Multilingual Off-line Handwriting Recognition in Real-world Images

Michał Kozielski, Patrick Doetsch, Mahdi Hamdani, Hermann Ney
Human Language Technology and Pattern Recognition Group
Chair of Computer Science 6
RWTH Aachen University, D-52056 Aachen, Germany
{kozielski,doetsch,hamdani,ney}@i6.informatik.rwth-aachen.de

***Abstract*—We propose a state-of-the-art system for recognizing real-world handwritten images exposing a huge degree of noise and a high out-of-vocabulary rate. We describe methods for successful image denoising, line removal, deskewing, deslanting, and text line segmentation. We demonstrate how to use a HMM-based recognition system to obtain competitive results, and how to further improve it using LSTM neural networks in the tandem approach. The final system outperforms other approaches on a new dataset for English and French handwriting. The presented framework scales well across other standard datasets.**

I. Introduction

In the handwriting community we are constantly presented with more and more challenging datasets. The datasets in one of the factors that drives the research in this field, as we are trying to understand the data we are dealing with. The increasing complexity of the data helps systems evolve. In can be noted that the publications are going in two main directions. The first one tries to directly adapt methods from speech recognition and analyze how they work in the context of handwriting. The second one aims to come up with accurate preprocessing schemes that have their roots in the computer vision community. Usually those two branches of research goes separately - for example there are different competitions on text line segmentation, binarization, and handwriting recognition. It would be very interesting to have publications that treat the whole process - from raw image to transcribed text - in a top to bottom manner. That kind of approach can address many critical problems: how do methods combine with each other, how do methods scale across other datasets, and so on. The goal of this work is to be such a step ahead.

II. State-of-the-art

IFN/ENIT [1] is a relatively small dataset that however requires going beyond explicit character segmentation and fostered the use of sliding window [2]. It consists of binarized images of handwritten Arabic words and the task is isolated word recognition with relatively small vocabulary that is closed over the test set. The good quality of the images encouraged preprocessing-free approaches [3], where any kind of hard decision about the data was postponed until the end of search process. Another publication exploits properties of the Arabic script by using explicit intra-word whitespace models [4].

The RIMES dataset [5] consists of handwritten French words. It comes with much bigger vocabulary than the IFN/ENIT corpus and introduces a problem of preprocessing, for example slant correction [6]. There have been many publications on this dataset and the error rates are relatively low, the best system scored below 5% word error rate [7]. This and other datasets were employed to show that the systems can get improvement from the use of context dependant models [6][8]. That dataset contains also the paragraph (block) recognition task, where the goal is to recognize whole sequences, spanning across multiple lines. The text lines are annotated in the corpus, so the systems do not have to perform the explicit text line segmentation.

The IAM dataset [9] consists of handwritten English sentences. It was a common approach in the literature to limit the vocabulary of the language model to produce a considerable amount of out-of-vocabulary (OOV) words in the test. That was a playground for the first open-vocabulary approaches that tried to address that problem with character-based language models [10][11]. Also many systems have been build that recognize the paragraphs line by line and have had the disadvantage of emptying the language model history at the beginning of every new line [12][13]. Another approach was to concatenate feature vectors across all lines and to recognize the paragraph as a whole [14].

The OpenHart evaluation [15] was the first continuous handwriting recognition task for Arabic. Arabic script is very interesting to work with, because a lot of work has to be put into the creation of a proper lexicon and vocabulary [1]. This dataset is also two orders of magnitude bigger than the other corpora, which poses a challenge to learn how to work with such a big data.

The Maurdor dataset[1] is the first one that in some way resembles real-world data. It introduces a whole new sort of preprocessing challenges, like for example: form-field artifacts, horizontal lines, complex background, highly-degraded characters, non-uniform skew. The OOV rate is considerably high because the annotations contain a lot of names, dates, numbers, and so on. It was also the first corpus to be annotated on paragraph and not line level, which incorporates the text line segmentation errors into recognition errors.

In this paper we present a system that addresses all the challenges introduced by the Maurdor dataset and achieves highly-competitive results.

