RIMES evaluation campaign for handwritten mail processing

Emmanuèle Grosicki¹ Matthieu Carré² Jean-Marie Brodin³ Edouard Geoffrois¹

emmanuele.grosicki@etca.fr matthieu.carre@itsudparis.eu Jean-Marie.Brodin@a2ia.com edouard.geoffrois@etca.fr

¹DGA/Centre d'Expertise Parisien ²Telecom sud-Paris/Artemis ³ Société A2IA

Abstract

The paper presents the French RIMES (Reconnaissance et Indexation de données Manuscrites et de fac similÉS / Recognition and Indexing of handwritten documents and faxes) project, which is an evaluation campaign dedicated to automatic systems of recognition and indexing of handwritten letters. Of particular interest are cases such as those sent by postal mail or fax by individuals to companies or administrations. This project is intended to all the key players of the handwritten recognition and document analysis communities. In particular, regular evaluation tests are provided to them (the first test took place in June 2007). In this framework, a new database of more than 5600 real mails completely annotated has been created as well as secondary databases of isolated characters, handwritten words (300,000 snippets) and logos. Beyond the evaluation, the other contribution of the RIMES project is to provide public access for the community to all these databases after the last test.

Keywords: Evaluation, metrics, databases, annotations.

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[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 in the framework of Techno-Vision 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. This project is intended towards all the key players of the document community where very few evaluation campaigns exist already and where there is an obvious lack of quality database publicly available. Among existing databases, one can enumerate the NIST database [2] which contains character and words in fixed format forms, the Senior & Robinson database [3] which corresponds to a single-writer and contains 25 pages, the IAM database [4] which is the largest cursive handwriting database, fixed forms, constrained text and non-native writers. As far as faxes are concerned, no public large database are available ; those existing are on a small scale (e.g. the Majordome project [5]). As far as existing evaluation campaigns in the field are concerned, they often focus on a specific task like recognition characters (NIST) or layout analysis (ICDAR).

The aim of the RIMES project is thus double :
-Firstly, create a new consequent annotated database composed of pages entirely handwritten (letters) and mixed (handwritten and typed) pages (fax, form).
-Secondly, set up the first large-scale evaluation campaigns in the document field proposing numerous varied tasks representative of industrial applications. The themes covered by RIMES are thus layout analysis, handwriting recognition, writer identification, logo identification and information extraction.

The first part of the article presents the database created by RIMES as well as its annotations. The second part of the article deals with the different tasks proposed in the framework of RIMES as well as the associated metrics. The third part focuses on the course of evaluation tests and in particular on the first test which were held from the 4th June to the 18th June 2007 and where 5 laboratories have participated on 6 tasks covering layout analysis, character recognition, writer identification and logo identification.

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 a real application, it was chosen to collect mails such as those sent by individuals to companies or administrations by fax or postal mail. Unfortunately, for legal and confidentiality reasons, it was not possible to collect existing mails. Despite this fact, a solution to create such a database has consisted in asking volunteers to write them in exchange of gift vouchers. Following the telephone speech campaign that drew volunteers to call free number recording their conversations, a website has been thus created dedicated to the database gathering (www.scribee.org) where volunteer writers registered on by giving their real identity (for gift vouchers) and miscellaneous information for database statistics (age, sex, right/left writer, ...). They were given a fictional identity (same sex) and up to 5 scenarios, one at a time. Each scenario has been chosen among 9 realistic following themes : change of personal information (address, bank account), information request, opening and closing (customer account), modification of contract or order, complaint (bad service quality…), payment difficulties (asking for a delay, tax exemption…), reminder letter, damage declaration with further circumstances and a destination (administrations or service providers (telephone, power, bank, insurances). The volunteer composed a 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.

The collect was a success with more than 1,300 people who have participated to the RIMES database creation by writing up to 5 mails.

The RIMES database thus obtained contains 12,723 pages corresponding to 5605 mails of two to three pages (cf fig 1):

- A first page corresponding to a handwritten letter
- A second page corresponding to a fixed form with information about the letter like the name of the sender, the receiver organism…
- An optional third corresponding to a fax cover sheet.

