International Journal of Document Analysis (2005)
DOI 10.1007/s10032-005-0147-6

REGULAR PAPER

Y. Fataicha · M. Cheriet · J. Y. Nie · C. Y. Suen

Retrieving poorly degraded OCR documents

Received: 8 February 2005 / Published online: 13 October 2005

Abstract A significant portion of currently available documents exist in the form of images, for instance, as scanned documents. Electronic documents produced by scanning and OCR software contain recognition errors. This paper uses an automatic approach to examine the selection and the effectiveness of searching techniques for possible erroneous terms for query expansion. The proposed method consists of two basic steps. In the first step, confused characters in erroneous words are located and editing operations are applied to create a collection of erroneous error-grams in the basic unit of the model. The second step uses query terms and error-grams to generate additional query terms, identify appropriate matching terms, and determine the degree of relevance of retrieved document images to the user's query, based on a vector space IR model. The proposed approach has been trained on 979 document images to construct about 2,822 error-grams and tested on 100 scanned Web pages, 200 advertisements and manuals, and 700 degraded images. The performance of our method is evaluated experimentally by determining retrieval effectiveness with respect to recall and precision. The results obtained show its effectiveness and indicate an improvement over standard methods such as vectorial systems without expanded query and 3-gram overlapping.

Keywords Document processing · Optical character recognition (OCR) · Information retrieval (IR) · Error-grams · Query expansion

Y. Fataicha (✉)
Laboratory for Imagery, Vision, and Artificial Intelligence (LIVIA), École de Technologie Supérieure, 1100 Notre-Dame West, Montreal, Quebec H3C 1K3, Canada; Department Informatique et Recherche opérationnelle, University of Montreal, CP 6128, succursale Centre-ville, Montreal, Quebec H3C 3J7, Canada; Centre for Pattern Recognition and Machine Intelligence (CENPARMI), Concordia University, Suite GM-606, 1455 de Maisonneuve Boulevard West, Montreal, Quebec H3G 1M8, Canada
E-mail: fyoussef@livia.etsmtl.ca

M. Cheriet
Laboratory for Imagery, Vision, and Artificial Intelligence (LIVIA), École de Technologie Supérieure, 1100 Notre-Dame West, Montreal, Quebec H3C 1K3, Canada
E-mail: mohamed.cheriet@etsmtl.ca

J. Y. Nie
Department Informatique et Recherche opérationnelle, University of Montreal, CP 6128, succursale Centre-ville, Montreal, Quebec H3C 3J7, Canada
E-mail: nie@iro.umontreal.ca

C. Y. Suen
Centre for Pattern Recognition and Machine Intelligence (CENPARMI), Concordia University, Suite GM-606, 1455 de Maisonneuve Boulevard West, Montreal, Quebec H3G 1M8, Canada
E-mail: suen@cenparmi.concordia.ca

1 Introduction

During the past 20 years, scientists have conducted extensive research on various aspects of electronic document processing. In practice, much information is still stored in paper documents, including technical reports, government files, newspapers, books, journals, magazines, letters, and bank checks, to name just a few. Many of these collections have been scanned, indexed, OCRed, and placed in corporate intranets to be used to gain competitive advantage. Retrieving them requires that their contents be recognized. Despite all the research that has been done in document image processing, several problems are still commonly encountered in this field [3]. Electronic documents produced by scanning and OCR software contain recognition errors, and the rate of errors increases significantly with the degradation of the document image. Such documents may then become inaccessible using conventional retrieval methods that affect the retrieval results significantly.

Information retrieval (IR) is the process of determining relevant documents from a collection of documents based on a query presented by the user. Research has been conducted on the interaction between OCR and IR since 1980 and has consistently shown that the results of operations based on the exact matching of string attributes are often of lower quality than expected. For example, consider a corporation maintaining various image databases. A specific customer name

might be present in more than one image. In one image, a customer name might be recorded as "riemannian", while in other image databases the same name may be recognized as "ricmanuinn" or as "licmamian". A request to correlate these images and create a unified view of users will fail to produce the desired output if exact string matching is used in the retrieval process.

The goal of this research is to design an error-gram (sequential confused characters within each erroneous word) tool that could be used to expand queries, and hence to OCR document images of varying quality. By confused character we mean characters from a document image that are wrongly recognized by an OCR system. In this study, we will take into account the generation of short erroneous substrings of n-grams of the confused characters in words and process them using standard methods available in IR. We have incorporated edit distance to determine possible OCR errors, which have occurred by confronting the OCR text with that provided by the ground truth, to collect frequent error-grams and construct their corresponding correction rules. Our method takes advantage of dynamic programming ability to generate error-grams derived from erroneous substrings, which are introduced to extend query terms. In addition, error-grams are weighted depending on their frequencies. These weights are used in an IR vector space model in which each document is represented by a vector and where each element reflects the importance of a particular term in the document and the collection. We compare the user's query to these document images through basic vector operations and rank the retrieved images in decreasing order of their similarity to the query.

To show the effectiveness of the system, different degradations are considered. Degraded images are obtained from ideal images through a degradation model, or by physically degrading (printing, scanning, faxing, etc.) a hard copy. The quality of the original document can be a problem for the following reasons.

1. The original is old and has suffered physical degradation.
2. The original is produced by a manual typewriter so the individual characters may show variations in print quality, contrast, and position.
3. The original is a low-quality photocopy and shows variations in toner density and character spread.

A commercial OCR was applied to different kinds of document images (article, newspaper, advertisement, business card, manual, form, and degraded images). Three sets were processed, with the first one considered for training and the next two for testing. Error-grams and correction rules were first generated using the training set and then combined to extend query terms. The experiments were performed and showed a marked improvement in retrieval performance when compared to standard methods such as vectorial systems without expanded query and 3-gram overlapping, as defined in Sect. 5.4.

In this paper, we describe an approach to enhancing the retrieval performance on OCR data obtained from document images of different levels of quality. The remainder of the paper is organized as follows. The next section presents previous studies. In Sect. 3, we define a framework of the retrieval process based on OCR errors. Section 4 categorizes various OCR errors and uses matching algorithms to construct error-grams. Section 5 presents the retrieval process and performance measures. Section 6 shows experimental results comparing the most efficient algorithms applied to three independent sets: training, test, and degraded-test sets of document images. The conclusions, discussion, and future work directions are presented in Sect. 7.

2 Related work

Information retrieval serves to search large textual databases and return documents the system considers relevant to the user's queries [6]. Smeaton [14] uses the approximate shape of words in a text to refine the retrieval process; however, this approach cannot disambiguate recognition errors. Text retrieval from document images is difficult because OCR errors derive from editing operations such as character substitution, deletion, and insertion [5, 7, 8, 22].

