2009 10th International Conference on Document Analysis and Recognition

Results of the RIMES evaluation campaign for handwritten mail processing

Emmanuèle Grosicki[1] emmanuele.grosicki@etca.fr
Matthieu Carré[2] matthieu.carre@it-sudparis.eu
Jean-Marie Brodin[3] brodin@a2ia.com
Edouard Geoffrois[1] e.geoffrois@etca.fr

[1]*DGA/Centre d'Expertise Parisien* [2]*Telecom & Management Sud-Paris* [3]*A2IA*

Abstract

This paper presents the results of the second test phase of the RIMES[1] evaluation campaign. The latter is the first large-scale evaluation campaign intended to all the key players of the handwritten recognition and document analysis communities. It proposes various tasks around recognition and indexing of handwritten letters such as those sent by postal mail or fax by individuals to companies or administrations. In this second evaluation test, automatic systems have been evaluated on three themes: layout analysis, handwriting recognition and writer identification. The databases used are part of the RIMES database of 5605 real mails completely annotated as well as secondary databases of isolated characters and handwritten words (250,000 snippets). The paper reports on protocols and gives the results obtained in the campaign.

Keywords: Evaluation, database, layout analysis, handwriting recognition, writer identification, metric

1. Introduction

Whatever the scientific research field considered, it is necessary to be able to compare objectively the performances of different developed systems. This comparison must be done on the same data set and in the same conditions. The most efficient solution is to organize evaluation campaigns where all systems are compared in the same way, on the same data and at the same time. In the speech recognition field, such evaluation campaigns are regularly organized and have shown the usefulness of such a methodology [1]. In particular, they have contributed to the rapid progress observed for several years by providing a large amount of high quality data difficult and expensive to obtain.

The RIMES project, funded by the French ministries of defense and research wishes to create a similar trend in the document analysis field by evaluating systems dedicated to the recognition and the indexing of mixed (handwritten and typed) documents. It is intended towards all the key players of the document community by proposing numerous various tasks covering layout analysis, handwriting recognition, writer identification, logo identification and information extraction. Two evaluation phases have already been organized using the new annotated databases created in the framework of the RIMES project and composed of more than 12600 pages entirely handwritten (letters) and mixed (handwritten and typed) (fax, form) as well as secondary databases: handwritten isolated characters (100,000) and words (250,000), paragraphs and logos (500).

The first part of the article presents briefly the RIMES annotated databases. The second part of the article describes the results of the recent second phase of the RIMES campaign.

2. The RIMES database

Automatic systems based on statistical methods need a lot of quality training data. The handwriting recognition field suffers from a definite lack of annotated data as their production is an important investment. The first RIMES challenge was then to create a new database. To obtain varied data representative of an industrial application, it was chosen to collect mails such as those sent by individuals to companies by fax or postal mail. Due to legal and confidentiality reasons, it was not possible to collect existing mails. Therefore, we have asked volunteers to write them in exchange of gift vouchers. Each volunteer writer received a fictional identity and up to 5 scenarios, one at a time, among 9

[1] *Reconnaissance et Indexation de données Manuscrites et de fac similÉS / Recognition and Indexing of handwritten documents and faxes*

978-0-7695-3725-2/09 $25.00 © 2009 IEEE
DOI 10.1109/ICDAR.2009.224

realistic themes like damage declaration or modification of contract. Each scenario was combined with various receivers (administrations or service providers). The volunteer composed his letter with those pieces of information using his own words. The layout was free and it was only asked to use white paper and to write in a readable way with black ink. 12,723 pages written by 1,300 volunteers have been collected corresponding to 5605 mails of two to three pages (see fig 1):

- A handwritten letter

-A fixed form with information about the letter

-An optional fax cover sheet.

