Combining Approaches to On-line Handwriting Information Retrieval

Sebastián Peña Saldarriaga[a], Christian Viard-Gaudin[b] and Emmanuel Morin[a]

[a]LINA UMR CNRS 6241, Université de Nantes, France;
[b]IRCCyN UMR CNRS 6597, École Polytechnique de l'Université de Nantes, France

ABSTRACT

In this work, we propose to combine two quite different approaches for retrieving handwritten documents. Our hypothesis is that different retrieval algorithms should retrieve different sets of documents for the same query. Therefore, significant improvements in retrieval performances can be expected. The first approach is based on information retrieval techniques carried out on the noisy texts obtained through handwriting recognition, while the second approach is recognition-free using a word spotting algorithm. Results shows that for texts having a word error rate (WER) lower than 23%, the performances obtained with the combined system are close to the performances obtained on clean digital texts. In addition, for poorly recognized texts (WER $>$ 52%), an improvement of nearly 17% can be observed with respect to the best available baseline method.

Keywords: On-line handwriting, word-spotting, noisy IR, rank fusion, retrieval models

1. INTRODUCTION

For several years, on-line handwriting was confined to the role of a convenient input method. With the recent evolutions of pen computers and digital pens, the production of on-line documents has become commonplace. As a result, algorithms for efficient storing and retrieval of on-line data are being increasingly demanded.

While the task of efficient retrieval of text documents has been addressed by researchers for many years,[1] the retrieval of handwritten documents - including off-line handwriting as well - has been investigated only recently.[2–7] Roughly, the existing methods for handwritten document retrieval can be divided into recognition-based[2,6] and recognition-free or word-spotting approaches[3,4,7,8](see Figure 1).

The aim of word-spotting approaches is to detect the words belonging to a query in the documents of a database. Whatever the type of the query is (electronic or handwritten text) the performances of recognition-free approaches substantially rely on the proper selection of image features and similarity measures. Thus retrieval errors are expected for words having similar shapes,[3] which results in the retrieval of non-relevant items. On the other side, standard information retrieval (IR) methods applied to noisy texts are most likely to be negatively influenced by high word recognition error rates. This may result in non-retrieval of relevant items, as well as the retrieval of non-relevant ones.

In practice, we cannot tell *a priori* which method performs better than the others under all circumstances. Furthermore, in our opinion, word-spotting and noisy IR do not address the same problem: the former can be seen as an extension of string searching algorithms (like the Knuth-Morris-Pratt algorithm) to the handwritten domain, while the latter should be more concerned with satisfying user information needs.

According to the previous observations, we can state that for the same query q, two different retrieval approaches should retrieve different sets of documents (both relevant and non-relevant). Usually some overlap does exist, and it can have a positive impact on the combined performances. Our hypothesis is that by combining the results of word-spotting and noisy IR we can get the best out of both techniques.

Further author information: (Send correspondence to S.P.S.)
S.P.S.: E-mail: sebastian.pena-saldarriaga@univ-nantes.fr
C.V-G.: E-mail: christian.viard-gaudin@univ-nantes.fr
E.M.: E-mail: emmanuel.morin@univ-nantes.fr

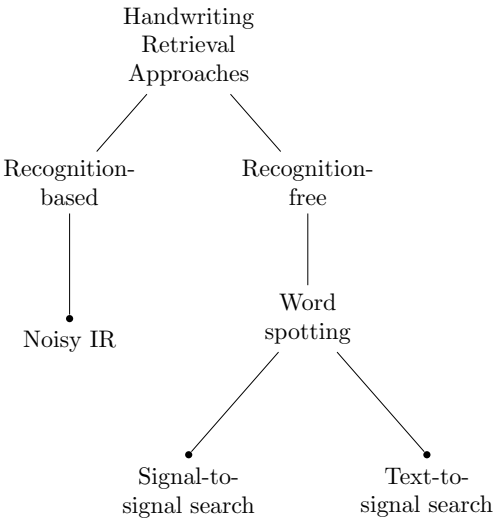

Figure 1. A hierarchical typology of handwritten document retrieval methods.