The obtained database is thus composed of more than 12,000 pages which was scanned and carefully annotated as the quality of the RIMES evaluation depends strongly on the quality of the annotation.

First, it was decided to scan each page of the database in different ways by a professional quality scanner (300 dpi, gray-level lossless compression) and by personal scanners in order to have at one’s disposal several digital versions of the same document. This will allow us to study the stability of the different algorithms evaluated in the framework of RIMES with respect to the quality of the digitalization. Then, each page has been then precisely 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 name of the company sender, the date of the letter (Information extraction). The ground-truth thus obtained is stored in XML files and serves as the reference to the RIMES evaluation. Fig. 2 shows the tool developed by RIMES's organizers to create and visualize image ground-truth.
Figure 2: Annotation of a letter

The important work made on the annotation of the RIMES database allows us to consider all kinds of tasks for evaluation covering layout analysis, handwritten recognition, writer and logo identifications and information extraction. In order to also propose easier recognition tasks with respect to those using the entire pages of the RIMES mails, isolated logos, alphanumeric characters and handwritten words snippets have been extracted from letters, forms and fax.

The secondary databases thus obtained are composed of:
- 100,000 snippets of Alphanumeric characters extracted from questionnaires
- 500 snippets of logos extracted from fax (representing 60 classes)
- 300,000 snippets handwritten word extracted from letters

Some samples are given below:

Each snippet and its corresponding ground-truth are generated automatically but are controlled manually in order to ensure a good quality of the databases. The handwritten word database is an asset to the RIMES evaluation as of what we know, such a database does not still exist. All these databases will be available to the scientific community after the last official test is finished, at the lowest possible cost.

3. The RIMES Campaign: tasks & metrics

The goal of the RIMES project is to evaluate systems dedicated to the recognition and the indexing of handwritten mails by proposing varied tasks close to industrial applications. The 5 themes covered by RIMES are thus:

- Document layout analysis
- Handwriting recognition
- Writer identification and verification
- Logo identification
- Information extraction.

Several tasks of various difficulties are thus defined in discussion with the evaluation participants for the following data: isolated characters, words, paragraphs, letters, forms, fax cover-sheets, and logos.

For each task, a principal metric is chosen in order to compare hypothesis output from evaluated systems to ground-truth. Its choice is important and difficult as it must reflect the most accurate possible automatic systems capabilities on the task. Secondary metrics called “diagnose metrics” are sometimes also considered to better analyze the obtained errors of automatic systems.


Two tasks are proposed with respect to this theme: one on letters (task LLa: Letter Layout analysis) and the other on faxes (task FLa: Fax Layout analysis). Their purpose is to localize fields in these pages useful for their automatic processing. As far as letters are concerned, 8 fields are thus defined: sender, destination, date & place, subject, opening (Dear Sir), body of the letter, signature, PS/enclosed.

Regarding faxes, the task is limited to only locate handwritten/printed and logos zones. This difference between fax and letter is explained by the fact that faxes have less layout constraints than letters and as a consequence additional processing similar to information extraction task would be necessary in the fax case to achieve the same field identification as the letters.

In the ground-truth, each field is delimited by one or several rectangular zones parallel to page axis. One box is sufficient most of the time. If writing is skewed, several boxes are necessary to cover the different fields and to avoid box overlapping corresponding to different fields.

As far as the metric is concerned, its choice was a subject of considerable discussions between the organizers of the
project and the evaluation participants. The metric should be accurate and objective to reflect system capability, at the same time simple. The chosen metric is based on a box overlapping measure weighting each pixel by its gray-level [6]. Each pixel thus receives the label of its covering box or the label “back-ground” if it is not covered by any box. The metric compares labels of each pixel in both hypothesis and ground-truth image, and counts the gray-level weighted pixel classification error rate. As a consequence, white pixels are not taken into account in the error rate.

Figure 3 presents the metric on a simple letter layout.

