Writing type and language identification in heterogeneous and complex documents

D. Hebert*, P. Barlas*, C. Chatelain†, S. Adam*, T. Paquet*
*Laboratoire LITIS - EA 4108, Universite de Rouen, FRANCE 76800
Email: david.hebert@univ-rouen.fr, philippine.barlas@insa-rouen.fr,{Sebastien.Adam,Thierry.Paquet}@litislab.eu
†Laboratoire LITIS - EA 4108, INSA Rouen, FRANCE 76800
Email: clement.chatelain@insa-rouen.fr

***Abstract*—This paper presents a system dedicated to automatic recognition of both the writing type and the language of text regions in heterogeneous and complex documents. This system is able to process documents with mixed printed and handwritten text, in various languages (French, English and Arabic). To handle such a problem, we divided it into two sub-tasks : the writing type identification and the language identification. The method for the writing type recognition is based on the analysis of the connected components while the language identification approach combines the analysis of connected components and the analysis of character distributions. We present the results obtained by the system during the second competition round of the MAURDOR campaign, and show that the performance of our system compares favorably with other participants.**

***Keywords*—writing type identification; language identification; document processing; codebook; character distribution;**

I. Introduction

Intelligent reading systems are still challenging for the document analysis community. Despite some great improvements made in the last years [1], it is still difficult to design a global OCR system able to read any character, in any script, with honorable performance. Furthermore, many documents mix printed and handwritten type. For example bank checks, application forms, annotated documents. These documents represent an additional difficulty in the automatic transcription since each writing type need to be processed using a specific recognition engine. Moreover, automatic transcriptions require complex language models to increase the transcription reliability. Therefore, writing type, script (alphabet) and language identification is a real need in order to design a generic document recognition system, able to automatically segment and transcript any kind of document.

In this paper, we present a system able to automatically recognize the writing type, and the document language between Arabic, French and English. Our system is based on morphological characterization between printed and handwritten text in order to detect the writing type and the alphabet (script). Bi-gram distribution of characters is used to discriminate between languages. As we do not have the transcription, the bi-gram profiles of languages are estimated on an OCR output. This strategy allows to take account of the OCR errors as an additional characteristic for the language description.

Our writing type and language identification system is a part of a complete processing chain dedicated to the automatic analysis of heterogeneous and complex documents. The overall system was evaluated in 2013 during the two MAURDOR campaigns [9]. These campaigns were led to evaluate the progress in automatic reading of heterogeneous documents and made an important step beyond existing ones [4], [13] regarding the variability of the documents to be processed. Indeed, the dataset contains heterogeneous documents (blank or completed forms, printed and manually annotated business documents, handwritten correspondence, maps, ID, newspapers articles, blueprints, etc.), with mixed printed and handwritten texts, in various languages (French, English and Arabic). Moreover, the MAURDOR campaigns assess not only a complete processing chain (starting from document segmentation to information retrieval, through text recognition) but also each module of the processing chain independently. In order to lead this evaluation of different tasks, document analysis problem have been divided into five subtasks respectively dedicated to segmentation, writing type identification, language identification, text recognition for each type/language, and information retrieval.

The paper is organized as follows. Section II presents the related works on the automatic writing type, script and language identification methods in the literature. In section III, we propose an overview of our processing chain. Then, the writing type and language identification approaches are detailed in section IV and V. Section VI presents experimental results obtained during the MAURDOR campaigns. Finally, the paper concludes with a brief summary and a discussion of future work.

II. Related Works

The writing type, the script and the language identification problems have received considerable attention in the past. From a global point of view, state-of-the-art methods for identifying writing type and the script alphabet of a document are rather similar, generally they are based on physical descriptors extracted from the text shapes. However, the identification becomes more difficult when several languages sharing the same alphabet are considered, as it is the case for French/English language identification for example. In this case, a recognition stage is often performed in order to capture some statistical language particularities.

