ICDAR 2009 Handwriting Recognition Competition

Emmanuèle Grosicki
DGA Centre d'Expertise Parisien,
16 bis avenue Prieur de la Côte d'Or,
94114 Arcueil cedex, France
Emmanuele.Grosicki@etca.fr

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

Abstract

This paper describes the handwriting recognition competition held at ICDAR 2009. This competition is based on the RIMES-database, with French written text documents. These document are classified in three different categories, complete text pages, words, and isolated characters. This year 10 systems were submitted for the handwritten recognition competition on snippets of French words. The systems were evaluated in three subtask depending of the sizes of the used dictionary. 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

The last years more than one competitions are organized at the International Conference On Document Analysis and Recognition (ICDAR). These competitions have shown how important to have a standard database to improve the classification techniques and to compare and evaluate different techniques and systems [6]. This experience with the benchmarking of systems has motivated us to organize such competition based on Latin script, and special with French handwriting recognition. Its goal is to evaluate automatic handwriting recognition systems on snippets of French handwritten words extracted from mails sent by individuals to companies or administrations. 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 [10] (Reconnaissance et Indexation de données Manuscrites et de fac similES) recognition and indexing of handwritten documents and faxes benchmarking has been used. It is composed of more than 12000 pages entirely annotated, 100000 snippets of characters and 250000 snippets of words. Automatic classification and recognition systems have been tested on three subtasks corresponding to three different sizes of the given dictionary (100, 1612, 5334). More information about the database and the project RIMES can be found in http://rimes.it-sudparis.eu.

This paper is organized as follows. In Section 2 the database and the test set are presented in some detail. Section 3 presents the participating groups and a short description of their systems. Section 4 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 had 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 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. 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 in-

978-0-7695-3725-2/09 $25.00 © 2009 IEEE
DOI 10.1109/ICDAR.2009.184

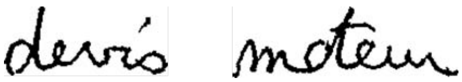

Figure 1. Examples of snippets of words.

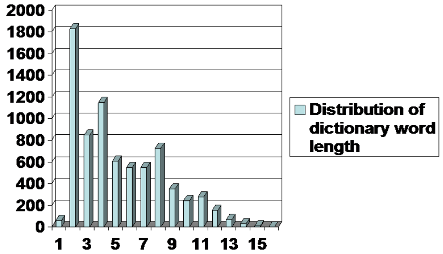

Figure 2. Distribution of word length

formation 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. The participants were given about 43000 snippets of words to train their system and a validation database of more than 7000 to test them. The unknown test dataset was composed of 7464 snippets. The distribution of the test word length is given in figure 2. Long words can correspond to two words linked by apostrophe or dash.

3. Participating Systems

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.

3.1. ParisTech

Three different systems were proposed by Anne-Laure Bianne (A2iA SA, Paris and Télécom ParisTech - TSI, Paris), Farès Menasri (A2iA SA, Paris), Ramy El-Hajj, Chafic Mokbel (University of Balamand, Lebanon), Laurence Likforman-Sulem (Télécom ParisTech - TSI, Paris), Christopher Kermorvant (A2iA SA, Paris):

- **Reference system**: The reference system proposed for the WR task is a combination of two different kinds of classifiers. One is based on HMMs and explicit segmentation and the others on HMMs (Hidden Markov Models) and sliding windows [1]. Three classifiers are combined with a Multi-Layer Perceptron (MLP) to build the reference system: $C1$ is a hybrid system composed of neural networks and hidden Markov Models [15]. This classifier has not been trained on the ICDAR'09 training database, but on a private database belonging to A2iA. It is case insensitive. $C2$ is based on the HCM-Toolkit provided by C. Mokbel [16] and sliding windows. These models are composed of left-right HMMs with skipping states and mixtures of Gaussians. This system uses 39 different classes and is case insensitive. $C3$ is based on HTK-Toolkit and sliding windows too. $C3$ is case sensitive, and is based on 65 models, transformed in tri-characters models. These models are then tied using a decision tree. The final number of models is about 1800.

- **Second system**: This system combines the $C2$ [15] and $C3$ classifiers, and another HTK based classifier, called $C4$. This classifier, as $C2$, is simply based on case insensitive models of words. A ranking method is applied to the outputs of the three classifiers. This method uses accumulation and comparison of the log-likelihoods given as the scores of the different classifiers.