[1]At the time of writing the Maurdor dataset was not public, but it was distributed to the participants of the Maurdor evaluation (http://www.maurdor-campaign.org/). The registration was open and free of charge. Participants were allowed to keep the data after the evaluation.

III. Proposed system

A. Preprocessing

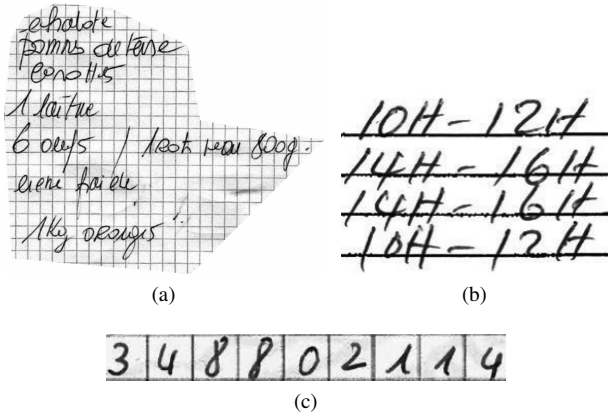
A crucially important issue that is usually omitted in the publications is to fix the spatial resolution of the images. Usually it is not a problem, because the documents are scanned in a consistent way and that process provides a one-to-one mapping between pixels and the sheet of paper. This is the case of most of the standard off-line handwriting recognition datasets. If the resolution of images (measured in dots per pixel) is not provided in the corpus (as it is usually the case), it can be usually retrieved from the image file properties, or guessed using some heuristics. For the Maurdor dataset we make the assumption that all documents are in A4 format (210×297mm) and scale them to the same size. In case of the images coming from "fax" the aspect ratio may be have been altered which has to be taken into consideration.

Binarization algorithms can be used for image denoising, although rather unwanted, because all our techniques should be designed to work with grey-scale images. In our work we use the well-know Otsu algorithm, which has this very appealing property of having zero parameters, in contrast to window-based algorithms. Before binarization the image should be blurred with a small window (3x3 pixels) to remove simple noise resulting from image acquisition and compression.

The next step is to deskew and deslant the image. In this work we correct the slant of images with a median of angle values estimated by three different deslanting algorithms [16][17][18]. The algorithms work by shearing the image with an angle from a certain range and evaluating those transformations with different objective functions. The deskewing algorithm follows the same design principle. We rotate the image with a certain angle and then we look for the biggest variance of its vertical projection.

We have applied the probabilistic Hough transform [19] to line removal with great success. It has a very useful ability to detect lines with gaps. Detection of such lines is however dangerous, because we might remove character strokes that accidentally form a line. Therefore we run the detection multiple times with different thresholds. The bigger gaps we allow, the longer the line has to be in total. Our implementation is based on OpenCV 2.3. The thresholds we use for the length of a line and a gap are: 75 and 0, 100 and 5, 200 and 10, 400 and 15. In reality a line has a certain width of a few pixels and can be covered with longer and shorter segments. The implementation of line removal follows [20].

Our implementation of text line segmentation is based on the CUBS algorithm [21]. We perform the run-length smoothing with an ellipse-shaped window, as described in the original paper. Our estimates of the parameters differ however, we use 270 and 6.5 pixels for the axis of the ellipse. Afterwards we apply Otsu binarization to the smoothed image. The connected components from the resulting binary map represent text line shapes. Connected components from the original image (character shapes) touching the text line shapes are thought to belong to that line. If a connected component touches more than one line or no line at all, it is left unclassified for a moment. We then grow the text line shapes from the binary map using morphological dilation. The leftovers are classified (and split) to a text line using the union operation between the line shape and the connected components (from the original image).



(a) (b) (c)

Fig. 1: Example images from the Maurdor dataset.

We use the sliding window approach with a constant shift to extract a sequence of overlapping frames from a line of text. We then independently normalize every frame using moments to make it shift and scale invariant [22][14]. A curious reader can compare this explicit approach with implicit approaches, where the invariance is incorporated int the model [23]. Every frame extracted with the sliding window is transformed into a single feature vector. The gray-scale values of all 256 pixels in a frame are used as features and are further reduced by means of Principal Component Analysis (PCA) to 20 components. The number of principal components has a small influence on the recognition performance. A final feature vector of size 24 is produced by adding the original moments of a frame to the output of PCA. Feature vectors extracted from a single paragraph of text (multiple text lines) are concatenated to form one segment. Note that because one paragraph usually contains multiple sentences, the language model has to be able to hypothesize the sentence boundary.