The combination of two different rankings, in such a way to optimize the performances of the resulting document rank list, is called *ranking fusion*, *ranking aggregation* or *data fusion* in the information retrieval litterature. We only consider the most popular methods in this paper, while a detailed survey about the subject can be found in works by Beitzel et al.,[9] and Farah and Vanderpooten.[10]

In this paper several ranking fusion strategies and their application to IR of on-line handwritten documents are evaluated. The fusion methods used are described in Section 2. These fusion models are applied to results returned by several baseline retrieval algorithms, including both, recognition-free and recognition-based approaches. A corpus of more than 2,000 on-line documents is used for experimental validation. Since this corpus was initially collected for text categorization (TC),[11] it was adapted to fit the needs of our IR experiments (see Section 3).

2. RANK FUSION STRATEGIES

Methods for merging document rank lists use information that is available from the ranking. In most cases, the only information that strategies can exploit are: (i) the ordinal rank of a document; and (ii) the score, or some transformation of the score, assigned to a document.

Let $\tau = [1 \dots |\tau|]$ be a linear ordering of documents given by a retrieval method in response to query q. We then introduce the following definitions.

- $\tau(i)$ is the rank of the document i in τ.
- $s^{\tau}(i)$ is the normalized score of the document i. $\forall \;\; i, j \in \tau$ that $i \neq j$, if $s^{\tau}(i) \geq s^{\tau}(j)$ then $\tau(i) \geq \tau(j)$.
- $\omega(i)$ is the fused score of the document i.
- $R = \left\{\tau_1, \tau_2, \dots \tau_{|R|}\right\}$ is the rank lists set to fuse.
- $h(i, R) = |\{\tau \in R : i \in \tau\}|$ is the number of rankings which had non-zero scores for the document i.

The first fusion method used in our experiments is the sum of scores for a given document called *CombSUM*:[12]

$$\omega(i) = \sum_{j=1}^{|R|} s^{\tau_j}(i) \tag{1}$$

The second fusion method is the sum of scores multiplied by $h(i, R)$, called *CombMNZ* (Multiply Non-Zero). In our case, as two retrieval strategies will be used, $|R| = 2$, and $h(i, R)$ is the number of times (0, 1 or 2) that a document i is retrieved by different approaches.

$$\omega(i) = h(i, R) \times \sum_{j=1}^{|R|} s^{\tau_j}(i) \tag{2}$$

As suggested by Lee,[13] using scores is equivalent to doing an independent weighting, i.e. without considering the whole result list. For this reason, the last two methods used in our experiences are based on ordinal ranks, i.e. considering the whole ranking. First we define a rank-derived score as follows:

$$r^{\tau}(i) = 1 - \frac{\tau(i) - 1}{|\tau|} \tag{3}$$

Then we use $r^{\tau}(i)$ in Equations (1) and (2). For convenience, the resulting methods will be called *RankCombSUM* (Equation (4)) and *RankCombMNZ* (Equation (5)) in our experiments.

$$\omega(i) = \sum_{j=1}^{|R|} r^{\tau_j}(i) \tag{4}$$

$$\omega(i) = h(i, R) \times \sum_{j=1}^{|R|} r^{\tau_j}(i) \tag{5}$$

3. FROM A TEXT CATEGORIZATION TO AN IR COLLECTION

Before describing the experimental setup and reporting the results, it is necessary to describe the Reuters-21578 collection and how it was used. The handwritten collection used in our experiments was previously used in text categorization experiments.[11] An obvious difference between TC and IR test collections is that TC collections do not have a standard set of queries with their corresponding relevant documents. However, documents are labeled with category codes. In our corpus, documents can belong to one out of ten categories.

Categories can serve as a basis for generating queries using relevance feedback. Previous works reported IR experiments with the Reuters-21578 collection[14] using an approach similar to the one that is described here.

In the following sections, we consider our corpus randomly partitioned in two subsets of equal sizes:

- Q is the set used for query generation
- T is the set used for retrieval

Selecting Relevant Terms

In order to provide relevant terms for possible query generation, we used and adaptation of the Porter and Galpin[15] formula. For a given category c, and a term t, the relevance score is given by the difference between the probability of t within the documents of c and the probability of t in Q

$$score_{t,c} = \frac{r}{R} - \frac{n}{|Q|} \tag{6}$$

Where r is the number of documents of c containing term t, R is the number of documents of c in the set Q, and n is the number of documents containing t. We keep the top 100 scoring terms as a basis for query generation as explained below.