Previous studies have tried to reduce errors through correction steps. Most approaches to the correction of scanning errors make use of the lexicon. Errors are detected by searching the text for words that do not appear in a lexicon [6, 17]. This leads to many false alarms, since a lexicon cannot possibly cover everything. Many studies in this area [5, 7, 8, 22] show that three common mistakes – character substitution, deletion, and insertion – make up 80–90% of all typing errors.

During the 1990s the Information Science Research Institute (ISRI) at the University of Nevada, Las Vegas conducted many experiments to study OCR accuracy and retrieval effectiveness from OCR-generated texts [19–22]. They showed the effects of OCR errors on ranking and feedback using the vector space model. Feedback is provided by an automatic process that uses information derived from known relevant and non-relevant documents to reformulate queries, but cannot be used to compensate for OCR errors caused by degraded documents. Taghva and Stofsky [22] developed OCRSpell, which uses a special parser, domain-specific dictionaries, and a statistical tool mapping word generator to create a list of word candidates to replace incorrect terms. OCRSpell is used in [20] to deal with typical OCR errors in texts, to avoid the extreme variability in ranked sets, and to improve retrieval effectiveness from poorly recognized document collections.

Expanded queries with error-generating tools such as "error-grams" for improving document retrieval of OCR texts have not been studied sufficiently. Automatic query expansion is a way of evaluating the potential usefulness of correcting OCR errors. Several algorithms are available in the current literature for automatic query expansion [10, 11, 16], and our goal is to include erroneous words which can have a relationship with the terms of a user's query. We believe that query expansion improves retrieval

effectiveness, especially when erroneous terms appear as proper nouns or in short documents, because of a lack of redundancy. Croft et al. [5] extend query terms by using n-grams contained in query words. This method needs better closeness measures in order to eliminate spurious terms in the expansion. Ohta et al. [8] present a probabilistic text retrieval method for carrying out full-text searches of English documents containing OCR errors. The validity of retrieved terms is determined based on the occurrence of confused characters and the connection with their preceding and succeeding characters. All possible error information included in the confusion matrices increases the recall rate but significantly decreases the precision rate. Suen [18] tabulated the growth in the number of distinct n-grams as a function of vocabulary size, their word-positional dependence, and the influence of the selected corpus.

In a recent work [4], we introduced error-grams along with a Boolean retrieval system. An error-gram refers to the part of the word which was not correctly recognized. The method needed a test collection and validation using degraded images. A survey of document image degradation models proposed in the existing literature can be found in [1]. Furthermore, as shown in [13], the vector space IR model improves the Boolean representation by synthesizing a document's content not only through a set of terms but also by considering the importance of the terms in documents and their specifics in the collection. After creating a set of augmented and weighted query terms with the correction rules, a SMART [12], vector-based retrieval system developed at Cornell University is used to retrieve relevant documents and to evaluate retrieval performance.

Finally, regarding the OCR package we used, our choice is based on a recent work by Souza et al. [15] in which different commercial OCRs were studied with the goal of assessing criteria for filter selection using a variety of image qualities. We used the best OCR because it obtained the highest recognition rate.

3 System architecture of the proposed approach

The system architecture is shown in Fig. 1. The approach is described by the following three stages.

1. In the first step, we start from the three sets of images (i.e. training, test, and degraded test). The first set contains images and associated ground-truth data (electronic text version) and is used for constructing error-grams and training retrieval systems. The ground truths associated with databases are the zones on each image and the corresponding ASCII text for all the text zones. The second is for system testing using 200 advertising and 100 Webpage images and ground truths (ASCII text) for each image, which is completely different from the training set. Furthermore, the latter collection was used to produce an

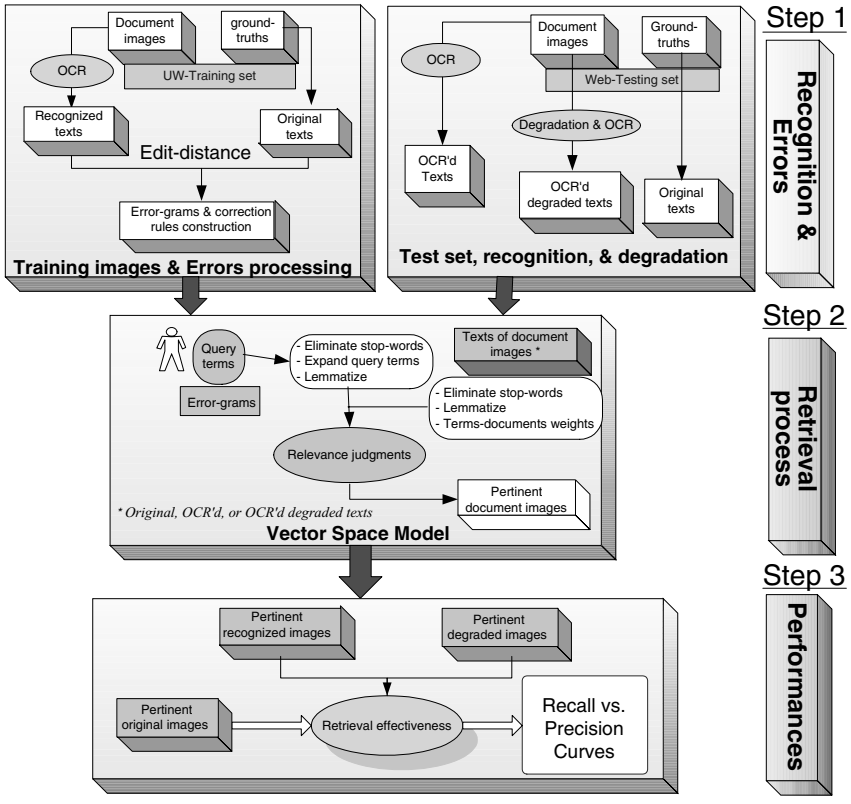

Fig. 1 Retrieval approach based on OCR errors

additional set of 700 images (third set) with various types of degradations. We used three independent techniques to produce an overall effect of degradation:

(a) Pixel noise: any number of black pixels added randomly to the content image.
(b) Blurred noise: each character (actually any connected set of black pixels) is grown by a small number of pixels along the boundary of the character. Each pixel is replaced by the average of pixels around it. This procedure generates the familiar blurring of character boundaries produced by a photocopier or a scanner.
(c) Repetitive photocopies: the spread of black and white pixels is also variable. White pixels can create broken characters, while black pixels can create touching characters.