B. Modeling

We train two systems using the preprocessing scheme described in the previous section.

The first one is a standard HMM-based system. Every character is represented by a Viterbi-trained left-to-right HMM with loop transitions. The models consists of six states. The transitions (treaded as penalties in negative log scale) are constant and fixed across all models. There is an additional penalty for exiting the HMM model, which controls the number of insertions during recognition. The whitespace model is a special one and has only one state and no loop penalty. We model the emission distributions using Gaussian mixtures with 128 densities and a globally-pooled diagonal covariance matrix. The parameters have been optimized experimentally on the development set.

We also trained an LSTM recurrent neural networks [24] using the alignment from the HMM training. The outputs of this network constitute a different feature set that can be used in combination with the standard HMM-based system (tandem approach) [25]. Another extension to the system allowed for

TABLE I: Statistics of the language models.

	French	English
Vocabulary size	11k	7.8k
OOV rate	26%	32%
In-lexicon PPL	3.0	4.0
Out-of-vocabulary PPL	16.2	19.8

altering the number of HMM states per character using a fixed alignment from a previous training procedure. As the final step of the training procedure we applied a discriminative training procedure to the HMM models using the M-MPE criterion. Those modification to the standard HMM-based system constitute what we call our improved system [14].

Our decoder is based on weighted finite-state transducers (WFST) [26]. We use an on-the-fly combination of two language models for out-of-vocabulary word recognition [11]. Additionally to a standard word-level language model we use a separate n-gram character-level language model. The probabilities assigned by those two models are combined into one decision rule.

Our system has been trained to recognize text written in French and English. Because those languages share almost the same alphabet, we trained the visual models (Gaussian mixtures) on the English and French data together. However as the vocabularies are different, the language models have been trained separately.

IV. Experiments

A. Dataset

The Maurdor dataset contains 3000 pages for training and 1000 for development. It is composed of French (50%), English (25%) and Arabic (25%) printed text (75%) and handwriting (25%). Documents are divided into five categories. This dataset is annotated at paragraph-level, which means that the explicit text line segmentation is not provided. The language and writing type of a every paragraph is annotated in the corpus. The handwritten French and English parts contain 359k and 136k running characters accordingly.

From the training set we removed text paragraphs annotated as ”signature”. Then we corrected the text line segmentation errors on the training set manually. We also manually removed badly binarized and deskewed images. No manual intervention was performed on the development and evaluation sets.

Figure 1 shows three example images from this dataset without keeping the original spatial resolution.

B. Training details

The systems have been trained using exclusively the data provided in the corpus. We have built the language models upon transcriptions of the training set. For that purpose we have used both the handwritten and typed parts of the dataset. The character inventory contains 104 characters plus two for whitespace and noise. We use the special noise character for characters that were annotated as unclassified (due to low-quality images) or for characters that occur less than ten times in the training set to avoid computational problems.

TABLE II: Results on the Maurdor development set.

System	French		English	
	CER	WER	CER	WER
Standard	16.9	39.2	23.2	50.8
Improved	13.8	34.3	17.0	45.3

TABLE III: Results with respect to category on the Maurdor development set for French using the standard system.

Category	running chars.	CER	WER
Forms	15K	22.5	52.8
Notes 1	16K	25.0	54.1
Letters	49K	10.6	28.0
Notes 2	9K	21.4	47.1
Other	2K	32.8	61.4

Table I shows basis statistics related to the language models. We report the perplexities (PPL) on character level separately for words from the lexicon and out-of-vocabulary words. The computation of perplexities includes the word boundary after every word (even the last one) as a special symbol. As word-level language model we use a standard 3-gram model with modified Kneser-Ney discounting built upon the training text source containing one sentence per line. We excluded the singletons and numbers from the vocabulary of the word-level language models. The 10-gram character-level language model has been built upon a list of out-of-vocabulary words extracted from the training set. The character-level language model was shared across languages. The word-level language models were trained separately for each language.