- **Third system**: This system corresponds to the classifier $C3$ and it is a case sensitive system.

3.2. LITIS

The proposed system is based on a multi-stream segmentation free HMM for the recognition of two scripts handwritten words (Arabic and Latin scripts). In the first step, a set of pre-processing is applied to the word image. Two feature sets have been tested in this work: contours based feature are extracted respectively from the lower and the upper contours, and density based feature are calculated on two different sliding windows each with a particular width. For each feature type two feature streams representing the input word image are computed. Each stream model is then separately trained using Baum-Welch algorithm, combination weights of the multi-stream model are then optimized using relative frequency strategy. The last step is recognition during which the two HMM models are simultaneous decoded according to the multi-stream formalism detailed in [13, 14]

3.3. IRISA

The proposed system is based on Continuous Densities HMMs. It is composed of three modules: preprocessing, feature extraction and recognition. The pre-prossing 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. In this phase, size normalization of each word image is carried out implicitly by using its pre-computed base- and upper-lines which define into it the ascender, descender and main body text zones. On each of these zones, the feature extraction is applied independently and their resulting feature vectors sequences are then merged into a one unique sequence. The way of the features are computed is described in [2].

As was 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.

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

3.4. Itesoft & LORIA (ITESOFT)

The proposed system is based on the Non-Symmetric Half-Plane Hidden Markov Model (NSHP-HMM). This model combines a Markov Random Field (MRF) and an HMM, and works directly on binary patterns. The MRF analyzes pixel columns considering a local neighborhood for each pixel; the HMM synthesizes the MFR information along the whole image, allowing an horizontal elasticity.

The keyword spotting approach uses one NSHP-HMM model for each dictionary word, plus one that models the writing universe. Word models are dynamically built by concatenating letter NSHP-HMM, giving global models that work without any segmentation. Each word score is calculated as the ratio between the corresponding NSHP score and the universe model score [5, 4].

3.5. University of Toulon & A2iA (LSIS)

LSIS system is built in collaboration with A2iA grapheme recognizer. A2iA generated the likelihoods for each image of each recognized letter. Then LSIS integrated these lists of likelihoods in a new architecture based on human reading and cognition models. LSIS is working on the reading features integration at the human vision and memory levels. Cognitive models show that human reading is not linear. We then propose at LSIS to modeling very simply the integration of pre-recognized letters as proposed by the cognitive models. The model is then a bigram association of the estimates. This topology tackles the usual Viterbi integration, and thus results in different errors and recognitions. Up to now, we just built a preliminary system, and many optimization are possible, that will be conducted at LSIS in further research.

3.6. Universidad Politécnica de Valencia (UPV)

The proposed handwriting recognition system is a combination, by voting, of different classifiers. Three hybrid HMM/MLP systems have been trained varying the number of states of the HMMs and other parameters of the MLPs. A holistic classifier based on MLP has been trained for a subset of 61 words from the training vocabulary. This subset has been selected using as selection criteria the number of running words (> 20), size of the words (< 4 letters), the aspect ratio (≤ 2) and the width (≤ 250) of the word images. The holistic MLP receives the word image scaled to a fixed size (60x30 pixels) and outputs the posterior probability of every word in the restricted vocabulary. At recognition time, the words are classified using the three different hybrid HMM/MLP and those whose aspect ratio is below 2 and width below 250 are also classified with the MLP. Every classifier gives a list of N-best and these lists are combined by voting.

3.7. SIEMENS

The script word recognizer for the RIMES handwriting recognition contest is based on the Siemens Hidden Markov Recognizer (HMR) for Latin script that is widely in use within Siemens AG for postal automation projects.

Most of the system is based on techniques developed in 1993 presented in [12] and [3]. A feature vector sequence is created by a sliding window, followed by an HMM Viterbi decoding. Image preprocessing includes binarization, correction of slant and rotation, determination of writing lines and thus partitioning of the sliding window. Structural features are extracted from smoothed contour or skeleton, depending on configuration. Word HMMs are constructed by concatenation of character HMMs; these contain parallel left-to-right state paths representing character allographs. A series of improvements has been applied to the system, e.g. the dynamic determination of optimal topology for character HMMs [17]. Standard Viterbi is performed on the lexicon trie.