To measure the mapping between the input image and its corresponding OCR output, we need a distance function to solve the problem of proximity matching. We use the edit distance between substrings. There is a match routine that detects any common segment between the original word (ASCII text in ground truth) and each of the OCR words. The output of the match routine is a distance which models the transformations that render the two words identical. In this step, we apply editing operations on OCR words, generating a collection of error-grams, and collecting original, OCR, and OCR-degraded texts.

2. The second step uses query terms, error-grams, ASCII texts in ground truth, and OCR texts to create searchable keywords, eliminate stop words, lemmatize words, identify appropriate matching terms, and apply a vector space model for indexing and determining relevant document images.
3. Finally, we measure the performance of the retrieval system and compare different methods to show the improvement in retrieved document images.

With the proposed method, given a scanned image, the user:

1. Locates and extracts text objects in it.
2. Compares OCR-recognized text with the original text (ground truth). Mistakes are modeled as a set of error-grams and correction rules.
3. Measures the effectiveness of retrieval systems using document ranking, recall, and precision.

Details of the algorithm related to Fig. 1 are presented in the following sections.

4 Matching OCR errors

A commercial OCR was used to read the located text in the training set to perform character recognition. The OCR engine made mistakes (Table 1 shows examples).

Our hypothesis is that differences in the observed frequency between the original input (ASCII text in ground-truth set) and recognized OCR output texts would indicate that the n-gram substring in question was incorrectly recognized. Since counting errors by hand is time consuming, a simple error measure – edit-distance – was adopted.

Table 1 Error groups with real examples

Error group	Correct word	Error example
Substitution	Light	Right
Deletion	Info	Nfo
Insertion	Kylie	Ikylie
Paste or split	n-gram	n-gram

4.1 Edit-distance

The edit-distance algorithm is based on dynamic programming and matches strings without lexicons or a priori information [2]. The distance between two words equals the number of editing operations required to transform one of the words into the other.

Let M_{ori} be the set of words contained in the original document, M_{ocr} the set of words contained in the recognized document, and $s = e_1; e_2; \ldots; e_n$ a sequence of edit operations for transforming a string x into another string y. The costs $c(s)$ of this sequence are given by $c(s) = \sum_{i=1}^{n} c(e_i)$, where $c(e_i)$ is the cost of the ith edit operation.

Given two strings x and y and given the cost of any edit operation which may be required to transform x into y, we define the distance between x and y by

$$\mathrm{d}(x, y) = \{\min\{c(s)\} : s \text{ is a sequence of edit operations which transform } x \text{ into } y\}$$

In set notation, we have correctly recognized words

$$M_{\text{rec}} = \{\text{words} \in M_{\text{ori}} \cap M_{\text{ocr}}\}$$

and remaining words

$$M_{\text{remo}} = \{\text{words} \in M_{\text{ori}} - M_{\text{rec}}\},$$
$$M_{\text{remr}} = \{\text{words} \in M_{\text{ocr}} - M_{\text{rec}}\}.$$

We now show how this measure is adapted to construct error-grams and correction rules in document images. The algorithm [23] used to compute the edit-distance () is based on dynamic programming. It fills the matrix $D_{0\ldots|x|,0\ldots|y|}$, where $D_{i,j}$ represents the minimum number of operations to match strings $x_{1\ldots i}$ to $y_{1\ldots j}$, x is a string, $|x|$ is its length, and x_i is the ith character of x. The costs relating to the editing operations are set to 1 in this study. The algorithm below calculates the distance gradually in order to match two strings:

$$\begin{cases} D_{i,0} = 0; \\ D_{0,j} = 0; \\ D_{i,j} = \{ \text{ if } (x_i = y_j), \quad \text{then } \{D_{i-1,j-1}\} \\ \qquad\qquad \text{else } (1 + \{\min(D_{i-1,j}, D_{i,j-1}, D_{i-1,j-1})\})\}. \end{cases}$$

4.2 Error-grams and correction rules

Error-grams are confused portions in erroneously recognized words (sequential confused characters inside each erroneous word). For example, if the word "schultz" is recognized as "sehnltz", one error-gram considered will be "chu". It can be replaced in OCR text by "ehn". Our algorithm processes the words which appear only in the remaining original words M_{remo}. It uses the edit-distance to find the nearest word in M_{remr}, to locate the erroneous substrings in the recognized word called error-grams, and to define the correction rules based on the corresponding substrings in the original word. We then verify the pairing, extract the immediate predecessor and successor for each confused character, and classify the n-grams extracted in order of occurrence. The algorithm works as follows:

1. For each word $x_i \in M_{\text{remo}}$
   Scan words $\in M_{\text{remr}}$ and Select $x_j \in M_{\text{remr}}$ that $d(x_i, x_j)$ is the minimum;
2. Locate errors and verify the matching accuracy between the badly recognized word and its correspondent. For each incorrectly recognized word, extract the confused characters and their neighbors to constitute the error-grams and its corresponding correction rules.
3. Calculate weights to quantify the importance of the errors and to evaluate their pertinence in the retrieval process. The correction rules contain the probabilities that any character A_i in the document image can be regarded as B_j in the OCR text, which is calculated using the total probability formula:

$$P(B_j) = \sum_{A_i} P(B_j \mid A_i) P(A_i)\,, \tag{1}$$

   where $P(B_j|A_i)$ denotes the conditional probability of B_j assuming that A_i has occurred,

$$P(B_j \mid A_i) = \frac{P(A_i \mid B_j) P(B_j)}{P(A_i)}\,. \tag{2}$$

5 Retrieval process

Information retrieval is about finding the relevant information in a large text collection, with string matching being one of its basic tools. However, exact string matching is not good enough for document image retrieval because a word, when recognized incorrectly in the database, can no longer be retrieved. When data are noisy or corrupted, as is the case with OCR texts, exact string matching becomes inappropriate, and another measure is needed to facilitate information retrieval from the collections of OCR text.

The main goal of our approach is to process document images and to expand the query into other possible terms. The document processor prepares, processes, and inputs the documents that users are searching for. It identifies potential indexable elements in documents, deletes stop words, stems terms, extracts index entries, computes weights, and creates and updates the main inverted file against which the search engine searches in order to match queries to documents. The expanded, weighted query is searched against the inverted file of documents obtained by the $M \times N$ document matrix, where M is the number of documents in the collection and N is the number of unique terms in the collection. The similarity of each document is calculated in the subset of documents, and the system presents an ordered list to the user.

Conceptually, our retrieval system is composed of three submodules which:

1. Add to the query words generated by initial query words, the error-grams, and the correction rules; assign weights to the obtained list for use in the retrieval process;
2. Extract document images relevant to the user's query. An indexing component is responsible for recognizing indexes from raw documents and constructing a searchable index structure (inverted file). Documents can only be added to the document collection by going through this component;
3. Employ a retrieval component that operates on query and document representations, decides what to return in response to a query, and in what order.