Writing type and script alphabet identification

[21] proposes a printed/handwritten text discrimination system based on the classification of a set of physical features

2167-6445/14 $31.00 © 2014 IEEE
DOI 10.1109/ICFHR.2014.75

(Gabor, run-length histogram, co-occurrence, . . .) and a filtering process by Markov Random Fields. [19] uses codebook of features based on character shapes to discriminate handwritten from printed text on Arabic documents. Physical features are also used for script alphabet identification. [20] proposes a classification scheme based on features extracted from connected components (centroid, number of white holes, . . .) for script alphabet identification on handwritten documents. These approaches that essentially focus on the script alphabet identification, globally follow a general scheme that we can synthesize as a physical feature extraction feeding a classifier.

Language identification

For language identification, some other works rely on language model and statistical analysis of characters [16], keywords/short words [15] or n-gram of characters [14], [15]. [17] made a combination of these three types of analysis with a ranking combination strategy to improve the identification rate on two digital documents databases. Also based on a n-gram modeling process, [16] defines Markov models of each language and try to find the best fitting model for a new sequence of characters. More recently, [18] has defined a method able to identify both the language of a small paragraph and the block of text of same language in a multi-language document. However, this approach is designed to deal with web pages, where the text transcription is available. Most of these approaches are language analysis systems that do not address the text transcription problem and assume to have a perfect text transcription. Finally, when transcription is available, n-gram statistical analysis gives reliable language identification, whereas shape or texture analysis are preferred when only the image of the text is available. To the best of our knowledge, language identification methods are limited to digital documents.

In this article, we propose a writing type and script alphabet identification method based on an original codebook-based feature set. We also describe a language identification method able to deal with document images. This approach relies on the statistical analysis of an OCR output. These two methods are described in the sections IV and V of this paper.

III. Overview of the processing chain

In the context of the MAURDOR campaigns, the LITIS has developed a complete processing chain corresponding to the five subtasks evaluated in the campaigns. As one can see on Fig 1, the first task consists in segmenting the document image into homogeneous areas, in particular by setting apart writing areas. The system developed for the tasks of document segmentation is composed of a text detector [2] based on both connected component information and a document segmentation method based on white zone. Tables are also detected using a line detection approach [3]. Once the text areas are identified, we proceed to the writing type and language identification according to the methods presented in this paper. This identification module allows to process each text block with the appropriate recognizer. Our recognition engines are based on Hidden Markov Models (HMM) of characters combined with the appropriate language model and language dictionary [22]. The final step consists in extracting the logical structure of a document finding semantic information in the text areas (title, legend, date, coordinates, object . . .) or, where applicable, finding a reading order between the various areas (for instance, a column sequence in a press article). The system developed for this task is a combination of a learning based approach for the semantic information detection and a simple rule-based approach for the detection of the reading orders.

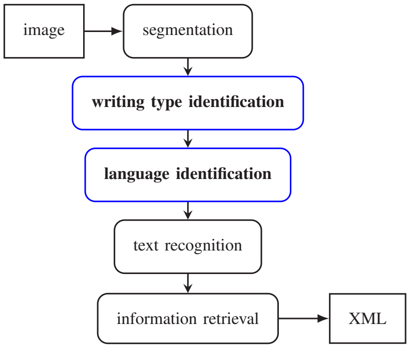

Fig. 1. Overview of the proposed system

The following sections focus on the two steps dedicated to the script and language identification as depicted in Fig 1.

IV. Writing type identification

The separation of text areas into printed areas and handwritten areas is an important step in the automatic transcription of complex documents. As the MAURDOR dataset includes three main languages : French, English and Arabic, the writing type identification in this multilingual context is further more complicated since unlike French and English, the Arabic writing printed is cursive. Moreover, the writing type identification must be performed on text blocks composed of single character or words as well as several paragraphs.

This section presents the proposed system to handle the issue of writing type identification. The classification between handwritten and printed components relies on a codebook based approach, inspired from the methods described in [6] and [7], both used for the writer identification. First, a collection of contour fragments, which are popular shape descriptors, is extracted from a first connected components learning dataset. This collection is used to build a codebook of contour fragments. Then a MLP classifier is trained using histograms of contour fragments on a second learning dataset.

