ICDAR 2011 - French Handwriting Recognition Competition

Emmanuèle Grosicki
DGA Ingnierie des Projets,
7-9 rue des Mathurins,
92220 Bagneux, France
emmanuele.grosicki@dga.defense.gouv.fr

Haikal El Abed
Technische Universitaet Braunschweig,
Institute for Communications Technology (IfN),
Braunschweig, Germany
elabed@tu-bs.de

Abstract

This paper describes the French handwriting recognition competition held at ICDAR 2011. This competition is based on the RIMES-database composed of French written documents corresponding to letters sent by individuals to companies or administrations. Two tasks have been proposed this year : the first one consists in recognizing isolated snippets of words with the help of a given dictionary; the second one consists in recognizing blocks of words segmented into lines. This year 9 systems were submitted for the different competition subtasks. A comparison between different classification and recognition systems show interesting results. A short description of the participating groups, their systems, and the results achieved are presented.

1. Introduction

Following the success of the ICDAR 2009 French handwriting recognition competition, a new French evaluation campaign has been proposed. Its goal is to evaluate automatic systems on two tasks of rising difficulties:

- The first one is similar to the one proposed in ICDAR 2009 and corresponds to recognition of isolated words with a given dictionary. It allows the community to see the performance improvements of automatic systems on this task.
- The second one represents a new task of handwritten word recognition in context. Its goal is to recognize blocks of words corresponding to the body of a mail sent by an individual to a company or an administration. Each block is segmented into lines and automatic systems are given in entry the coordinates of the corresponding line polygons.

.

The organization of an evaluation session where all automatic systems are compared in the same way, on the same data and at the same time appears to be the most efficient solution to be able to compare objectively the performances of different developed systems. Moreover, evaluation campaigns allow participants to have quality training data which are difficult to obtain as their production is an important investment.

In this competition, the high-quality database created in the framework of the RIMES [7] (Reconnaissance et Indexation de données Manuscrites et de fac similES/ recognition and indexing of handwritten documents and faxes) project has been used. It is composed of more than 12000 pages entirely annotated, 100000 snippets of characters, 250000 snippets of words and 6500 blocks of words. More information about the database and the project RIMES can be found in http://rimes.it-sudparis.eu. and http://www.rimes-database.fr.

This paper is organized as follows. In Sections 2 and 3 the database and the test set are presented in some details. Section 4 presents the participating groups and a short description of their systems. Section 5 describes the tests and the results achieved with the different systems. Finally the paper ends with some concluding remarks.

2. The RIMES Database

Automatic classification and recognition 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 RIMES database is composed of 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, the RIMES organizers have asked to 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 realistic themes like damage declaration or modification of con-

1520-5363/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDAR.2011.290

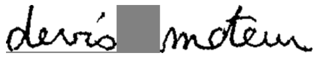
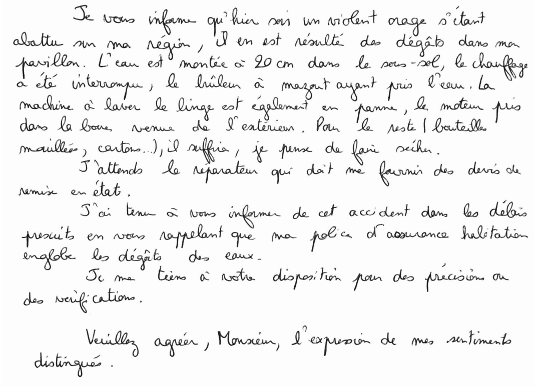


Figure 1. Examples of snippets of words.



Je vous informe qu'hier soir un violent orage s'étant
abattu sur ma région, il en est résulté des dégâts dans mon
pavillon. L'eau est montée à 20 cm dans le sous-sol, le chauffage
a été interrompu, le brûleur à mazout ayant pris l'eau. La
machine à laver le linge est également en panne, le moteur pris
dans la boue venue de l'extérieur. Pour le reste (bouteilles
mouillées, cartons...), il suffira, je pense de faire sécher.
J'attends le réparateur qui doit me fournir des devis de
remise en état.
J'ai tenu à vous informer de cet accident dans les délais
prescrits en vous rappelant que ma police d'assurance habitation
englobe les dégâts des eaux.
Je me tiens à votre disposition pour des précisions ou
des vérifications.

Veuillez agréer, Monsieur, l'expression de mes sentiments
distingués.