Finally, we measure the performance of the retrieval system and compare different methods.

5.1 Query expansion and selection

For every query word, we add the words generated by substituting all error-grams contained in the term with their corresponding correction rules. Let us first give an example of the query expansion. Suppose that an original query contains the word "light". It is statistically uncertain because OCR confuses "i" with "l" and "g" with "e", etc. Through the error-grams we know that the words "*llght*", "*lighl*", "*right*", etc. (32 words) are strongly related to "light". Then, the expanded query with a probability higher than 0.001 will be
$\langle light; llght; ligit; lighd; lieht; iight; lighl; ligbt; right\rangle$
and the weights of the expanded query are
$\langle 1; 0.096; 0.0092; 0.0086; 0.009; 0.0036; 0.0032; 0.004; 0.001\rangle$.

Some generated words can cause noise effects and confusion in the answers. For example, the word "right" above is used as an extended term, and its use harms the meaning of the user's request. This can, however, also help in identifying the documents in which "light" has been mistakenly recognized as "right". As we use the vectorial model, the affected weight 0.001 with the word "right" in our example will influence the order of relevance for the documents which contain it.

5.2 Indexing component

Documents are usually described through a set of terms. A common automatic indexing strategy is to take the set of all words found in the document as terms, remove the most

common words, such as "the" and uninteresting terms such as "thing", and stem the remaining terms to get "image" from "images" and "imaging". The remaining terms constitute the set of index terms.

The vector space model uses vectors to represent documents in a database and queries. A vector is obtained for each document and query from sets of terms with associated weights. The weights could be the frequency of occurrences of the word in the document. We could also assign a more sophisticated weighting scheme for each term. One of the most popular ways of creating weighting vectors is through the tf*idf family of weighting schemes. The term frequency component (tf) of a term t_i for a document d_j is calculated according to

$$\mathrm{tf}_{ij} = \frac{\mathrm{frequency}_{ij}}{\mathrm{Max}_l \ \mathrm{frequency}_{lj}} ,$$

where $\mathrm{Max}_l \ \mathrm{frequency}_{lj}$ is the frequency of the most common term in the document. The inverse document frequency idf is computed as follows:

$$\mathrm{idf}_i = \log \frac{N}{n_i} ,$$

where N is the total number of documents in the database and n_i is the number of documents that contain the term t_i. In the tf*idf weighting scheme, the component of the weighting vector for the document d_j at position i (i.e. for term t_i) is

$$d_{ij} = \mathrm{tf}_{ij} \ . \ \mathrm{idf}_i \ .$$

5.3 Similarity calculation

With the above measures we can use cosine similarity $\mathrm{sim}(q, d_i)$ between a query $q = q_1; q_2; \ldots; q_t$ and a document $d_i = d_{i1}; d_{i2}; \ldots; d_{it}$ to determine how close they are to each other geometrically. The cosine similarity is calculated using the following formula:

$$\mathrm{sim}(q, d_i) = \frac{\sum_{j=1}^{t} q_j \ d_{ij}}{\sqrt{\sum_{j=1}^{t} q_j^2} \sqrt{\sum_{j=1}^{t} d_{ij}^2}} ,$$

where d_{ij} is the weight of the term t_j in the document d_i and q_j is a query term determined as follows:

$$q_j = \begin{cases} 1 & \text{if } q_j \text{ is a query term,} \\ \prod_{j=1}^{s} \Pr(q_j) \ . \ \mathrm{idf}_j & \text{if } q_j \text{ is an extended term,} \\ 0 & \text{otherwise .} \end{cases}$$

In the above formula, s is the number of error-grams used in the expanded query term and $\Pr(q_j)$ corresponds to the probability of the error-gram used to construct q_j.

Documents are then ranked in the order of their similarity to the query. Documents whose similarities exceed a certain threshold are retained in the response list, while all others are rejected as being irrelevant.

5.4 Performance measures

Performance is determined by the retrieval of randomly selected queries. The lists of relevant document images on the basis of original texts are compared with those obtained by using OCR texts. The evaluation of the different methods is based on the retrieval effectiveness using average values of the rated recall and precision, which are calculated from the following equations:

(i) RECALL is a measure of the ability of the system to present all relevant images. It is calculated as

$$\mathrm{Recall} = \frac{\text{total of relevant images retrieved}}{\text{total number of relevant images}} .$$

(ii) PRECISION is a measure of the ability of the system to present only relevant images. It is calculated as

$$\mathrm{Precision} = \frac{\text{total of relevant images retrieved}}{\text{total number of images retrieved}} .$$

(iii) Quality-distance (QD) is used to measure the performance of the 3-gram overlap approach. This approach consists of decomposing the string T into a following succession of three characters $((c_i, c_{i+1}, c_{i+2}))_{i \in [1, n-2]}$, where T is a query term with a length of n characters and c_i its ith character. This difference between two strings x and y is measured as the quality distance QD; $QD(x, y)$ is the number of 3-grams contained in two words versus the number they share:

$$QD(x, y) = G(x) + G(y) - 2(G(x) \cap G(y)) ,$$

where $G(x)$ represents the number of 3-grams contained in a word x.

(iv) The precision averages (at 11 standard recall levels – from 0 to 100%) are used to compare the performance of different methods and as the input for plotting the recall-precision graph.

6 Experimental results

6.1 Data collection

Three sets of document images were chosen. The first was a training set which was used to construct error-grams; the document corpus was the technical document database of the University of Washington [9], with 979 journal pages from a wide variety of journals covering diverse subject areas and publishers. The average size per page was 510 words.

The second was a Test 1 set of 100 images obtained and printed from the Web (the average size per page was 410 words). This corpus was then degraded using an image processing model and by repeated photocopying. We

(a) A portion of an original image

(b) Degraded with 2 times photocopies

(c) Added noise and blurring using 3x3 window

(d) Added noise and blurring using 5x5 window

Fig. 2 Example of original Web page and its degraded versions

used a combination of noise and blurring to form an additional set of degraded images. Noise degradation was applied first, using three different standard deviations, and then blurred using average operations in the convolution of the input image with two window sizes [the value of each pixel in the output image is based on a comparison of the corresponding pixel in the input image with its 8 (24) neighbors for a 3×3 (5×5) window size]. Repeated photocopying caused character breaks and added black and white areas at random throughout the page. The new degraded-test set contained 700 images obtained from the above test set of 100 Web-page images and seven types of degradation. Figure 2 shows an original Web page and three types of degradations.

The third was a Test 2 set of 200 images obtained from the Media Team document image database provided by the University of Washington. They came from various fields (business cards, advertisements, manuals, and forms). The average size per page was 44 words for advertising and 304 words for manual and form images. This corpus was used to test the robustness and sensitivity with images containing names of people or places and short texts.