A. Codebook construction

Fragment extraction and representation: An efficient way to discriminate writing type is to extract fragments of external contour of connected components. A contour fragment is defined by a fixed length l and an overlapping area of fixed size s, moving along the external contour of the connected component as illustrated on Fig 2. The overlapping area represents the number of pixels shared between fragment i and fragment $i + 1$. Fragments are extracted over the whole contour of the connected component. We choose to represent fragments using the ChainCode Histogram (CCH) described in [8] which is a translation and scale invariant shape descriptor.

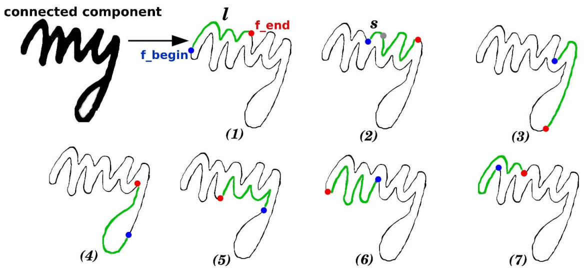

Fig. 2. Fragments extraction on a connected component

Codebook generation: The codebook generation step aims at finding a collection of similar contour fragments in a first learning dataset. In the proposed system, this stage is realized through a 2D Self-Organizing Map (SOM) [5] trained on the chaincode histogram feature vectors. In order to tackle the difficulty of discriminating printed and handwritten text in the presence of both Latin and Arabic script, a set of fragments in four kinds of text (Latin printed, Arabic printed, Latin handwritten and Arabic handwritten) has been extracted on the MAURDOR database. Different sizes of fragments were tested as well as various overlapping parameter values. We have experimentally selected two sizes of fragments : $l = 10$ and $l = 8$ pixels with an overlap of $s = 5$ pixels. The number of fragments extracted for each class is about 200,000 fragments. Several codebooks of different sizes were tested and we chose empirically to use 20×20 codebooks. Finally, we have selected two codebooks (one 20×20 codebook with $l = 10$ pixels fragments and one 20×20 codebook with $l = 8$ pixels fragments) for the MAURDOR campaigns. The connected components are then classified regarding these two codebooks.

B. *Connected component classification*

Once the codebook built, a feature vector is extracted from each connected component of the MAURDOR dataset in Arabic and Latin for both printed and handwritten. For each connected component of this dataset, fragments are extracted and for each fragment of the connected component, the nearest fragment in the codebook is identified using an euclidean distance. Then, the number of occurrences of each codebook fragment in the external contour of the component is computed. This leads to a vector of 400 features. This step was carried out for the two codebooks.

Two MLPs were thus trained (one per codebook) on this dataset containing approximately 200,000 samples of each class (Arabic printed, Latin printed, Arabic handwritten and Latin handwritten). The writing type decision is taken at the connected component level and the result is mapped into two classes : printed and handwritten. The writing type of a connected component is determined regarding the decision of the MLP with the highest confidence (the responses of the MLP are included in a range of $[-1, 1]$, the higher the value of the response, the better the confidence). The final step consists in identifying the writing type of a text area making a majority vote on the decisions taken for the connected components. Experimental results obtained during the MAURDOR campaigns will be presented in section VI.

V. Language identification

We now describe the language identification method. As already said, it is a difficult problem since there are many languages with strong similarities between them. Indeed, there are many alphabet (*i.e.* Arabic, Latin, Japanese, Cyrillic, ...) which are rather easy to discriminate, but some languages globally share the same alphabet, making them difficult to discriminate, such as English and French languages. In this latter example, the small specificities (presence or absence of accentuated character) are not enough to reliably discriminate the shapes based on physical descriptors. Therefore, we have turned toward the use of statistical textual descriptors to discriminate the language. For that, a reliable recognition engine is needed.

The proposed language identification system is made of two sequential steps. The first one discriminates between distinct script alphabets and then, the second step discriminates between languages that use the identified alphabet.

A. *Script alphabet identification*

The first stage is dedicated to the discrimination between script alphabets. Letters are different, but ligatures between characters and words can also be discriminative. Consequently, the aim of this first stage is similar to the writing type identification problem described in section IV. Moreover, the related works presented in section II show that these two tasks are generally made by the same kind of approaches. Therefore, the system presented in the previous section for printed/handwritten letters discrimination has been adapted to perform the script alphabet discrimination.