No algorithmic changes specific to the competition have been applied to the system. Various configurations have been trained and tested with the specified training and test sets. (Final training have been performed on the joint sets.) Trained configurations differ in preset preprocessing, feature extraction and modelling. The final system is a voted combination of 10 configurations, using a voting scheme which combines result confidences of the single systems. Three configurations are from different projects and are not based on the contest data, but do not contribute much to the final result. Optimization has been performed on the overall recognition rate of the validation data only. No detailed single image error analysis has been performed. But this would be useful for further adaptation of the system, especially by adjusting feature extraction.

3.8. TU München (TUM)

The submitted multilingual handwriting recognition system is based on a hierarchy of multidimensional recurrent neural networks [8, 7]. It can accept either on-line or off-line handwriting data, and in both cases works directly on the raw input (i.e. the pixel values or the sequence of pen positions) with no preprocessing or feature extraction. It uses the multidimensional Long Short-Term Memory network architecture [8, 7], an extension of Long Short-Term Memory [11] to data with more than one spatio-temporal dimension. In the case of handwriting recognition the networks are either one-dimensional (for online data) or two-dimensional (for off-line images). The basic structure of the system, including the hidden layer architecture and the hierarchical sub-sampling method is described in [9]. However the exact parameters (e.g. size of hidden layers, size of sub-sampling blocks etc.) varied from experiment to experiment. In addition the choice of output layer and objective function used for training depend on the network task.

Connectionist Temporal Classification [8, 7] is an recurrent neural network output layer designed for labeling sequences of data whose segmentation is ambiguous or difficult to determine, e.g. speech signals or cursive handwriting. It trains the network to map directly from the input sequence to a probability distribution over output label sequences, and therefore does not require either pre-segmentation or post-processing. This output layer was used for the online and off-line Arabic recognition competitions as well as the French recognition competition.

4. Tests and Results

At the start of the test period, each participant had access to the unknown test dataset composed of about 7000 snippets to run his own software on them in his own hardware environment. The result files in the expected format had to be sent back before the end of the test phase. 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.

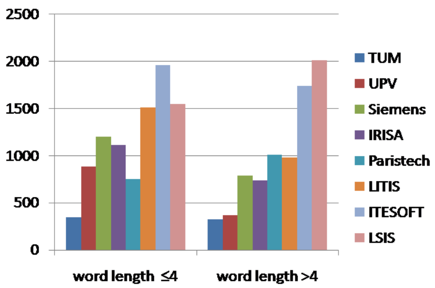

Figure 3. Number of errors vs. word length.

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 equal to 10) is added.

Moreover as some systems does not return any accent, the results have also been compared by normalizing the ground-truth. This step removed accents in the results files when needed.

We have thus evaluated the performance of 10 different handwriting recognition systems on three subtasks corresponding to three sizes of dictionaries:

- the subtask (WR1) where each word of the test is associated to a list of 99 words chosen randomly among test words in addition to the right transcription.
- the subtask (WR2) where the given dictionary is composed of all the test words (1612 words).
- the subtask (WR3) where the given dictionary contains also words of the training dataset (5334 words).

Most important results are given in tables 1, 2 and 3. More details will be presented at the ICDAR 2009 conference.

A comparison of the number of errors of the different systems with respect to the test word length is also shown in Figure 3. Some systems give more errors in short words whereas some others in long words. This may depend on different features and different normalization.

5. Conclusions

This competition has allowed us to evaluate 10 systems on 3 tasks of French handwritten recognition. The best results were achieved with a multilingual handwriting recognition system based on a hierarchy of multidimensional recurrent neural networks (**TUM**). 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 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.

Table 1. Recognition rate (%) on task WR1.

System	GT		System	normalized GT	
	top_1	top_{10}		top_1	top_{10}
TUM	98.43	99.92	TUM	98.46	99.92
UPV	96.45	99.60	ParisTech(1)	97.17	99.76
SIEMENS	95.30	99.71	UPV	96.49	99.60
LITIS	92.40	99.26	SIEMENS	95.39	99.71
IRISA	91.24	96.33	LITIS	92.46	99.28
ParisTech(1)	86.37	88.79	ParisTech(2)	92.46	99.16
ParisTech(2)	82.05	88.25	IRISA	91.29	96.41
ParisTech(3)	79.90	87.98	ParisTech(3)	90.13	98.85
ITESOFT	74.57	85.28	ITESOFT	84.74	96.14