C. Experimental results

Error rates are calculated using the case-sensitive Levenshtein alignment between reference and hypothesis. The word boundary after every word (even the last one) is included as a special symbol for the computation of the character error rate (CER). In addition to recognition experiments we perform a visual assessment of the preprocessing algorithms.

Table II show the results for the French and English handwriting recognition task. The tandem approach with use of LSTM neural network is better than the standard HMM-based system, however we have to point out that the standalone HMM approach is not much worse. Table III shows a breakdown of the results for the recognition of French with respect to document category (using the standard system). Results in the category ”letters” are significantly better, because this category contains long sentences sentences that are easier to segment, have better context, and a small OOV rate.

Although the design principles of the text line segmentation algorithm are pretty simple, we found out that this approach is very good in practice. In our experiments around 10% of paragraphs had problems with text line segmentation, however this did not affect the recognition performance too much, because images that were difficult to segment were also difficult to recognize.

By visual assessment the deslanting algorithm produces almost perfect results. The deskewing algorithm fails on im-

TABLE IV: Evaluation of the preprocessing scheme and the standard HMM-based system on standard datasets for off-line handwriting recognition.

Dataset	CER	WER
IFN/ENIT	2.4	4.0
RIMES word	4.0	10.4
IAM	4.7	12.6
RIMES paragraph	5.5	15.7
OpenHart	10.7	25.8
Maurdor French	16.9	39.2
Maurdor English	23.2	50.8

ages with short lines. To circumvent this issue we apply the deskewing only to images of a certain size. The line finding algorithm is very accurate and robust, however the line removal had problems removing lines that went through the character strokes. As for the denoising with Otsu we were not able to remove all artifacts and that hurt the performance. In our opinion one needs more sophisticated, well-tuned, several-stage algorithms to remove complex background. The application of Otsu binarization reduced the word error rate (WER) for French from 47.8% to 42.8%. Line removal improved the result to further 39.2%. We analyzed the influence of deslanting and moment normalization on the error rate in the following publication [14].

An important modeling question was to answer whether it is better to train the models for English and French together or separately. It turned out that one common visual model is better, which is quite obvious as latin scripts share practically the same set of characters. When we trained the system for English using the English data exclusively the WER increased by 15% relative in comparison to the system trained using all data. Modeling high OOV rates was a challenging issue, especially because the perplexity for out-of-vocabulary words was very high. The open-lexicon approach was only partially successful as many of the OOVs were not real words (but e.g. serial numbers) and thus there were little dependence between characters.

We performed several different experiments that in the end did not improve the results. The recognition of form-field paragraphs was difficult because of the presence of artifacts (See Figure 1c). We tried to learn them explicitly (as an additional HMM model), as they were annotated in the corpus, but the error rate increased. Another experiment was to train a writer-dependant system using the CMLLR transformation [27] with document category as writer information, however again without success.

D. Comparison on standard datasets

Another important question was to find out how our preprocessing scheme scales across other standard datasets. Otsu binarization performs at least as good as simple contrast normalization methods [14] and has the advantage of having no parameters. The deslanting algorithm produces almost perfect results, however we do not apply it to Arabic script. The deskewing algorithm fails on images with short lines, this can be however circumvented using simple heuristics. Line detection is very accurate, which means that the line removal is not fired on clean images. Text line segmentation performs well on clean images. If there is only a single line it misbehaves rarely. A certain source of problem is if the lines are of a different width (due to run-length smoothing).

TABLE V: Official results of the handwriting recognition task of the first Maurdor evaluation.

System	French		English	
	CER	WER	CER	WER
Our system	20.8	34.5	20.0	38.0
Participant 1	41.1	71.7	32.5	59.0
Participant 2	67.7	98.1	85.8	119.1
Participant 3	75.5	98.6	84.6	103.4
Participant 4	102.0	173.8	124.7	201.8

In Table IV we report the results on standard datasets for off-line handwriting recognition. For IFN/ENIT we train using the sets a,b,c,d and test on the set e. For RIMES word we test using the test set from ICDAR 2009 competition, for RIMES paragraph using the test set from ICDAR 2011 competition, for IAM using the validation set, for Maurdor using the development set from the first evaluation, for OpenHart using the development set from the 2013 evaluation.