For query collection, we evaluated 50 queries, randomly selected from the content of documents. Each query contained an average of three words.

To measure the performance, we considered the documents returned by SMART for each query and based on electronic texts provided by the U-W-1 database as relevant. The documents retrieved by different systems were then compared with the supposed relevant documents to determine, for each query, whether or not they were relevant.

6.2 OCR recognition

6.2.1 Training phase

The results obtained using an edit-distance are presented in Table 2. In the original images, we had 614 non-text fields, which explains the higher number of words present in the OCR texts than the original texts: 499,123 words were present in the original documents while the OCR extracted 528,315 words. Only 468,619 words out of the 499,123 were correctly recognized. The dynamic programming algorithm with distance 2 matched 5,185 words, and we used them to build the error-grams. Note that we can improve recognition by reducing noise and using features acquired to distinguish text that is considered as noise.

We obtained 6,933 substitutions, 2,216 deletions, and 2,319 insertions. The output of the edit-distance algorithm will serve as the input for the error rule-building algorithm to construct the error-grams and the correction rules. The algorithm constructed 2,822 error-grams and correction rules.

Table 2 Text recognition using edit-distance on the training set; 979 scanned images were recognized by commercial OCR

	Words	Characters	Recognition
Original image	499,123	2.9 MB	
OCR extracts	528,315	3 MB	
Correct words	468,619	2.74 MB	93.8%
Distance ≤ 2	5,185	30,591	1.03%
Total recognition	473,804	2.78 MB	94.83%

Table 3 Top 20 error-grams with their correction rules. A_i and B_j are n-grams in the original and recognized word, respectively. P is the probability that A_i can be regarded as B_j

A_i	B_j	P	A_i	B_j	P
th	di	0.12	i	1	0.064
i	l	0.096	th	dh	0.059
h	i	0.092	l	1	0.05
th	ti	0.089	l	i	0.036
t	d	0.086	y	v	0.034
r	l	0.086	z	s	0.033
th	dh	0.077	t	l	0.032
he	ie	0.066	the	die	0.029
the	tie	0.065	e	0	0.026
t	1	0.064	ize	ise	0.024

Table 4 Text recognition and OCR errors on Test 1 collection images. One hundred Web-page images degraded by photocopying twice

	Without degradation	Repetitive photocopies
Number of words	40,642	40,642
OCR extracts	34,318	22,600
Number of correct words	29,368	15,278
% of correct words	72.26	37.59

Table 3 shows the top 20 error-grams and the probability P that error-gram A_i in the original image can be regarded as B_j in the OCR texts.

6.2.2 Test phase

Tables 4 and 5 show the results of the recognition of the test and the degraded-test collections. We note the decrease in the performance of the recognition on the degraded images and observe that added Gaussian noise does not affect recognition accuracy. Figure 3 shows that the recognition by OCR resists noise, but the performance falls as the blur of the characters grows.

The recognition rate in the Test 1 set (73%) is lower than that in the training set (93%) because the resolution and the quality of the printed Web images are worse.

A degraded document image will yield a low recognition rate when it is submitted to OCR software. Many factors such as font size, broken character, touching characters, and white speckles are used to indicate the image quality. Experimental results on the test and degraded sets exhibit a significant decrease in the recognition rate from 93.8% on the training set to 72.26% with the test collection and to about 10% with a high blur degradation. Two-pass photocopying produces a recognition rate of about 38%, which has the same effect as degradation blurred by eight neighbour pixels.

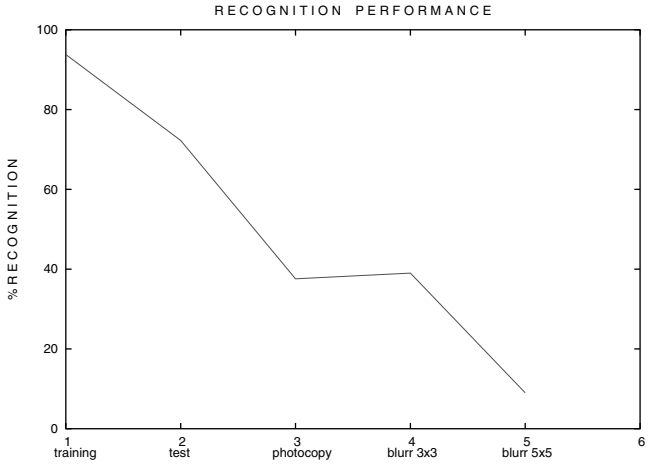

Fig. 3 Recognition degradation on test and degraded images

6.3 Retrieval effectiveness

The recall-precision graph is the most commonly used method for comparing systems. The plots of different runs can be superimposed on the same graph to determine which system is superior. Comparisons are best made in three different recall ranges: 0–0.2, 0.2–0.8, and 0.8–1. These ranges characterize low-recall, mid-recall, and high-recall performance, respectively.

6.3.1 On training set

In the retrieval process, performance is determined by the retrieval of 50 randomly selected queries. Each query contains several words. Our method is compared with the SMART-based vectorial model without expanded query (called Smart in different figures) as well as with Ukkonen's [23] Q-gram for quality distance = 1 and 4. Figure 4 shows that, in the training phase, our approach achieves an improvement in terms of recall and precision.

In Table 6, for the vectorial model without expanded query (Smart), the average precision is between 97.81 and 99.84% for the low-recall, between 83.62 and 96.95% for

Table 5 Text recognition and OCR errors on Test 1 collection images

Blurring window		3×3			5×5	
Gaussian noise σ		0.01			0.01	
Number of words	0.1	40,642	0.001	0.1	40,642	0.001
OCR extracts	20,356	20,528	20,680	10,870	9,980	10,294
Correct words	15,906	15,730	16,170	3,728	3,310	3,530
% correct words	39.14	38.70	39.79	9.41	8.35	8.91

100 Web-page images degraded by a Gaussian noise with three standard deviations and a blurring with two window sizes

Table 6 Recall-precision results on training set

	Low-recall	Mid-recall	High-recall	Average precision
Smart	97.81–99.84	83.62–96.95	23.53–76.08	86.53
Best 3-gram overlaps	94.64–99.9	38.98–89.22	0–29.35	65.99
Our method	97.50–99.72	87.07–95.35	36.61–84.44	87.68