**Figure 3**: Application of the LLa metric on a letter

### 3.2 Handwriting recognition.

This theme concerns three kind of documents: letters, forms, fax cover-sheets. Several tasks of variable difficulties are thus proposed covering characters, words, paragraphs recognition to full pages recognition. Ground-truth are faithful to what is written including spelling and grammar errors. They also contain the corrected version of the transcription. The following 5 types of tasks are defined:

- Isolated Character Recognition (task CR) from snippets extracted from forms: digit recognition (task CR1), alphabetic letter recognition (task CR2), and alphanumeric character recognition (task CR3).
- Handwritten Word Recognition from snippets extracted from letter body (task WR).
- Handwritten word recognition applied on - word blocs where reading order presents no ambiguity (task BR: Block Recognition), - forms where all fields have to be recognized (task FR: Form Recognition)

**Pixel error rate = 17%**

### 3.3 Writer recognition

We group here 3 tasks: writer verification and writer identification on isolated word and bloc of words. Writer verification confirms or not that two isolated word snippets of the same word were written by the same person. The system may also access to the whole training database with all possible writers. Writer identification with reject is an open-world question, i.e. the writer may be outside the training database. It applies on isolated word (task WIW) or block of words (task WIB), given the whole training database. This task is designed for forensic systems (not using any semantic information, thus not using sender bloc of letters).

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

### 3.4 Logo recognition

This theme was less studied, because another Techno-Vision project called EPEIRES is more dedicated to these systems. The task (LgI) is an identification with reject: recognize a given snippet given the training database. For the record, logo localization is covered by the first theme: fax cover sheet layout analysis. Like writer identification, we measure the snippet recognition rate, counting the reject class (for all logo outside training database) like a N+1 class.

### 3.5 Information Extraction

The first obvious task is the letter scenario identification, among the 9 classes. The standard techniques used for TREC competition on press news need to be adapted to the very confusing handwriting recognition, which is a challenge of this project. The second information extraction task is the writer identification, given the whole letter or fax with the sender address bloc. The letter recognition task defined before requires the labeling of each zone among 8 classes (destination, sender, date/place …), which is quite related with information extraction. The metrics are simply the recognition rates.
4. The First RIMES evaluation test

A first evaluation test was held in June 2007 in which 5 laboratories have participated: ENST (Paris), LITIS (Rouen), DGA/CEP Arcueil, IRISA (Rennes) and CRIP V (Paris V). 6 tasks have thus been evaluated covering layout analysis, character recognition, writer identification and logo identification. All the tasks planned in the project RIMES could not be considered here as at the moment of this test all the RIMES database was not yet available. This first test allows us to validate most of the RIMES metric such as the task SL as well as the whole evaluation protocol. It also allows us to organize the first workshop gathering all participants at the evaluation where results and used algorithm were analysed and discussed.

Table 1 summarizes the list of tasks covered by the first RIMES evaluation campaign as well as the participation of the laboratories.

<table>
<thead>
<tr>
<th>Task</th>
<th>Participation for each task</th>
</tr>
</thead>
<tbody>
<tr>
<td>Labo</td>
<td>LLa</td>
</tr>
<tr>
<td>ENST</td>
<td>x</td>
</tr>
<tr>
<td>LITIS</td>
<td>x</td>
</tr>
<tr>
<td>CEP</td>
<td>x</td>
</tr>
<tr>
<td>IRISA</td>
<td>x</td>
</tr>
<tr>
<td>CRIP V</td>
<td></td>
</tr>
</tbody>
</table>

For each task, a training database set was 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 2).