We use our standard HMM-based system in all experiments. For OpenHart we use a slightly modified version of our system [28].

E. Official results

Table V shows the official results of the first Maurdor evaluation for the recognition of handwritten French and English text. There were five participants. The evaluation set without transcriptions was given to the participants two weeks before deadline for submission of the recognition results. Our system outperformed other approaches and scored the first place for Latin script recognition. The language of a given text paragraph was given, our system did not have to recognize it. We did not participate in the Arabic handwritten text as well as in the typed text recognition task. The official results were obtained from the organizers of the evaluation. Following the request from organizers we anonymized the names of other participants.

V. Conlusions

We have demonstrated that recognizing real-world images suit well into our standard HMM framework. Many of the existing algorithms can be adapted and combined to build a systems that can successfully address increasing challenges. Good results can be obtained using the more sophisticated discriminative framework as well as model trained in the Maximum-likelihood fashion. Furthermore the Maurdor dataset is a great opportunity for different researchers inside the handwriting community to get together and work on improving approaches to recognizing the real-world data.

Acknowledgments. This work was partially supported by a Google Research Award and by the Quaero Programme, funded by OSEO, French State agency for innovation.

REFERENCES

[1] M. Pechwitz, S. S. Maddouri, V. Märgner, N. Ellouze, and H. Amiri, "IFN/ENIT - database of handwritten arabic words," in *Colloque International Francophone sur l'Ecrit et le Document (CIFED)*, Oct. 2002, pp. 129–136.

[2] M. Pechwitz and V. Maergner, "HMM based approach for handwritten arabic word recognition using the IFN/ENIT - database," in *Document Analysis and Recognition, 2003. Proceedings. Seventh International Conference on*, 2003, pp. 890–894.

[3] P. Dreuw, G. Heigold, and H. Ney, "Confidence- and margin-based MMI/MPE discriminative training for off-line handwriting recognition," *Int. J. Doc. Anal. Recognit.*, vol. 14, no. 3, pp. 273–288, Sep. 2011.

[4] P. Dreuw, S. Jonas, and H. Ney, "White-space models for offline arabic handwriting recognition," in *International Conference on Pattern Recognition*, Tampa, Florida, USA, Dec. 2008, pp. 1–4.

[5] E. Grosicki and H. El Abed, "ICDAR 2009 handwriting recognition competition," in *Document Analysis and Recognition, 2009. ICDAR '09. 10th International Conference on*, Jul. 2009, pp. 1398–1402.

[6] A.-L. Bianne-Bernard, F. Menasri, R.-H. Mohamad, C. Mokbel, C. Kermorvant, and L. Likforman-Sulem, "Dynamic and contextual information in HMM modeling for handwritten word recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 10, pp. 2066–2080, 2011.

[7] F. Menasri, J. Louradour, A.-L. Bianne-Bernard, and C. Kermorvant, "The A2iA French handwriting recognition system at the Rimes-ICDAR2011 competition." in *Document Recognition and Retrieval Conference*, ser. SPIE Proceedings, C. Viard-Gaudin and R. Zanibbi, Eds., vol. 8297. SPIE, 2012.

[8] G. Fink and T. Plotz, "On the use of context-dependent modeling units for HMM-based offline handwriting recognition," in *Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on*, vol. 2, 2007, pp. 729–733.

[9] U.-V. Marti and H. Bunke, "The IAM-database: an English sentence database for offline handwriting recognition," *International Journal on Document Analysis and Recognition*, vol. 5, pp. 39–46, Nov. 2002.

[10] F. Zamora-Martinez, M. Castro-Bleda, S. España-Boquera, and J. Gorbe-Moya, "Unconstrained offline handwriting recognition using connectionist character n-grams," in *Neural Networks (IJCNN), The 2010 International Joint Conference on*, Jul. 2010, pp. 1–7.