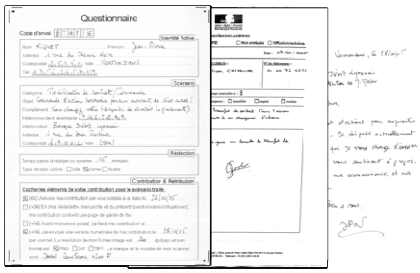

Figure 1 *:* A mail sample

The obtained database was then scanned by a professional quality scanner (300 dpi, gray-level lossless compression) and carefully annotated in order to extract the maximum information which could be useful for evaluation such as its layout structure (logic structure), its textual content (transcription), and more general information about the document like the writer, the date or the object of the letter (information extraction). The ground-truth thus obtained serves as the reference to the RIMES campaign. In order to also propose easier recognition tasks with respect to those using the entire pages of the RIMES mails, isolated logos (500), alphanumeric characters (100,000), handwritten words snippets (250,000) and paragraphs have been extracted from letters, forms and fax.

Some samples are given bellow:

All these databases will be available to the scientific community after the last official test is finished, at the lowest possible cost.

3.The second RIMES evaluation phase

The second RIMES evaluation test phase covers the following themes: document layout analysis, handwriting recognition and writer identification by proposing 7 tasks:

-Letter Layout Analysis (**task LLa**)

-Character Recognition: digit (**task CR1**), alphabetic letter (**task CR2**), alphanumeric character (**task CR3**).

-Word recognition with a given dictionary of different sizes : 100 words for the task **WR_100** and 1636 words corresponding to the test dictionary size for the task **WR_1636**.

-Writer identification from Blocks of words (**WiB**)

For each task, a principal metric is chosen in order to compare hypothesis output from evaluated systems to ground-truth. Secondary metrics called "diagnose metrics" are sometimes also considered to complete the performance analysis.

For each task, a training database set is distributed to the participants as well as a validation set necessary for the tuning of systems and whose features are close to the test set (see table 1).

Tasks	Training set	Validation set	Test set
LLa	1050 letters	100 letters	100 letters
RC1	7671 chars (650 forms)	5937 chars (500 forms)	5124 ch (500 f.)
RC2	2779 ch.	2098 ch	3096 ch
RC3	10450 ch	8035 ch	8220 ch
WR	36588 words	7786 words	7542 wds
WiB	382 letters	100 blocks	100 blocks

Table 1: Data sets for the second RIMES test

3.1. Document layout analysis

The task proposed on this theme is called Letter Layout analysis (LLa) and consists in localizing key-fields in pages of letters useful for their automatic

processing. 8 fields are thus defined: sender, destination, date & place, subject, opening ("Dear Sir"), body of the letter, signature, PS/enclosed (see fig 2).

In the ground-truth, each field is delimited by one or several rectangular zones parallel to the page axis. Most of the time, a single box is sufficient. If writing is slanted, several boxes may be necessary to cover a field without overlapping its neighbours.

For the metric, each pixel receives the label of its covering box or the default label "back-ground" if it is not covered by any box. The metric compares label of each pixel in both hypothesis and ground-truth image, and counts the bi-level weighted pixel classification error rate [2] (see fig.2). As a consequence, white pixels are not taken into account in the error rate.

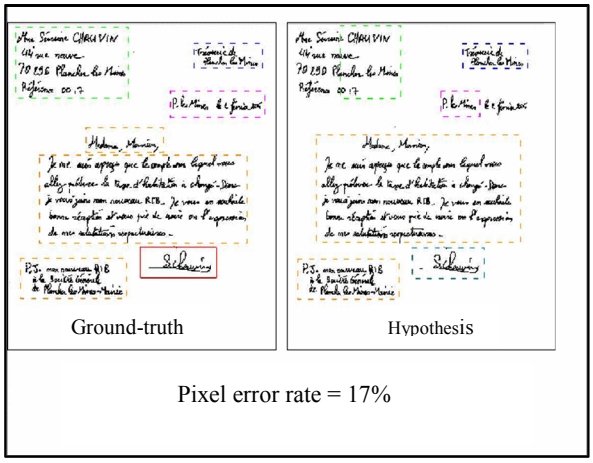

Figure 2: LLa metric

Four systems have been evaluated on this task:

-The first system proposed by CEP Arcueil / Telecom SudParis combines a MRF-based (Markovian Random Field) model with the optimal 2D Dynamic programming. A rule-based post-processing is then used to correct possible errors [4].

-The second system proposed by IRISA Rennes is based on the DMOS method and uses a grammatical EPF language to define some rules supposed to be at the heart of the letter layout [3].

-The third system proposed by LITIS is a MRF-based approach using multi-resolution pixel density and position features. The inference is done by the ICM sub-optimal relaxation-based method [5].