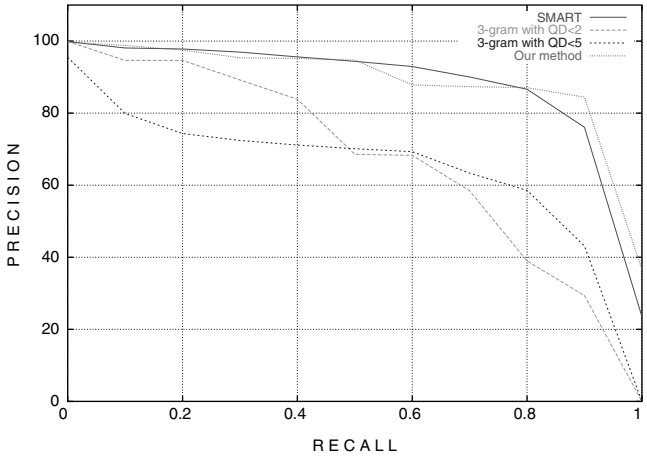

Fig. 4 Recall-precision averages on training collection

Table 7 Retrieval effectiveness when searching for Test 2 set

Retrieval method	Average recall (%)	Average precision (%)
Our approach	68.40	88.82
Smart without expansion	41.88	91.32
3-gram overlaps		
QD ≤ 1	63.46	70.08
QD ≤ 4	75.76	60.64

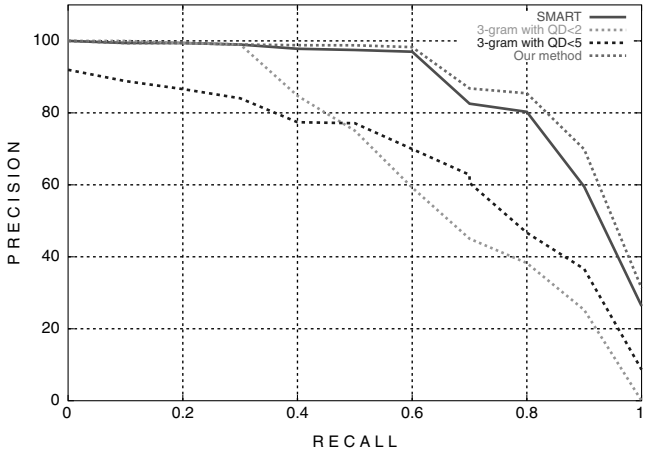

Fig. 5 Recall-precision averages on Test 1 collection

the middle-recall, and between 23.53 and 76.08% for the high-recall performance. The "3-gram overlap" technique extracts a broad range of words and results in a fall of precision at high recall levels. For 3-gram overlaps with a large QD distance, the recall increases but the precision decreases. In addition, the average precision for all relevant documents (averaged over queries) is 87.68% for our approach but 86.53% for a vectorial search without expanded query and does not exceed 65.99% for 3-gram overlap methods.

Our approach has the advantage of using only the parts of the required word likely to be incorrectly recognized. It presents the best performance at high recall. We observe that our approach achieves better overall retrieval effectiveness compared with other methods. This is due to the statistical characteristics of extracting and classifying expanded words based on their importance, relative to the confusions, in training. An example with the query input "Is schultz in some journals?": after removing common words and stemming, query terms become "schultz journ" and the expanded query using 3-gram overlap is "schultz journ sch chu hul ult ltz jou our urn". If the word "schultz" is recognized in a document as "sehnltz", the QD distance between these words is 4. But with our method the expanded query contains the terms "sehnltz" with 0.002 probability and the document image is ranked in the relevant list.

To better appreciate the performance of our approach, we should mention that in the second range (middle recall), our system outperforms Smart by 4%, and it goes up to 13.08% in the third range (high recall). This will help to understand the importance of modelling errors and to exploit this misrecognition in the retrieval process.

6.3.2 On Test 1, Test 2, and degraded-test sets

It is interesting to extend these experiments to a wider range of document images. To do this, we tested our system and compared the results for original and degraded collections based on retrieval performance. The Test 1 and Test 2 collections used to obtain the results presented in Fig. 5 and Table 7 show the same tendency as that of the training set, except for the 3-gram overlap approach, with a small distance, which is able to perform well in the high precision field. This is due to the smaller number of test document images we have compared with the training set.

For the Test 1 set, the average precision for all relevant documents (averaged over queries) is 87.90% for our approach, but 85.33% for the vectorial search without expanded query, and it does not exceed 65.99% for other methods.

For the Test 2 set, the results obtained in Table 7 show that our approach has the best overall precision and concurs with the recall of the best 3-gram overlaps. The explanation for this is obtained from weak frequency words in the document, such as people or place names, and which are badly recognized. Only through our method the example "Tauberian", which appears twice in a document and is recognized as "Tauoenan", can be retrieved with a 0.0009 probability. We mention that query expansion improves retrieval effectiveness, especially when erroneous terms appear as proper nouns or in short documents due to a lack of redundancy. The influence of the added terms starts when the recall becomes high.

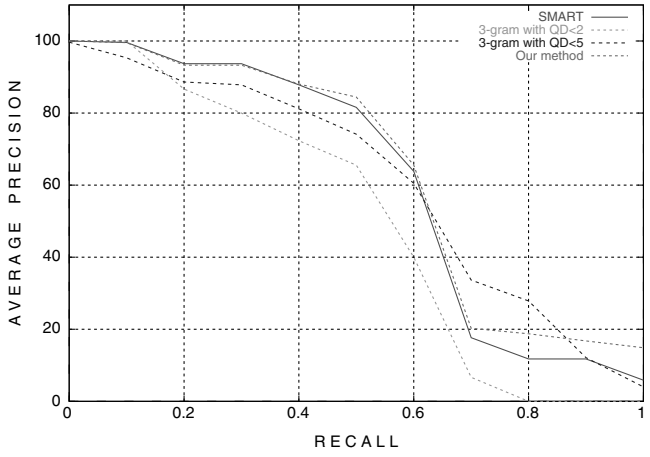

Fig. 6 Recall-precision averages on photocopy collection

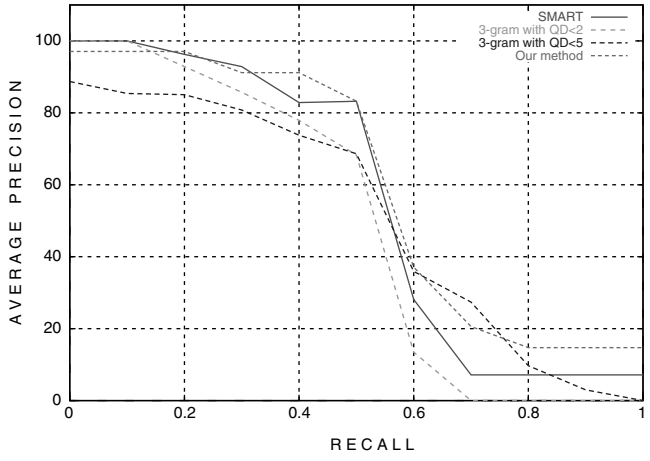

Fig. 7 Recall-precision averages on degraded collection

For degraded document images, we know that recognition accuracy is very low. We can see in Figs. 6 and 7 that the precision is maintained with an upper limit of 65.54% for recalls lower than 50%. Beyond that the performance decreases and all the methods follow the same tendency. We note that the performance rate is better with our expanded queries at high-recall performance for photocopied sets. The average precision for all relevant documents over all queries is 63.16% for our method, but decreases to 60.44% for 3-gram overlaps with large distances and to 60.66% for standard vectorial searches without expanded queries. We see the same tendency for degraded images, where the average precision for all relevant documents over all queries is 59.86% for our method and 55.62% with the standard vectorial system. However, the rate decreases to 50.74% for 3-gram overlaps with wide distances. This is due to the power of 3-gram overlaps in extracting any parts of words which appear in the text and to the low number of images in the test collection.