Figure 2. Example of block of words.

tract. 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. 12723 pages written by 1300 volunteers have been collected corresponding to 5605 mails of two to three pages corresponding to a handwritten letter, a fixed form with information about the letter and an optional fax cover sheet.

The obtained database was then scanned by a professional quality scanner (300 dpi, gray-level lossless compression). Isolated handwritten words snippets (250000) have been then extracted from handwritten letters. Some samples are shown in figure 1.

Each snippet has been associated to a transcription faithful to what is written including spelling and grammar errors. Each snippet with its transcription (Ground-Truth, GT) have been examined manually in order to insure a good quality of this database.

Blocks of words have also been extracted from letters. An example is shown in figure 2.

3 Recognition tasks:

3.1 Evaluation protocol:

For both tasks, the participants are given a training database and a validation database to train and test their system. At the start of the test period, each participant has access to the unknown test dataset to run his own software on them in his own hardware environment. Participants commit themselves not to modify their system during the test phase. Multiple runs are accepted, but participants must identify one of them as their primary one. The result files in the expected format have to be sent back before the end of the test phase.

3.1.1 Task 1

The training database is composed of more than 50.000 snippets of words, the validation test and the test set respectively about 7.000 snippets of words. At the beginning of the test period, participants were given a dictionary composed of more than 5740 words containing the test words. As far as the metric is concerned, the chosen primary error rate measure consists in counting word error rate. As most of word recognition tools return not a single answer but a list of words with confidence score, a measure of the presence of correct answer in the N-best recognition list (N up to 10) is added.

3.1.2 Task 2

The training database is composed of 1.500 blocks of words, the validation test and the test set respectively about 100 blocks of words. The error measure consists in counting substitutions, deletions and insertions in the alignment between the ground-truth and the hypothesis. Percent of correct words given by (1) is also given. The measure will be done without punctuation and case. We will use the popular tool ScLite from NIST (www.nist.gov/speech), used in speech recognition.

$$\text{Corr} = 100 * \frac{\text{\# Correct words}}{\text{\# Reference words}} \qquad (1)$$

4. Participating systems description

The following section gives a brief description of the systems submitted to the competition. Each system description has been provided by the system's authors and edited (summarized) by the competition organizers. The descriptions vary in length due to the level of detail in the source information provided.

4.1 Telecom ParisTech

4.1.1 Word Recognition system

The proposed system is based on an HMM-based approach. It is composed of three steps : preprocessing, features extraction and recognition. First, grayscale images are deslanted and the background is whitened using Otsu threshold. Then the feature extraction module, based on the sliding-window approach transforms the image into a sequence of 28 features [1] . Recognition system uses 81 different character models, which are transformed into tri-graph models to take into account contextual information. Then those tri-graph models are tied using a decision tree [3]. Models are left-right HMMs using mixtures of Gaussian probability density functions. Training and Recognition are based on the HTK toolkit. Training of models was made on train and validation databases provided for the competition.

4.1.2 Block Recognition system

Telecom ParisTech has proposed three systems(1), (2), (3) all based on an HMM approach applied at the line level. They differ by their language modeling. They are composed of three steps : preprocessing, feature extraction and recognition.

First, lines are cut from grayscale documents. Then, peripheral noise belonging to other lines is removed using a connected components extraction. In addition, line images are deslanted and the background is whitened using Otsu threshold.

Then the feature extraction module, based on the sliding-window approach transforms the image into a sequence of 28 features [1]. Recognition system uses 91 different character models, which are transformed into tri-graph models to take into account contextual information. Then those tri-graph models are tied using a decision tree [3]. Models are left-right HMMs using mixtures of gaussian probability density functions. Training and Recognition are based on the HTK toolkit. Recognition uses also a bigram language model built with SRILM on train transcriptions.

Reference system (1) and second system (2) language models both include backed-off probabilities in order to model bigrams which don't appear on the train data. Those systems only differ by the weight given to computed language model probabilities. Third System (3) does not include backed-off probabilities.

4.2 IRISA

IRISA has proposed two systems based on Continuous Densities HMMs [8]. They are composed of four modules: pre-processing, feature extraction, recognition and verification.

The pre-processing module, suitable for HMMs-based approach and aimed at correcting/normalizing the handwritten styles attributes of the samples, involves the following steps: noise reduction, skew and slant corrections. All these steps were performed using standard state-of-the-art techniques found in the literature.