<table>
<thead>
<tr>
<th>Test set</th>
<th>Validation set</th>
<th>Test set</th>
</tr>
</thead>
<tbody>
<tr>
<td>LLa</td>
<td>950 letters</td>
<td>100 letters</td>
</tr>
<tr>
<td>CR1</td>
<td>4773 ch. (400 forms)</td>
<td>2881 ch. (250 forms)</td>
</tr>
<tr>
<td>CR2</td>
<td>1597 ch. (400 forms)</td>
<td>1182 ch. (400 forms)</td>
</tr>
<tr>
<td>CR3</td>
<td>6370 ch. (400 forms)</td>
<td>4063 ch. (400 forms)</td>
</tr>
<tr>
<td>WIB</td>
<td>950 letters</td>
<td>100 letters</td>
</tr>
<tr>
<td>Lgl</td>
<td>236 logos</td>
<td>100 logos</td>
</tr>
</tbody>
</table>

The procedure of evaluation which has been chosen among the possible ones [7] consists in sending the test database to each participant, running his own software in his own hardware environment, and sending back his result file in the expected format as soon as processing is finished. Participants commit themselves not to modify their system during the test phase.

All collected results are sent back to all participants with the ground-truth and the metric results of measuring tools. The ground-truth error correction are done in agreement between evaluation organizers and the participants. The results obtained by automatic systems for the 6 tasks are summarized in table 3.

Table 3: Results of the first RIMES test

<table>
<thead>
<tr>
<th>Tasks</th>
<th>Error rate x in %</th>
<th>Some references of the tested methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>LLa</td>
<td>9.13 &lt; x &lt; 11.66</td>
<td>HMM 2D [11], DMOS method (based on grammatical EPF language) [8]</td>
</tr>
<tr>
<td>CR1</td>
<td>2.26 &lt; x &lt; 2.37</td>
<td>HMM 2D [11], SVM classification [12], SVM/Random Forest classification [10]</td>
</tr>
<tr>
<td>CR2</td>
<td>5 &lt; x &lt; 6.39</td>
<td>Local features &amp; Bayesian classification [9]</td>
</tr>
<tr>
<td>CR3</td>
<td>6.97 &lt; x &lt; 7.27</td>
<td>HMM 2D [11], Correlation measure [12]</td>
</tr>
<tr>
<td>WIB</td>
<td>Top 1: 68*</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Top 10: 49*</td>
<td></td>
</tr>
<tr>
<td>Lgl</td>
<td>1 &lt; x &lt; 3</td>
<td></td>
</tr>
</tbody>
</table>

*The proposed system has not taken into account the fact that the writer may be outside the training database.
Results are given here anonymously as it was a first test and the evaluation protocol was not yet validated. Despite that, the obtained results are interesting as they show that the different proposed methods have very close error rates for the considered tasks. The analysis of the obtained errors does not show any correlation between the different systems in particular for the LLa task and as a consequence one can imagine the possibility of combining different systems in order to reduce the error rates. Moreover, a workshop gathering all participants was organized at the end of the test where results and used methods were presented and discussed. Some references of the proposed methods are given in table 3. Beyond the obvious scientific interest, the workshop allows us also to improve some aspects of the evaluation. For example, the analysis of the errors for the task LLa have shown the impact of the pixels of the background assumed to be white on the error rate and some modifications are therefore forecasted for next evaluations. Moreover, the weak error rate obtained for the task LgI have shown that it is more interesting to extend this task to a more complete task including localization in fax cover-sheets.

5. Conclusion

The paper presents the project RIMES whose goal is to evaluate system dedicated to handwriting recognition and indexing documents. The idea is to create a trend similar to that existing in speech recognition field where important progress have been observed since regular evaluation campaigns are organized. The RIMES project has already succeeded in creating a large database of more than 5600 free handwritten mails from multiple writers which will be put at the end of the last test at the scientific community’s disposal in a field where very few annotated databases are for the moment publicly available. The tasks proposed in the framework of the RIMES evaluation are related to document layout analysis, handwriting recognition, writer identification and verification, logo identification and information extraction. A first test was held in June 2007 where 5 laboratories have participated on 6 tasks. This first test allowed us to validate the evaluation protocol and some metrics and to obtain the first benchmarks. This test was also the opportunity to organize a scientific workshop where all participants have presented their different methods. A new test is planned in June 2008 with more complete tasks for which all laboratories and industries are welcome. More information about the RIMES project are available on the web site http://rimes.it-sudparis.eu

References