The alphabet identification system is then similar to the writing type identification system, except for the codebook feature set. Codebooks representing the most frequent fragments of contours in the four kinds of text (Latin printed, Arabic printed, Latin handwritten and Arabic handwritten) are used to characterize the connected components. MLPs are used to classify the connected components into one of the four classes and the results are mapped into Arabic and Latin. Finally, the alphabet of a text region is decided using a majority vote on the alphabet of its connected components.

B. *Language identification*

The method used for discriminating languages sharing the same set of characters is to analyse their character sequences. Some characters are more frequently used depending of the language. For example, the character '*W*' is used in a lot of common words in the English language, whereas there are less than 230 french words (that are not everyday words) containing this character. The same phenomenon can be observed for couples of characters. Moreover, the language analysis literature shows that n-gram analysis are commonly used for electronic document language identification.

Based on this observation, the proposed language identification system relies on the analysis of bi-gram (couple of two characters) of an OCR output. We assume that the frequencies of some particular bi-grams are strong characteristics of a language, even if the recognition engine is not perfect. n-gram with $n > 2$ can be even more discriminative but need to ensure having correct sequences of n characters, that is not guaranteed in our case due to OCR errors. The key idea is to always use the same OCR in order to replicate the same transcriptions errors. For our system, we use the LITIS OCR based on HMM with variable state number, described in [22]. Since the language is unknown during recognition, this OCR is a language free version working at the character level (without any language model and dictionary).

Language profile estimation: To select the appropriate language according to the bi-gram distribution, we need to estimate the language profiles (the distribution of bi-grams for each language). A language profile is estimated by OCRing the content of a document set of this language and estimate the bi-gram frequencies on the resulting transcription. Thanks to the previous printed/handwritten discrimination, we can refine the representation by defining two profiles for each language : a printed profile and a handwritten profile. In the Latin alphabet, we have to discriminate French from English. Hence, we get 4 profiles: French-hand, French-printed, English-hand and English-printed. These profiles are estimated on the documents from the MAURDOR training dataset, respectively made of 375,196; 1,576,820; 132,526 and 1,095,497 bi-grams of characters.

Decision process: First of all, the text content of a document is OCRised by the same OCR used for language profile estimation. Then, the document profile of bi-gram is generated for both handwritten and printed characters. Handwritten document profile is compared with the set of hand-profiles (here, the French-hand and the English-hand) and the printed one, with the set of printed profiles. The profile comparison is made by a weighted χ^2 like score to measure the distance between the document profile Pr_{doc} and the languages ones Pr_{lang}:

$$Score_{lang} = \sum_{b \in Pr_{doc}} \frac{(Pr_{doc}(b) - Pr_{lang}(b))^2}{Pr_{lang}(b)} \times weight(b)$$

The $weight(b)$ is the difference between frequencies of bi-gram b in the French and the English profiles. More generally, this is a coefficient that tries to maximize the contribution of most discriminative bi-grams.

At this moment, a global language decision has to be taken using the handwritten decision and the printed decision. We have chosen to trust the printed decision, unless there is significantly more handwritten content in the document. Indeed, the OCR is more accurate on printed than on handwritten texts. An empirical good hand/printed ratio is 5, obtained by evaluating several values. The complete working scheme is depicted on Fig 3.

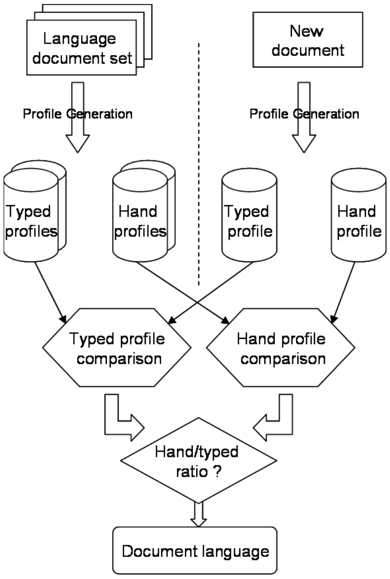

Fig. 3. Overview of the language identification system

VI. Experimental results

The system was evaluated during the first and the second MAURDOR campaigns respectively in March and November 2013. In this section, the MAURDOR dataset is presented, the metrics are described, and the results are exposed.

