. **b** Linear combination with $f1$ and $f3$ only (image system), $f2$, $f4$ and $f5$ only (textual system), all features. The two lines corresponding to System-I (WAT = 2 and 3) are superimposed. **c** Linear combination using features $f1$–$f5$ and features $g1$–$g5$. **d** Linear and non-linear combination with features $g1$–$g5$. **e** All systems

Recall increases using the MLP model (see Table 4) but precision decreases. Recall is improved because capitalized words are more likely to be classified as searched names using the MLP. Concerning precision, contextual post-processing ensures that words lying in the neighborhood of a highly-scored word are retrieved: these words were ignored by the MLP and this makes the precision rate decrease. However, many words with low scores that were ignored by the neural network (for instance initials) may be detected using contextual post-processing. This makes the recall rate increase. Thus, contextual analysis improves the recall rate and retrieves complete sender name strings which is qualitatively better than retrieving isolated words.

For journal title pages, recall and precision rates in the case of an MLP combination are given in Table 4. Both recall and precision rates reach higher values for journal pages than for printed facsimile. Concerning recall, some author names may be missed because of segmentation errors. In our scheme, the address/affiliation block should be segmented into individual text lines separated from author text lines, which is not always the case when the document is tilted. Moreover, the OCR sometimes fails to produce accurate word

positions. Missed names also occur when authors belong to name sequences included into the general English dictionary (for instance Asian names such as Pang, Fang,...) which lowers the scores of these names. Concerning precision, some of the false alarms for author names are due to words from the title which are in capital letters, and may include technical names which do not belong to the general dictionary. Part of addresses may be retrieved, especially in the neighborhood of state names or city names which do not belong to our dictionaries or could even be first names (such as Austin for instance). Finally author titles on the same textline as the author names are also retrieved (such as ''Member'', ''IEEE'', ''Senior'') because of the contextual post-processing.

Results on journal images are better than on facsimile images. This result may be surprising as journal title pages usually contain many more words than fax cover pages. This is explained easily: in facsimile, a lot of words are written with capital letters, leading to false alarms, unlike articles where capital letters are left for titles, names and addresses. The facsimile layouts may also be more varying than those of journals, which follow professional typographic conventions for readability and visibility.

The different systems presented here may be used for indexing or filtering messages depending on their performance. For indexing messages, the precision rate must be high enough to minimize the amount of manual indexing. To filter messages, a high recall rate is necessary to spot key information (Fig. 11c). The BCL fax routing system presented above reaches recall/precision rates of 78.24%/99.73% on the BCL printed fax database. A high precision rate is achieved because using a recipient database allows the system to discard all the words that do not belong to it. Similarly, the system presented in [17] correctly routes 95% of printed faxes using also a database of recipients (the alias database). In the case of routing systems, using a recipient database including for instance first names, initials and family names is necessary. And building such a database is possible because the set of recipients is closed. The name extraction system presented here is a general-purpose name extraction system, not related to a finite set of users. Our system reaches recall/precision rates of 86%/13% for fax images and 92%/41% for journal images. Our system could be boosted by collecting a general database of family names. But this would make comparison between our system and another difficult as performance would be highly related to the size and coverage of this family name dictionary.

Table 4 MLP combination: results for printed facsimile ($EAT = 0.7$) and journal title pages ($EAT = 0.4$)

MLP combination	Recall (%)	Precision (%)
Facsimile	86.28	13.78
Journal title pages	92.22	41.33

7 Conclusion

We have demonstrated in this paper the usefulness of a combined approach for the extraction of names in document images. The approach relies on a small set of image-based and textual features which are easy to extract and thus suitable for practical systems. The method considers large classes of documents and is effective even in the case of degraded documents such as facsimile and photocopied technical journals. Image features consider regions of interest near Anchor components and the writing type. To enhance this detection, Anchor components are grouped into Pairs using visual clues. One main advantage of using Pairs is that regions of interest are located with more robustness and, in the case of fax images, sender and recipient regions are clearly separated. Textual features consider name patterns and typographic attributes of words of the OCR stream. All features extracted are then combined through a machine learning scheme.