An indication of the results obtained for the very degraded set can be seen in Fig. 8, which shows a drop in the precision rate. We note that the "3-gram overlaps" remain

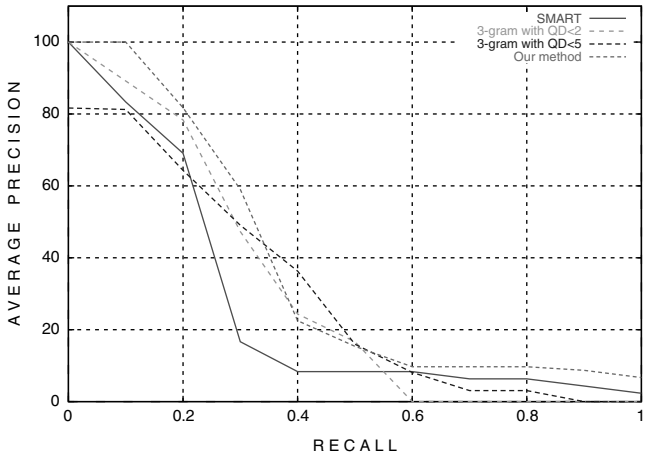

Fig. 8 Recall-precision averages on very degraded collection

better in the middle recall. The problem with the 3-gram approach is the fast drop in precision when the quality distance (QD) increases. Between the high-precision and high-recall fields, the average precision becomes lower than 50% for the middle-recall and less than 3% for the high-recall performance, except for our method, which is maintained at 6.69%. The average precision for all relevant documents over all queries is 38.46% for our method but decreases to 33.64% for 3-gram overlaps with a small quality distance and 28.48% for standard vectorial search.

We can see from the retrieval performance figures that our algorithm performs better than other methods on degraded-quality images. We have a better closeness method for eliminating spurious terms in the expansion query, and it has improved the retrieval performance in document images. It is interesting to see how the image quality affects retrieval performance. All approaches yield almost the same results for very degraded images.

7 Conclusion and perspectives

This work presents an approach to processing textual information contained in document images and for performing effective retrieval. String processing in a textual corpus is a very fertile and useful research area. Current OCRs do not work well on poor-quality or scanned document images. Three different sets of document images were used for training, testing, and validating. The proposed method collects frequent error-grams and correction rules that can be used to extend query terms and to improve retrieval performance. We have shown that an n-gram and its corresponding probability can greatly influence the results returned by the standard cosine measure. Furthermore, investigating the OCR of poor-quality documents is important for document images generated from archives of originals created before the dawn of the digital age. We have discussed the use of one OCR engine and how the nature of the degradation can affect the accuracy of the resulting OCR effort.

One hundred Web pages along with their degraded images and 200 advertising and manual images were tested in order to validate any increase in retrieval effectiveness by using error-grams to expand query terms. Experimental results indicate a noticeable improvement in the retrieval effectiveness as compared to exact, partial, and 3-gram overlap matching. Further research is currently being undertaken to outperform our approach. The aim is to investigate the ability of this approach to improve retrieval effectiveness. Possible techniques include:

– Additional image preprocessing, which may increase the OCR recognition rate.
– An iterative edit-distance technique with various cost functions to match more words and to increase the recognition rate.
– Use of different training sets of document images to create other kinds of error-grams and hence correction rules to increase information retrieval performances.
– Possible changes to some of the correction rules on which we based our approach. A potential solution is to rebuild these "error-grams" for very degraded images in order to produce better results.
– Use of a combination of a number of OCR engines and the derivation of a voting system for erroneous words to increase the quality of results.