Generating Queries

The queries are generated using relevant terms for a given category using the Ide dec-hi[16] relevance feedback formula. The Ide dec-hi method merges vectors of positive and negative documents for a given category c. The prototypical query q_c of c is computed as follows.

$$q_c = \sum_{i=1}^{|c|} C_i - \sum_{j=1}^{|c|} S_j \tag{7}$$

Where C_i is the vector for the i-th document of c, S_j is the vector for the j-th document not belonging to c, and $|c|$ is the number of documents of c. It is worth to note that the vector space dimension is 100, and that documents are indexed using the term frequency. Document term weights are directly substracted without normalization. All the documents of c are used but only $|c|$ random negative samples. The queries generated are 5 terms long as suggested by Sanderson's results.[14]

For each category in our corpus, the relevance feedback query is given in Table 1. By choosing a category, we can now perform a retrieval on T using the generated query. The documents tagged with the choosen category are considered as relevant.

Table 1. Relevance feedback queries for each of the 10 categories represented in the test collection. Query terms are stemmed.

Category	**Query terms**
Earnings	vs ct net shr loss
Acquisitions	acquir stake acquisit complet merger
Grain	tonn wheat grain corn agricultur
Foreign Exchange	stg monei dollar band bill
Crude	oil crude barrel post well
Interest	rate prime lend citibank percentag
Trade	surplu deficit narrow trade tariff
Shipping	port strike vessel hr worker
Sugar	sugar raw beet cargo kain
Coffee	coffe bag ico registr ibc

Event though the generated queries are convincing from a lexical point of view, i.e. they are likely to be representative of their corresponding category, it is not clear if they make sense from a human perspective. Nevertheless, within the scope of our experiments, what is important is that all the retrieval methods perform on the same task. Moreover, the query generation process can be seen as an iteration of relevance feedback during an interactive retrieval session.[14]

4. EXPERIMENTS

4.1 Baseline Scores

Several existing retrieval systems were used as baseline systems to be combined. On the recognition-free side, we use **InkSearch®** (IS)*, which is a stable and out-of-the-box system. It allows searching of text in handwriting, and does not need training.

*InkSearch® is part of MyScript Builder SDK.

On the recognition-based side, two classical models, namely, **cosine** and **BM25**[†] are used. These two methods are considered as among the most effective ones and are regularly used as reference systems. In the cosine model, for each query-document pair (q, d), the Retrieval Status Value (rsv) is given by the cosine of the angle between the query and document vectors. In our experiments documents are indexed using tf and $tf \times idf$ weightings.

$$rsv\,(q,d) = \frac{q \cdot d}{\|q\| \cdot \|d\|} \tag{8}$$

In the probabilistic BM25 model,[17] the rsv is calculated as follows:

$$rsv\,(q,d) = \sum_{i=1}^{|q|} idf(i) \times \frac{tf(i,d) \times (k_1+1)}{tf(i,d) + k_1 \times (1 - b + b \times \frac{|d|}{avgdl})} \tag{9}$$

Where $|q|$ is the number of query terms; $tf(i,d)$ is the frequency of term i in d; $|d|$ is the length of d in terms; $avgdl$ is the average document length in the collection; k_1, and b are free parameters allowing to control the effect of term frequencies and document lengths respectively. We use these parameters as they are usually set in the litterature: $k_1 = 2$ and $b = 0.75$.

For the recognition-based methods, the recognition engine of MyScript Builder[‡] is used. Recognition is performed on a character-level basis (termed as *free*) or on a word-level basis (termed as *text*). For information, the word and term error rates[18] for each recognition strategy are reported in Table 2.

Table 2. Recognition error rates for the *free* and *text* strategies.

Recognition Type	**Word Error Rate**	**Term Error Rate**
text	22.19%	22.45%
free	52.47%	55.80%

Unsurprisingly word-level recognition clearly outperforms character-level recognition. Half the information is lost with the latter, while the former misses about one word out of five. The WER and TER obtained with the *text* strategy can be considered as low since this resource contains no prior linguistic knowledge specific to this kind of documents, which yet feature a lot of out of lexicon terms.