This method has been applied to two completely different classes of documents, facsimile and journals, with small adaptations from one class to the other. Within each class of document, layouts may vary strongly and there is no attempt to refer to a learned model including block positions. The generality of the method is well-suited to general purpose systems, which have to deal with a lot of incoming documents with unknown layouts. Further improvements may be obtained as OCRs improve, and by the use of technical dictionaries. Also, we could use scored matches against dictionaries, which would yield features consisting of real numbers instead of binary numbers.

Acknowledgements We thank the French Ministry of the Economy, Finance and Industry (MINEFI) which has been supported this work under Grant no : 01.2.93.0268. This work could not have been possible without the competent help of François Yvon of the ENST Computer Science Department, who devoted much of his time during the first stage of this project to provide us with advice, guidance, and scientific experience. The authors also wish to thank Noura Azzabou for her assistance in the experiments.

References

1. Vinot R, Yvon F (2001) Semi-automatic response in a Mail Center. In: Proceedings of the 10th international symposium on applied stochastic models and data analysis. ASMDA 2001, Compiègne (France), pp 992–997
2. Sakkis G, Androutsopoulos I, Paliouras G, Karkaletsis V, Spyropoulos CD, Stamatopoulos P (2001) Stacking classifiers for anti-spam filtering of e-mail, 6th conference on empirical methods in natural language processing, Carnegie Mellon University, Pittsburgh, pp 44–50
3. Gravier G, Yvon F, Ettore G, Chollet G (1997) Directory name retrieval using HMM modelling and robust lexical access. In: Proceedings of the IEEE Workshop on automatic speech recognition and understanding, Santa Barbara
4. Leibowitz-Taylor S, Fritzon R, Pastor JA (1992) Extraction of data from preprinted forms. Mach Vis Appl 5(3):211–222
5. Casey R, Ferguson D, Mohiuddin K, Walach E (1992) Intelligent forms processing system. Mach Vis Appl 5(3):141–155
6. Koch G, Heutte L, Paquet T (2005) Automatic extraction of numerical sequences in handwritten incoming mail documents. Pattern Recogn Lett 26:1118–1127
7. Baumann S, Ali M, Dengel A, Jäger T, Malburg M, Weigel A, Wenzel C (1997) Message extraction from printed documents: a complete solution, 4th ICDAR. Ulm (Germany), pp 1055–1059
8. Cesarini F, Gori M, Marinai S, Soda G (1998) INFORMys : a flexible invoice-like form reader system. IEEE PAMI 20(7):730–745
9. Cesarini F, Francesconi E, Gori M, Soda G (2003) Analysis and understanding of multi-class invoices. IJDAR 6:102–104
10. Liang J, Doermann D (2002) Logical Labeling of Document Images using layout graph matching with adaptive learning. In: Lopresti D, Hu J, Kashi R (eds) DAS, Princeton, pp 224–235
11. Dengel A, Barth G (1988) High level document analysis guided by geometric aspects. IJPR 2(4):641–655
12. Kim J, Le DX, Thoma GR (2001) Automatic labeling in document images. In: IS&T/SPIE conference on document recognition and retrieval VIII, San Jose, pp 111–122
13. Lin X (2005) DDR research beyond COTS OCR software: a survey. In: IS&T/SPIE conference on document recognition and retrieval XII. San Jose, 2005, pp 16–20
14. De Silva GL, Hull J (1994) Proper noun detection in document images. Pattern Recogn 27(2):311–320
15. Lii J, Srihari SN (1995) Location of name and address on fax cover pages, 3rd ICDAR. Montréal (Québec, Canada), pp 756–759
16. Alam H, Hartono R, Sugono Y, Tran T (2000) FaxAssist : an automatic routing of unconstrained fax to email location. In: IS&T/SPIE conference on document recognition and retrieval XI, San José, pp 148–156
17. Viola P, Rinker J, Law M (2004) Automatic fax routing. In: Proceedings of document analysis systems, DAS 2004, pp 484–495