A. The MAURDOR dataset

The MAURDOR dataset is composed of heterogeneous documents distributed according the following categories :

C1 (12%) : Blank or completed (by hand) forms;
C2 (40%) : Printed business documents (invoice, bill, receipt, catalogue page, newspaper article, contract, legal or administrative document, map, drawing, etc.);
C3 (25%) : Private handwritten correspondence (invitation letter, post-it, etc.);
C4 (20%) : Printed business correspondence (medical receipt, fax header, etc.);
C5 (3%) : Other documents (schemes, plans, tables).

Fonts and handwriting are different across documents and documents are digitized according to different methods. The documents are either in French, Arabic or English but they can occasionally contain text in other languages. The Fig 4 contains some examples of documents and the Fig 5 shows some examples of text regions. The corpus for the second campaign was composed of 6000 training documents and 1000 other documents for the evaluation.

B. The metrics

The metrics proposed by the French National Metrology and Testing Laboratory (LNE) to evaluate the tasks of writing type and language identification are the following. First of all, a reject metric is defined as ”Silence”. The Silence rate is the proportion of text areas that has been rejected by the algorithm (if the algorithm has a rejection ability). Then a classical Precision/Recall measure is used to evaluate the writing type and language identification.

$$Prec = \frac{nb\ correct\ areas}{nb\ areas\ hypothesis}, Rec = \frac{nb\ correct\ areas}{nb\ areas\ groundtruth}$$

$$Sil = \frac{nb\ areas\ rejected}{nb\ areas\ groundtruth}$$

C. The MAURDOR campaign results

In this section, the results of the proposed systems obtained on the documents of the second MAURDOR campaigns are given and

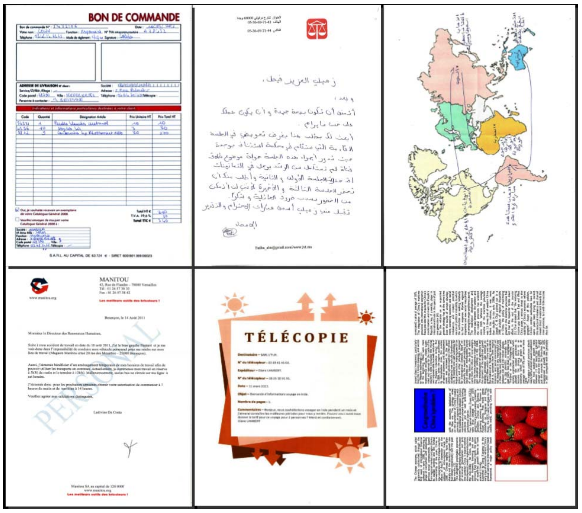

Fig. 4. Example of documents used in the MAURDOR campaigns

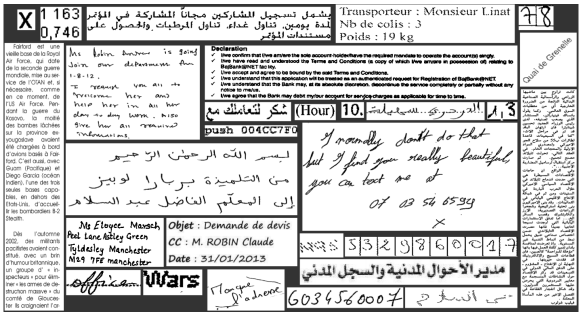

Fig. 5. Example of text areas used in the MAURDOR campaigns

compared with the results of the other participants. These results[1] are given using the Precision, Recall and Silence metrics mentioned above.

1) Results of the writing type identification system: For the writing type identification task, inputs are documents with the position of all text blocks. We evaluate 2 configurations during the second campaign (System A and B) and another configuration were evaluated after the campaign (System C). The distinction between the three systems corresponds to the variation of a threshold used to reject blocks regarding the value of the outputs of the MLP. We have empirically chosen the threshold in order to maximize the precision in the **System A**, while the threshold in the **System B** is fixed in order to reduce the reject. Finally, the **System C** (which was evaluated after the campaign) corresponds to the system without any reject. ”Participant_1” denotes the other participant of the MAURDOR campaign.