-The fourth tested system proposed by CEP Arcueil/LITIS system is a CRF-based (Conditional Random Field) approach using multi-level features [6].

The results obtained for this task are given below:

Lab	CEP (1)	IRISA (2)	LITIS (3)	CEP/LITIS (4)
error rate %	**8.53**	**8.97**	**12.62**	**12.88**

One can see that the methods based on rules (systems (1) with its post-processing and (2)) give better performance than those entirely statistic (systems (3) and (4)). However, the analysis of the obtained errors on each image shows that methods based on rules generate sometimes very high error rates with respect to statistical methods showing the rigidity of this kind of approach, very specific for a given kind of document layout. Moreover, since the analysis of the obtained errors have not shown any correlation between these different approaches, one can imagine increasing the performance by combining systems.

3.2. Handwriting recognition

This theme includes two different tasks: handwritten character and word recognition. Concerning character recognition from snippets extracted from forms, three subtasks have been proposed: digit recognition (task CR1), alphabetic letter recognition (task CR2), and alphanumeric character recognition (task CR3). As far as recognition of snippets of handwritten words with a given dictionary is concerned, two versions of the task have been proposed corresponding to two different sizes of the given dictionary. In the first case (task WR_100), each word of the test was associated to a list of 99 words chosen randomly among test words in addition to the right transcription. In the second sub-task (WR_1636), the test dictionary was composed of 1636 words.

The ground-truth of snippets of words is faithful to what is written including spelling errors. The primary metric is the recognition error rate. As most of word recognition tools return not a single answer but a list of words with a confidence rate, a secondary metric measures the presence of the correct answer within the N first elements of the recognition list (N equal to 10).

3.2. 1 Isolated Character recognition

For the second RIMES test, three methods have been evaluated (see table 2).

Lab.	Proposed methods
LITIS	Random Forest classification [8]
IRISA	SVM classifier using a vector of 93 features (Zernike moments of order up to 8 and a chaincode feature set extracted from the contours of the connected components)
Itesoft / LORIA	Expert voting combination of MLP (Multilayer Perceptron) classifiers using structural morphological and global features

Table 2: Participants to the 2nd RIMES test

The error rates obtained for the first RIMES evaluation phase were quite high compared to those published for example on the MNIST database. For example, for the task CR1, the error rate was about 2.3% [7].

This could be explained by the size of the training database available for this first test. As a consequence, for the second test, participants could use other training databases in addition to the given training RIMES database.

The error rates in percentage obtained by using either the RIMES database or other training database (like IRONOFF or MNIST) are given below:

Tasks	**RC1**		**RC2**		**RC3**	
Training set	RIMES	other	RIMES	other	RIMES	other
ITESOFT	**0.51**	**0.33**	2.87	**0.61**	4.72	4.59
IRISA	0.76	0.37	**1.91**	1.16	**3.59**	**3.26**
LITIS	0.94		3.68		6.97	

Performance obtained for the task RC1 in this second test phase are closer to those published on MNIST database. However, one can observe that the best results for the three sub-tasks are obtained when participants have used another training database in addition to the RIMES database. This can mean that the latter is still too small to be representative of the test set and should be increased.

3.2.2 Word recognition

Two laboratories have participated to the two sub-tasks proposed on this theme:

-LITIS whose word recognition system is based on a multi-stream segmentation free HMM. Two feature vector sequences are created using a sliding window, and are simultaneously decoded according to the multi-stream formalism. One stream is composed of density features while the other is made of contour features[9].

-ITESOFT/LORIA whose word recognition system is based on the Non-Symmetric Half-Plane Hidden Markov Model (NSHP-HMM). The latter combines a MRF model for each pixel column and a HMM to synthesize the MRF information along the whole image [10].

The recognition rates obtained are given below:

Tasks	**WR_100**		**WR_1636**	
Labs	**Top 1**	**Top 10**	**Top 1**	**Top 10**
LITIS	**91.33%**	**99.12%**	**72.53%**	**93.79%**
Itesoft	84.66%	94.33%	63.99%	86.86%

The results given for Itesoft system have been obtained by normalizing the ground-truth where accents were removed as the proposed system does not return any accent. Moreover, this system had not been trained on the RIMES training set which can explain the difference in performance with the other system.
For the next evaluation tests, it would be interesting to increase the size of the given dictionary to analyze its influence on performance. We could also propose a new task corresponding to recognition of handwritten paragraphs.