18. Faure C (2000) Extracting the tables of contents from the images of documents. In: Proceedings of RIAO, Paris
19. Klink S, Kieninger T (2001) Rule-based document structure understanding with a fuzzy combination of layout and textual features. IJDAR 4:18–26
20. Xerox (1994) ScanWorX API release notes. Xerox imaging systems
21. Wong KY, Casey R, Wahl F (1982) Document analysis system. IBM J Res Dev 6:642–656
22. Palumbo P, Srihari S, Soh J, Sridhar R, Demjanenko V (1992) Postal address block location in real time. Computer 25(7):34–42
23. Fan K-C, Wang L-S, Tu Y-T (1998) Classification of machine printed and handwritten texts using character block layout variance. Pattern Recogn 31(9):1275–1284
24. Bishop C (1995) Neural networks for pattern recognition. Oxford University Press, Oxford
25. Lowe D, Webb AR (1990) Exploiting prior knowledge in network optimization: an illustration from medical prognosis. Network 1:299–323
26. Faussett L (1994) Fundamentals of Neural Networks. Prentice Hall, Englewood Cliffs
27. Bruce V, Green P, Georgeson M (2003) Visual perception: physiology, psychology and ecology. Psychology Press, Hove (East Sussex), UK
28. Holstege M, Inn Y, Tokuda L (1991) Visual parsing: an aid to text understanding. In: Proceedings of RIAO'91, Barcelone, pp 175–193
29. ABU, Association des Bibliophiles Universels, on http://www.abu.cnam.fr/
30. Kelk B (2003) UK English wordlist with frequency classification, version 1.0, 1 February 2003, on http://www.bck-elk.uklinux.net/menu.html
31. Bikel D, Schwartz R, Weischedel R (1999) An algorithm that learns what's in a Name. Mach Learn 34:1–3, 211–231
32. Likforman-Sulem L, Chollet G, Vaillant P, Azzabou N, Blouet R, Renouard S, Mostefa D (2004) Reconnaissance de noms propres et vérification d'identité dans un système de messagerie, convention Minefi no 01.2.93.0268, Final Report, January 2004, 100 p
33. Askilrud ES, Haralick RM (1993) A quick guide to uw english document image database I. Department of Electrical Engineering, Department of Computer Science/Software Engineering, University of Washington
34. Alvarez S (2002) An exact analytical relation among recall, precision and classification accuracy in information retrieval. Technical Report, Computer Science Department, Boston College

Author Biographies

Laurence Likforman-Sulem is graduated in engineering from Ecole Nationale Supérieure des Télécommunications-Bretagne (ENST) in 1984, and received her Ph.D. from ENST-Paris in 1989. She is Assistant Professor at ENST in the Department of Signal and Image Processing since 1991 where she serves as a senior instructor in Pattern Recognition and Document Analysis. Laurence Likforman is a founding member of the francophone Groupe de Recherche en Communication Ecrite (GRCE), association for the development of research activities in the field of document analysis and writing communication. Her research area concerns document analysis dedicated to handwritten and historical documents, document image understanding and character recognition.

Pascal Vaillant graduated as a telecommunication engineer from France's National Telecommunication Institute (INT) in 1992. He received a Ph.D. in Cognitive Science from the University of Paris-Sud (Orsay, France) in 1997. He now teaches computer science and linguistics at the University of the French West Indies and Guiana, in Cayenne. His research is about semantics of texts and images, and automatic learning from text corpora.

Aliette de Bodard de la Jacopière is graduated in engineering from Ecole Polytechnique in 2004. She now studies for a Master in Signal and Image Processing at Ecole Nationale Supérieure des Télécommunications (ENST). Her interests are pattern recognition, document analysis and more particularly information extraction.