Table 2. Recognition rate (%) on task WR2

System	GT		System	normalized GT	
	top_1	top_{10}		top_1	top_{10}
TUM	93.17	98.95	TUM	93.37	98.95
UPV	86.11	97.95	ParisTech(1)	90.53	98.83
SIEMENS	81.30	96.37	UPV	86.44	98
ParisTech(1)	80.20	87.94	ParisTech(2)	81.71	96.03
IRISA	79.56	92.78	SIEMENS	81.66	96.38
LITIS	74.10	94.88	IRISA	80	93.34
ParisTech(2)	72.36	85.44	LITIS	74.57	94.94
ParisTech(3)	63.80	80.05	ParisTech(3)	72.70	90.37
ITESOFT	59.39	76.71	ITESOFT	68.07	86.91

Table 3. Recognition rate (%) on task WR3

System	GT		System	normalized GT	
	top_1	top_{10}		top_1	top_{10}
TUM	91.02	98.34	TUM	91.39	98.35
UPV	83.17	96.84	ParisTech(1)	86.08	97.68
ParisTech(1)	76.34	86.86	UPV	83.76	96.86
IRISA	74.69	90.29	ParisTech(2)	76.67	93.93
SIEMENS	73.24	93.39	IRISA	75.71	91.45
ParisTech(2)	67.99	83.57	SIEMENS	74.36	93.45
LITIS	66.65	91.61	LITIS	67.42	91.68
ParisTech(3)	58.70	77.21	ParisTech(3)	66.76	87.23
LSIS	52.37	54.86	ITESOFT	57.60	82.57
ITESOFT	50.44	72.94	LSIS	57.33	60.36

References

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

[2] I. Bazzi, R. Schwartz, and J. Makhoul. An omnifont open-vocabulary ocr system for english and arabic. *T-PAMI*, 21(6):495–504, June 1999.

[3] T. Caesar, J. Gloger, and E. Mandler. Preprocessing and feature extraction for a handwriting recognition system. In *2nd ICDAR*, pages 408–411, 1993.

[4] C. Choisy. Dynamic handwritten keyword spotting based on the nshp-hmm. In *Proc. 9^{th} ICDAR*, volume 1, pages 242–246, 2007.

[5] C. Choisy and A. Belaid. Cross-learning in analytic word recognition without segmentation. *IJDAR*, 4:281–289, 2002.

[6] H. El Abed and V. Märgner. Base de données et compétitions - outils de développement et d'évaluation de systèmes de reconnaissance de mots manuscrits arabes. In *10th Colloque International Francophone sur l'Ecrit et le Document , CIFED*, 2008.

[7] A. Graves. *Supervised Sequence Labelling with Recurrent Neural Networks*. PhD thesis, IDSIA (Istituto Dalle Molle di Studi sull'Intelligenza Artificiale), 2007.

[8] A. Graves, S. Fernàndez, and J. Schmidhuber. Multi-dimensional recurrent neural networks. In *Proceedings of the International Conference on Artificial Neural Networks*, 2007.

[9] A. Graves and J. Schmidhuber. Offline handwriting recognition with multidimensional recurrent neural networks. In *Advances in Neural Information Processing Systems 21*, 2009.

[10] E. Grosicki, M. Carré, J.-M. Brodin, and E. Geoffrois. RIMES evaluation campaign for handwritten mail processing. In *Proc. 10th*, pages 574–578, 2008.

[11] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[12] A. Kaltenmeier, T. Caesar, J. Gloger, and E. Mandler. Sophisticated topology of hidden markov models for cursive script recognition. In *2nd ICDAR*, pages 139–142, 1993.

[13] Y. Kessentini, T. Paquet, and A. Benhamadou. Multi-script handwriting recognition with n-streams low level features. In *Proc. 19th Inter. Conf. on Pattern Recognition (ICPR)*, pages 1–4, 2008.

[14] Y. Kessentini, T. Paquet, and A. Benhamadou. A multi-stream hmm-based approach for off-line multi-script handwritten word recognition. In *Proc. 10th*, pages 147–152, 2008.

[15] S. Knerr and E. Augustin. A neural network-hidden markov model hybrid for cursive word recognition. In *Proc. Fourteenth International Conference on Pattern Recognition*, volume 2, pages 1518–1520 vol.2, 1998.

[16] C. Mokbel, H. A. Akl, and H. Greige. Automatic speech recognition of Arabic digits over telefone network. In *RTST*, 2002.

[17] M.-P. Schambach. Model length adaptation of an HMM based cursive word recognition system. In *7th ICDAR*, pages 109–113vol.1, 3-6 Aug. 2003.