4.2 Results and Discussion

In this section, results on combining IS and the noisy IR strategies described above are presented. The scores reported correspond to the Mean Average Precision (MAP). For a given query, the average precision is defined as follows:

$$avgPrec(q) = \frac{\sum_{r=1}^{n} prec(r) \times rel(r)}{N} \tag{10}$$

Where n is the number of documents retrieved, $prec(r)$ is the precision at position r, $rel(r)$ is the indicator function of the relevance of the r-th document, and N is the number of relevant documents in the entire collection. The individual measures are averaged on a query basis, thus leading to the single MAP measure.

[†]Also known as the Okapi formula

[‡]MyScript Builder SDK can be found at `http://www.visionobjects.com/products/software-development-kits/myscript-builder`

Combining IS and Cosine with tf weighting

In the first experience, we have applied the combining methods for cosine with tf weighting and IS. Table 3 shows the effectiveness levels obtained. As shown in this table the CombMNZ method performs slightly better than CombSUM. Score-based methods clearly outperform rank-based strategies, this is particularly obvious for the runs with the *text*-recognized documents.

Table 3. Improvement in tf/IS scores after fusion. Bold numbers indicate improvements with respect to the best of the baseline methods and italic numbers indicate degradations in performances

Retrieval Method	**MAP**
IS	0.4918
Upper bound[§]	0.7973
tf-free	0.5564
CombMNZ-free	**0.7228**
CombSUM-free	**0.7202**
RankCombMNZ-free	**0.6854**
RankCombSUM-free	**0.6756**

(a) free

Retrieval Method	**MAP**
IS	0.4918
Upper bound	0.7973
tf-text	0.7455
CombMNZ-text	**0.7834**
CombSUM-text	**0.7777**
RankCombMNZ-text	*0.7389*
RankCombSUM-text	*0.7355*

(a) text

Combining IS and Cosine with $tf \times idf$

In the experiments, whose results are reported in Table 4, we also considered the cosine affinity, but using the $tf \times idf$ weighting this time. The same behaviour is observed, except that the upper bound is a little lower than for tf alone. This is due to the query generation process (see Section 3).

Table 4. Improvement in $tf \times idf$/IS scores after fusion. Bold numbers indicate improvements with respect to one of the baseline methods and italic numbers indicate degradations in performances

Retrieval Method	**MAP**
IS	0.4918
Upper bound	0.7714
$tf \times idf$-free	0.5579
CombMNZ-free	**0.7213**
CombSUM-free	**0.7191**
RankCombMNZ-free	**0.6807**
RankCombSUM-free	**0.6715**

(a) free

Retrieval Method	**MAP**
IS	0.4918
Upper bound	0.7714
$tf \times idf$-text	0.7172
CombMNZ-text	**0.7568**
CombSUM-text	**0.7508**
RankCombMNZ-text	*0.7039*
RankCombSUM-text	*0.7011*

(a) text

Combining IS and BM25

We have also applied the combining methods for the probabilistic BM25 model. Serious improvements are also observed for every score-based configuration as shown in Table 5. It is worth to note, that for *text*-recognized documents, the performances after combination are very close to the upper bound.

[§]The upper bound is the MAP obtained with the original electronic texts

Table 5. Improvement in BM25/IS scores after fusion. Bold numbers indicate improvements with respect to one of the baseline methods and italic numbers indicate degradations in performances

Retrieval Method	MAP
IS	0.4918
Upper bound	0.7799
BM25-free	0.5552
CombMNZ-free	**0.7221**
CombSUM-free	**0.7214**
RankCombMNZ-free	**0.6826**
RankCombSUM-free	**0.6734**

(a) free

Retrieval Method	MAP
IS	0.4918
Upper bound	0.7799
BM25-text	0.7283
CombMNZ-text	**0.7689**
CombSUM-text	**0.7659**
RankCombMNZ-text	*0.7227*
RankCombSUM-text	*0.7214*

(a) text

5. CONCLUSION

In our contribution, we presented a variety of strategies to combine approaches to on-line handwriting information retrieval. The analysis of experimental results showed that the CombMNZ function, using a weighted sum of scores given by the word spotting module and the IR system, provided better retrieval performance than the others. However, the performances obtained with CombSUM are still very close to CombMNZ. We also observed that score-based fusion clearly outperform rank-based strategies. This is to the best of our knowledge, the first time that rank fusion methods are explored at the crossroads of IR and pattern recognition, and it opens a wide spectrum of perspectives.