3.3 Writer identification

In the first RIMES test, the proposed task has consisted in identifying writers of blocks of words with reject, i.e. the writer might be absent from the training database.

The metric used is the recognition error rate. The reject rate is counted like any other class, and encloses all writers outside the training database.

The test set was composed of 100 blocks of words whose 28 were written by unknown writer. The proposed system based on local features and Bayesian classification [12] had not considered the rejection case. The error rates obtained are given below on top 1 and top 10

	Top 1	Top 10
Test 1	68%	49%

In the second test phase, we have simplified the task by considering only blocks of words whose writer belongs to the training database composed of 382 writers. One system has been evaluated which compares the test images to the reference images by using a mutual information criterion and segmented graphemes computed by a clustering method [11]. The error rates obtained are given below on top 1 and top 10.

	Top 1	Top 10
Test 2	27%	16%

A way to compare the results of these two tests would consist in removing in the first test the errors corresponding to unknown writers. The error rates would be then for the first test

	Top 1(bis)	Top 10 (bis)
Test 1	55%	29%

However, to compare objectively the performances of these two systems, experiments should have been done on the same data set and in the same conditions which is not the case here.

For the next evaluation, we wish to propose again the task “writer identification with reject”.

4. Conclusion

The paper presents the results of the second RIMES test phase where 7 tasks have been proposed covering handwriting recognition, writer identification and layout analysis. 5 French laboratories (LITIS, CEP Arcueil, Telecom Sudparis, LORIA and IRISA Rennes) and 1 company (Itesoft) have participated. First benchmark results have been obtained on specific tasks such as recognition of isolated handwritten words. New tests with more tasks are scheduled in 2009. They are open to all key-players of the handwritten recognition and document analysis community. More information about the RIMES campaign are available on the web site http://rimes.it-sudparis.eu.

5. References

[1] D. Pallett, “A Look at NIST’s Benchmark ASR Tests: Past, Present, and Future”, Proc 2003 IEEE Workshop on Automatic Speech Recognition and Understanding.

[2] B.A. Yanikoglu et L. Vincent, "Pink panther : a complete environment for ground-truthing and benchmarking document page segmentation", Pattern Recognition, Vol. 31, N°9, 1998.

[3] A. Lemaitre, J. Camillerapp, B. Coüasnon, “A generic method for structure recognition of handwritten mail document”, Document Recognition and Retrieval 2008.

[4] M.Lemaitre, « Approche markovienne bidimensionnelle d’analyse et de reconnaissance de documents manuscrits », Phd thesis analysis of University of Paris V, Nov. 2007

[5] S. Nicolas, T. Paquet, L. Heutte, "A Markovian Approach for Handwritten Document Segmentation", International Conference on Pattern Recognition, *2006.*

[6] F. Montreuil, E. Grosicki, L. Heutte, S. Nicolas, “Unconstrained Handwritten Document Layout Extraction using 2D Conditional Random Fields”, *to be published ICDAR 2009*

[7] E. Grosicki, M. Carré, J-M. Brodin, E. Geoffrois “ RIMES Evaluation campaign for handwritten mail processing", International Conference on Frontiers in Handwritting Recognition *ICFHR 2008*

[8] S. Bernard, L. Heutte, S. Adam "Using Random Forests for Handwritten Digit Recognition", *ICDAR 2007*

[9] Y. Kessentini, T. Paquet, A. Benhamadou, “A Multi-Stream HMM-based Approach for Off-line Multi-Script Handwritten Word Recognition”, *ICFHR 2008*

[10] C. Choisy, “Dynamic Handwritten keyword spotting based on the NSHP-HMM”, *ICDAR* 2007.

[11] A. Benséfia, T. Paquet, L. Heutte, “A Writer Identification and Verification System”, *Pattern Recognition Letters*, vol. 26, no. 13, pp. 2080-2092, 2005.

[12] I. Siddiqi, N. Vincent, “Writer Identification in Handwritten Documents”, *ICDAR 2007*