[11] M. Kozielski, D. Rybach, S. Hahn, R. Schlüter, and H. Ney, "Open vocabulary handwriting recognition using combined word-level and character-level language models," *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2013.

[12] H. Bunke, S. Bengio, and A. Vinciarelli, "Offline recognition of unconstrained handwritten texts using HMMs and statistical language models," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 26, no. 6, pp. 709–720, 2004.

[13] S. España-Boquera, M. Castro-Bleda, J. Gorbe-Moya, and F. Zamora-Martinez, "Improving offline handwritten text recognition with hybrid HMM/ANN models," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 4, pp. 767–779, Apr. 2011.

[14] M. Kozielski, P. Doetsch, and H. Ney, "Improvements in RWTH's system for off-line handwriting recognition," *International Conference on Document Analysis and Recognition*, Aug. 2013.

[15] V. M. A. Tong, M. Przybocki and H. E. Abed, "Nist 2013 open handwriting recognition and translation (openhart'13) evaluation," in *Proceedings of the NIST 2013 Open Handwriting and recognition Workshop*, Washington DC, Aug. 2013.

[16] M. Pastor, A. Toselli, and E. Vidal, "Projection profile based algorithm for slant removal," in *Image Analysis and Recognition*, ser. Lecture Notes in Computer Science, A. Campilho and M. Kamel, Eds. Springer Berlin / Heidelberg, 2004, vol. 3212, pp. 183–190.

[17] A. Vinciarelli and J. Luettin, "A new normalization technique for cursive handwritten words," *Pattern Recognition Letters*, vol. 22, no. 9, pp. 1043–1050, 2001.

[18] M. P. i Gadea, A. H. Toselli, V. Romero, and E. Vidal, "Improving handwritten off-line text slant correction," in *Procc. of The Sixth IASTED international Conference on Visualization, Imaging, and Image Processing (VIIP 06)*, 2006.

[19] J. Matas, C. Galambos, and J. Kittler, "Robust detection of lines using the progressive probabilistic hough transform," *Comput. Vis. Image Underst.*, vol. 78, no. 1, pp. 119–137, Apr. 2000.

[20] S. Zhixin, S. Setlur, and V. Govindaraju, "Removing rule-lines from binary handwritten arabic document images using directional local profile," in *Pattern Recognition (ICPR), 2010 20th International Conference on*, 2010, pp. 1916–1919.

[21] ——, "A steerable directional local profile technique for extraction of handwritten arabic text lines," in *Document Analysis and Recognition, 2009. ICDAR '09. 10th International Conference on*, 2009, pp. 176–180.

[22] M. Kozielski, J. Forster, and H. Ney, "Moment-based image normalization for handwritten text recognition," in *International Conference on Frontiers in Handwriting Recognition*, Bari, Italy, Sep. 2012, pp. 256–261.

[23] T. Bluche, H. Ney, and C. Kermorvant, "tandem HMM with convolutional neural network for handwritten word recognition," *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2013.

[24] A. Graves, M. Liwicki, S. Fernandez, R. Bertolami, H. Bunke, and J. Schmidhuber, "A novel connectionist system for unconstrained handwriting recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 5, pp. 855–868, May 2009.

[25] H. Hermansky, D. P. W. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems," in *Acoustics, Speech, and Signal Processing (ICASSP), 2000 IEEE International Conference on*.

[26] D. Rybach, H. Ney, and R. Schlüter, "Lexical prefix tree and WFST: A comparison of two dynamic search concepts for LVCSR," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 6, pp. 1295 –1307, Jun. 2013.

[27] P. Dreuw, D. Rybach, C. Gollan, and H. Ney, "Writer adaptive training and writing variant model refinement for offline Arabic handwriting recognition," in *International Conference on Document Analysis and Recognition*, Barcelona, Spain, Jul. 2009, pp. 21–25.

[28] M. Hamdani, P. Doetsch, M. Kozielski, H. Pesch, A. El-Desoky Mousa, and H. Ney, "The RWTH large vocabulary arabic handwriting recognition system," *NIST 2013 Open Handwriting and Recognition Workshop*, Aug. 2013.