The sliding-windows-based feature extraction module transforms each preprocessed word image into a sequence of 60-dimensional real-valued feature vectors, normalized by implicitly using the base and upper lines to define the ascender, descender and main body zone where the features are independently extracted.

As mentioned, the recognition process is based on HMMs, where character classes (the basic recognition units) are modeled as a continuous left-to-right HMMs, using a variable number of states for each of them. That is, the number of states for a particular HMM character class is in function of the average length of feature vector sequences used to train it. Moreover, each HMM state was assumed to generate feature vectors following a mixture of Gaussians densities.

Finally each hypothesis of the N-best list produced by the HMM recognition is used to segment the image into corresponding characters and re-score each of them with a trained SVM. The new character scores are added to assign a global SVM score to the hypothesis, to be combined with the HMM score to re-sort the hypotheses and therefore to label the word with an optimal and verified hypothesis according to SVM and HMMs.

All training process was carried out using only the RIMES Database provided by the competition organization.

The two proposed systems differ by the way the dictionary is reordored : for system 1, it is done according to accents and for system 2 randomly.

4.3 A2iA systems

4.3.1 Isolated word recognition system

The system submitted by A2iA is a combination of three different kinds of word recognizers. All these systems were trained independently using only the data provided for the competition. Symbol models were trained for letters with accent and case, digits and punctuation symbols.

Each recognizer outputs a list of 10-best word candidates along with confidence scores, which are combined using a weighted Sum-Rule voting algorithm. Then the case is normalized so as to provide the final answer.

- Hybrid MLP-HMM recognizer based on grapheme extraction:
  This system is based on a hybrid MLP-HMM using features computed on graphemes extraction [9]. Unlike the two other systems, this recognizer relies on a segmentation of the word into a sequence of graphemes. A vector of 74 features (statistical and geometric) are extracted on each grapheme. A Multi-Layer Perceptron embedded in a HMM is trained to compute the posterior probability of each grapheme class with respect to its feature vector.

- GMM-HMM recognizer based on sliding window feature extraction :

This system is based on an HMM modeling of characters using Gaussian Mixtures. A sliding window is used to extract 33 statistic, geometric and directional features which are subject to a first-order regression. We use context-dependent character models (trigraph). In order to reduce the number of parameters needed to learn the trigraph models, a state clustering is performed, where a state can be shared between all trigraphs centered on a same character for a given state position. The clustering is based on original binary decision trees [3]. For an initial number of characters equal to 78, 1942 trigraphs are modeled using 3586 state clusters.

- MDLSTM recurrent Neural Network :
  This system is a Two Dimensional Long-Short Term Memory recurrent neural network [6]. Unlike the two other systems, this recognizer does not rely on HMMs. It also directly takes the image as input (the raw values of pixels), and therefore trains its own embedded feature extraction given the data. Given an input image, the MDLSTM provides a sequence of output activations (80 characters + blank symbol) corresponding to character posterior probabilities (softmax outputs), therefore creating an output lattice [5]. We used our own modified version of Alex Graves' RNNLib library to perform the trainings and decodings.

4.3.2 Block recognition system

The system proposed by A2iA is based on an explicit word segmentation, and uses the combined system trained on isolated words previously described. For each line, 8 options of word segmentation are considered. For each word in each segmentation option, a combination of recognizers is used to provide 15 word recognition candidates. Both the word segmentation options and the word recognition options are then stored in a weighted finite state transducer (WFST) [2]. A language model was trained using SRILM [**?**] and converted into a WSFT. The language model was trained on the competition's line-level training set only. The lexicon for the language models and the recognizer was build from the competition's line-level training set and was composed of 6556 words. The recognition WFST and the language model WFST were composed and the recognition result was extracted using a best-path algorithm.

4.4 Jouve

In order to improve performance of the state of the art HMM recognition engine, JOUVE has developped a complementary system. The proposed system is then a combination of two different kinds of classifiers.

The first is based on a state of the art HMM recognition engine (recognition rate of 76.34% on the second half of the validation set). This classifier has not been trained on the RIMES database. The second used the open-source RNNLIB[4] which is based on a hierarchy of multidimensional recurrent network. This classifier has been trained only on the training set of the RIMES database with no preprocessing or feature extraction. As RNNLIB doesn't implement the decoding with a dictionary, a post-processing is done to output a dictionary word. An edit distance dedicated to handwriting has then been trained on the first half of the validation set to favor frequent substitutions (recognition rate of 77.45% on the second half of the validation set). Lists of words proposed by both systems have then been combined with the Borda Count method. An improvement of more than 10% in absolute is reached (recognition rate of 88.58%) which demonstrates the complementary of both systems.