For the time being, the combining methods need to be validated against human prepared queries. The queries should be prepared based on the contents of the database and the relevance of the documents judged by human assessors. Experiments should be conducted using our database and other available corpora as well.

Even though our methods were applied to on-line handwriting, in the middle term, we hope that studies are conducted using different types of document images. Document images can include, but are not limited to, the following types: (*i*) off-line handwritten documents, including historical manuscripts for which achieving acceptable recognition rates is still challenging; (*ii*) complex documents including hand drawn graphics, where graphic and text searches could be combined; and even (*iii*) low quality typeset document images on which OCR performances are very poor.

ACKNOWLEDGMENTS

This work was supported in part by the French National Research Agency grant ANR-06-TLOG-009. Additional support was provided by *La Région Pays de la Loire* under the MILES Project.

REFERENCES

[1] Baeza-Yates, R. and Ribeiro-Neto, B., [*Modern Information Retrieval*], Addison-Wesley (1999).

[2] Russell, G., Perrone, M., and Chee, Y. M., "Handwritten document retrieval," in [*Proceedings of the 8th International Workshop on Frontiers in Handwriting Recognition*], 233–238 (2002).

[3] Jain, A. K. and Namboodiri, A. M., "Indexing and retrieval of on-line handwritten documents," in [*Proceedings of the 10th International Conference on Document Analysis & Recognition (ICDAR 2003)*], 655–659 (2003).

[4] Rath, T. M., Manmatha, R., and Lavrenko, V., "A search engine for historical manuscript images," in [*Proceedings of the 27th Annual ACM SIGIR Conference (SIGIR 2004)*], 369–376 (2004).

[5] Srihari, S. N., Huang, C., and Srinivasan, H., "A search engine for handwritten documents," in [*Proceedings of Document Recognition & Retrieval XII (DRR 2005)*], 66–75 (2005).

[6] Vinciarelli, A., "Application of information retrieval techniques to single writer documents," *Pattern Recognition Letters* **26**(14), 2262–2271 (2005).

[7] Jawahar, C. V., Balasubramanian, A., Meshesha, M., and Namboodiri, A. M., "Retrieval of online handwriting by synthesis and matching," *Pattern Recognition* **42**(7), 1445–1457 (2009).

[8] Rath, T. M. and Manmatha, R., "Word image matching using dynamic time warping," in [*Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition (CVPR 2003)*], 521–527 (2003).

[9] Beitzel, S. M., Jensen, E. C., Chowdury, A., Grossman, D., Frieder, O., and Goharian, N., "On fusion of effective retrieval strategies in the same information retrieval system," *Journal of the American Society of Information Science & Technology* **50**(10), 859–868 (2004).

[10] Farah, M. and Vanderpooten, D., "An outranking approach for rank aggregation in information retrieval," in [*Proceedings of the 30th Annual ACM SIGIR Conference (SIGIR 2007)*], 591–598 (2007).

[11] Peña Saldarriaga, S., Viard-Gaudin, C., and Morin, E., "On-line handwritten text categorization," in [*Proceedings of Document Recognition & Retrieval XVI (DRR 2009)*], 724709 (2009).

[12] Shaw, J. A. and Fox, E. A., "Combination of multiple searches," in [*Proceedings of the 2nd Text REtrieval Conference (TREC-2)*], 243–252 (1994).

[13] Lee, J. H., "Analysis of multiple evidence combination," in [*Proceedings of the 20th Annual ACM SIGIR Conference (SIGIR 1997)*], 267–276 (1997).

[14] Sanderson, M., "Word sense disambiguation and information retrieval," in [*Proceedings of the 17th Annual ACM SIGIR Conference (SIGIR 1994)*], 142–151 (1994).

[15] Porter, M. and Galpin, V., "Relevance feedback in a public access catalogue for a research library: Muscat at the scott polar research institute," *Program* **22**(1), 1–20 (1988).

[16] Ide, E., [*The Smart Retrieval System*], ch. New Experiments in Relevance Feedback, 337–354, Prentice-Hall, Inc. (1971).

[17] Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M., and Gatford, M., "Okapi at trec-3," in [*Proceedings of the 3rd Text REtrieval Conference (TREC 1994)*], 109–126 (1994).

[18] Vinciarelli, A., "Noisy text categorization," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **27**(12), 1882–1895 (2005).