References

1. Baird, H.S.: Document image quality: making fine discriminations. In: Proceedings of the IAPR International Conference on Document Analysis and Recognition, vol. 11, pp. 1209–1223. Bangalore, India (1999)
2. Bunke, H., Csirik, J.: Parametric string edit-distance and its application to pattern recognition. IEEE Trans. Syst. Man Cybern. **25**, 202–206 (1975)
3. Fataicha, Y., Cheriet, M., Nie, J.Y., Suen, C.Y.: Content analysis in document images: A scale space approach. In: Proceedings of the 16th IEEE International Conference on Pattern Recognition. vol. 3, pp. 335–338. Quebec (2002)
4. Fataicha, Y., Cheriet, M., Nie, J.Y., Suen, C.Y.: Information retrieval based on OCR errors in scanned documents. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recogntion Workshops (CVPR'03), vol. 3. Madison, WI (2003)
5. Harding, S.M., Croft, W.B., Weir, C.: Probabilistic retrieval of OCR degraded text using n-grams. In: Proceedings of the 1st European Conference ECDL. vol. 1324, pp. 345–359. Pisa, Italy, Research and Advanced Technology for Digital Libraries (1997)
6. Mäkinen, V., Baeza-Yates, R., Riberro-Neto, B.: Modern Information Retrieval, p. 513. Addison-Wesley, Reading, MA (1999)
7. Mäkinen, V., Navarro, G., Ukkonen, E.: Algorithms for transposition invariant string matching. In: Proceedings of the STACS 20th Annual Symposium on Theoretical Aspects of Computer Science. Lecture Notes in Computer Science, vol. 2607, pp. 191–202. Springer, Berlin Heidelberg New York (2003)
8. Ohta, O., Takasu, A., Adachi, J.: Probabilistic retrieval methods for text missrecognized OCR characters. IEEE Trans. Pattern Anal. Mach. Intell. **22**(11), 1224–1240 (1998)
9. Phillips, I.T.: User's reference manual for the uw english/technical document image database. UW-I English-Technical Document Image Database, University of Washington (1993)
10. Rijsbergen, C.J., Harper, D.J., Porter, M.F.: The selection of good search terms. Inf. Process. Manage. **17**, 77–91 (1981)
11. Rijsbergen, C.J., Harper, D.J., Porter, M.F.: The selection of good search terms. Inf. Process. Manage. **17**, 77–91 (1981)
12. Salton, G.: The Smart Retrieval System-Experiments in Automatic Document Processing. Prentice Hall, Englewood Cliffs, NJ (1971)
13. Salton, G., McGill, M.: Introduction to Modern Information Retrieval, p. 1. McGraw-Hill, New York (1983)
14. Smeaton, A.F.: Retrieval images of scanned text documents. In: Proceedings of the Optical Engineering Society of Ireland and Irish Machine Vision and Image Processing Joint Conference (OESI-IMVIP), pp. 271–286 (1998)
15. Souza, A., Cheriet, M., Naoi, S., Suen, C.Y.: Automatic filter selection using image quality assessment. In: Proceedings of the 7th International Conference on Document Analysis and Recognition ICDAR'03, pp. 508–512. Edinburgh, UK (2003)
16. Spink, A., Saracevic, T.: Interactive information retrieval: Sources and effectiveness of search terms during mediated online searching. J. Am. Soc. Inf. Sci. **48**(8), 741–761 (1997)
17. Strohmaier, C.M., Ringlstetter, C., Schulz, K.U., Mihov, S.: Lexical postcorrection of OCR-results: The web as a dynamic secondary. In: Proceedings of the 7th International Conference on Document Analysis and Recognition ICDAR'03, pp. 1133–1137. Edinburgh, UK (2003)
18. Suen, C.Y.: N-gram statistics for natural language understanding and text processing. IEEE Trans. Pattern Anal. Mach. Intell. **1**, 164–171 (1979)
19. Taghva, K., Borsack, J., Condit, A.: Effects of OCR errors on ranking and feedback using the vector space model. Inf. Process. Manage. **32**(3), 317–327 (1996)
20. Taghva, K., Borsack, J., Erva Condit, S.: Hairetes: A search engine for OCR documents. In: Proceedings of the 5th IAPR International Workshop on Document Analysis Systems, pp. 412–422. Princeton, NJ (2002)
21. Taghva, K., Borsack, J., Condit, A.: Evaluation of model-based retrieval effectiveness with OCR text. ACM Trans. Inf. Syst. **14**(1), 64–93 (1996)
22. Taghva, K., Stofsky, E.: Ocrspell: An interactive spelling correction system for OCR errors in text. Int. J. Doc. Anal. Recog. **3**(3), 125–137 (2001)
23. Ukkonen, E.: On approximate string matching. In: Proceedings of the International Conference on Foundations of Computer Theory, pp. 487–495. Lecture Notes in Computer Science, vol. 158. Springer, Berlin Heidelberg New York (1983)

Youssef Fataicha received his B.Sc. degree from Université de Rennes1, Rennes, France, in 1982. In 1984 he obtained his M.Sc. in computer science from Université de Rennes1, France. Between 1984 and 1986 he was a lecturer at the Université de Rennes1, France. He then served as engineer, from 1987 to 2000, at Office de l'eau potable et de l'électricité in Morocco. Since 2001 has been a Ph.D. student at the École de Technologie Supérieure de l'Université du Québec in Montreal, Québec, Canada. His research interests include pattern recognition, information retrieval, and image analysis.

Mohamed Cheriet received his B.Eng. in computer science from Université des Sciences et de Technologie d'Alger (Bab Ezouar, Algiers) in 1984 and his M.Sc. and Ph.D., also in computer science, from the University of Pierre et Marie Curie (Paris VI) in 1985 and 1988, respectively. Dr. Cheriet was appointed assistant professor in 1992, associate professor in 1995, and full professor in 1998 in the Department of Automation Engineering, École de Technologie Supérieure of the University of Québec, Montreal. Currently he is the director of LIVIA, the Laboratory for Imagery, Vision and Artificial Intelligence at ETS, and an active member of CENPARMI, the Centre for Pattern Recognition and Machine Intelligence. Professor Cheriet's research focuses on mathematical modeling for signal and image processing (scale-space, PDEs, and variational methods), pattern recognition, character recognition, text processing, document analysis and recognition, and perception. He has published more than 100 technical papers in these fields.

He was the co-chair of the 11th and the 13th Vision Interface Conferences held respectively in Vancouver in 1998 and in Montreal in 2000. He was also the general co-chair of the 8th International Workshop on Frontiers on Handwriting Recognition held in Niagara-on-the-Lake in 2002. He has served as associate editor of the International Journal of Pattern Recognition and Artificial Intelligence (IJPRAI) since 2000. Dr. Cheriet is a senior member of IEEE.

Jian Yun Nie is a professor in the computer science department (DIRO), Université de Montreal, Québec, Canada. His research focuses on problems related to information retrieval, including multilingual and multimedia information retrieval, as well as natural language processing.

Ching Y. Suen received his M.Sc. (Eng.) from the University of Hong Kong and Ph.D. from the University of British Columbia, Canada. In 1972 he joined the Department of Computer Science of Concordia University, where he became professor in 1979 and served as chairman from 1980 to 1984 and as associate dean for research of the Faculty of Engineering and Computer Science from 1993 to 1997. He has guided/hosted 65 visiting scientists and professors and supervised 60 doctoral and master's graduates. Currently he holds the distinguished Concordia Research Chair in Artificial Intelligence and Pattern Recognition and is the Director of CENPARMI, the Centre for Pattern Recognition and Machine Intelligence.

Professor Suen is the author/editor of 11 books and more than 400 papers on subjects ranging from computer vision and handwriting recognition to expert systems and computational linguistics. A Google search on "Ching Y. Suen" will show some of his publications. He is the founder of the International Journal of Computer Processing of Oriental Languages and served as its first editor-in-chief for 10 years. Presently he is an associate editor of several journals related to pattern recognition.

A fellow of the IEEE, IAPR, and the Academy of Sciences of the Royal Society of Canada, he has served several professional societies as president, vice-president, or governor. He is also the founder and chair of several conference series including ICDAR, IWFHR, and VI. He has been the general chair of numerous international conferences, including the International Conference on Computer Processing of Chinese and Oriental Languages in August 1988 held in Toronto, International Conference on Document Analysis and Recognition held in Montreal in August 1995, and the International Conference on Pattern Recognition held in Québec City in August 2002.

Dr. Suen has given 150 seminars at major computer companies and various government and academic institutions around the world. He has been the principal investigator of 25 industrial/government research contracts and is a grant holder and recipient of prestigious awards, including the ITAC/NSERC award from the Information Technology Association of Canada and the Natural Sciences and Engineering Research Council of Canada in 1992 and the Concordia "Research Fellow" award in 1998.