5. Results

5.1 Task 1 : Isolated Word recognition

5 systems have been evaluated on this task. Table 1 gives the obtained recognition rates.

Table 1. Error rate in % on task Word Recognition

System	top_1	top_{10}
A2IA	5.13	0.44
Jouve	12.53	2.04
IRISA 1	21.41	11.51
ParisTech(1)	24.88	6.85
IRISA 2	25.46	16.08

Figure 5.1 shows error rate versus top number.

5.2 Task 2 : block Word recognition

4 systems have been evaluated on this task. Tables 2 and 3 give respectively the results at word and character levels.

Table 2. Word error rate (task BR)

Systems	**A2IA**	Telecom 1	Telecom 2	Telecom 3
% corr	86.1	73.2	69.6	63.7
%sub	10.0	24.4	27.9	33.8
% del	3.9	2.4	2.5	2.4
% ins	1.3	4.4	6.6	8.2

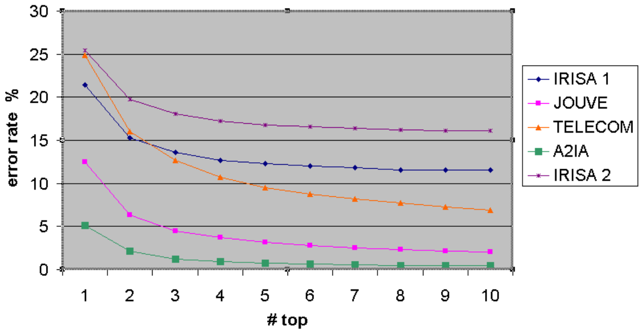


Figure 3. Example of block of words.

Table 3. Character error rate (task BR)

Systems	A2IA	Telecom Paritech 1
% corr	93.7	84.2
%sub	2.9	9.8
% del	3.4	6
% ins	0.9	2.2

6. Conclusions

This competition has allowed us to evaluate 9 systems on 2 tasks of French handwritten recognition. The best results were achieved with a combination of three word recognizers : two based on HMMs and one on a Long-Short Term Memory recurrent neural network (**A2IA**). But others methods based on HMM or classifiers achieved even good recognition rates. Moreover, it is important to notice that some systems have used other training sets than the RIMES one.

For the next evaluation tests, it would be interesting to extend the task 2 to the recognition of complete text document

References

[1] R. M. Al-Hajj, L. Likforman-Sulem, and C. Mokbel. Combining Slanted-Frame Classifiers for Improved HMM-Based Arabic Handwriting Recognition. 31, 2009.

[2] C. Allauzen, M. Riley, J. Schalkwyk, W. Skut, and M. Mohri. OpenFst: A general and efficient weighted finite-state transducer library. In *Ninth International Conference on Implementation and Application of Automata, (CIAA 2007)*, volume 4783 of *Lecture Notes in Computer Science*, pages 11–23. Springer, 2007. `http://www.openfst.org`.

[3] A.-L. Bianne, F. Menasri, R. A.-H. Mohamad, C. Mokbel, C. Kermorvant, and L. Likforman-Sulem. Dynamic and contextual information in HMM modeling for handwritten word recogni tion. 99, 2011.

[4] A. Graves. Rnnlib: A recurrent neural network library for sequence learning problems. In *http://sourceforge.net/projects/rnnl/*.

[5] A. Graves, F. J. Gomez, and J. Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *International Conference on Machine Learning*, pages 369–376, 2006.

[6] A. Graves and J. Schmidhuber. Offline handwriting recognition with multidimensional recurrent neural networks.

[7] E. Grosicki, M. Carré, J.-M. Brodin, and E. Geoffrois. RIMES evaluation campaign for handwritten mail processing. In *Tenth International Conference on Frontiers in Handwriting Recognition*, pages 574–578, 2008.

[8] L. Guichard, A. H. Toselliand, and B. Couasnon. Handwritten word verification by svm-based hypotheses re-scoring and multiple thresholds rejection. 2010.

[9] S. Knerr and E. Augustin. A neural network-hidden markov model hybrid for cursive word recognition. *Fourteenth International Conference on Pattern Recognition*, 2:1518–1520, May 1998.