Global results on the writing type identification are presented on Table I. Our system A obtains the best precision but rejects more often, reducing its recall performance. However, if we look at the system C, we can notice that without reject our system still achieves better performances than the other campaign participant. Table II proposes a finer analysis by presenting the system performance per writing types. One can see that all the systems achieve better performances on printed writing type.

TABLE I. WRITING TYPE IDENTIFICATION:RESULTS ON THE DOCUMENTS OF THE SECOND CAMPAIGN

System	Precision (%)	Recall (%)	Silence (%)
System A	96.11	85.43	11.12
System B	95.55	86.39	9.58
System C	93.34	93.34	0.0
Participant_1	93.30	93.16	0.15

TABLE II. WRITING TYPE IDENTIFICATION:RESULTS ON THE DOCUMENTS OF THE SECOND CAMPAIGN PER WRITING TYPE

	Printed			Handwritten		
System	P (%)	R (%)	S (%)	P (%)	R (%)	S (%)
System A	96.92	89.85	8.40	93.18	72.10	19.30
System B	96.19	91.13	7.11	93.16	72.09	17.04
System C	95.02	96.16	0.00	88.01	84.82	0.00
Participant_1	94.93	96.15	0.08	88.10	84.14	0.38

2) Results of the language identification system: For the language identification task, inputs are documents with the position of all text blocks and the associated ground truth writing type. We evaluate two configurations during the second campaign, and two others that are evolutions of these systems:

- **System A :** The system described above, made of script alphabet identification (Arabic/Latin) by codebook and language identification (French/English) by bi-gram distributions of Latin OCR output
- **System B :** Script Alphabet identification (Arabic/Latin) and language identification (French/English) are both performed using the bi-gram distributions of Latin OCR output
- **System A+ and B+ :** Evolutions of systems A and B with an improvement of the text line separation algorithm used before the OCR, and some additional preprocessing steps (rule lines removal and correction of the inverse video).

”Participant_1” denotes the other participant of the MAURDOR campaign.

The Tables III and IV present the global and per language performances respectively. The systems A and B outperform the other campaign participant. Our system B was ranked first for this competition. However, we can see that the evolutions made after this campaign increase significantly the results. An interesting fact is that the discrimination between Arabic and Latin alphabet is more accurate by analysing the Latin OCR output on Arabic documents. Even if the recall for Arabic is better with codebook, the best overall performances are obtained only using bi-gram analysis.

TABLE III. LANGUAGE IDENTIFICATION:RESULTS ON THE DOCUMENTS OF THE SECOND CAMPAIGN

System	Precision (%)	Recall (%)	Silence (%)
System A	78.95	71.99	8.97
System B	83.65	83.65	0.00
System A+	80.48	73.46	8.89
System B+	86.78	86.78	0.00
Participant_1	57.88	55.66	4.00

3) Results of the chain: In order to improve the experiment quality, we evaluate the chain made of both the writing type and the language identification described in this paper. Therefore, the difference with respect to the previous results is that the language identification does not benefit from the ground truth writing type, but only from the output of our writing type method. The results

[1] Obtained with the version 1.11 of the evaluation tool supplied by the LNE

TABLE IV. Language identification:Results on the documents of the second campaign per language

	Arabic			English			French		
System	P (%)	R (%)	S (%)	P (%)	R (%)	S (%)	P (%)	R (%)	S (%)
System A	58.42	96.03	2.34	92.18	56.17	10.89	88.97	70.17	10.20
System B	75.64	86.92	0.00	85.04	58.47	0.00	86.10	92.37	0.00
System A+	62.76	95.40	3.09	93.32	51.32	10.70	87.61	74.70	9.90
System B+	70.75	91.46	0.00	86.21	75.27	0.00	94.02	89.88	0.00
Participant_1	29.24	4.96	3.42	25.00	0.05	4.53	58.90	93.16	4.00

given in Tables V and VI show the robustness of the system B+ for handwritten/printed misclassification. The drop of precision for the system A is induced by the precision decrease of the Arabic identification.

TABLE V. Writing type + Language identification:Results on the documents of the second campaign

System	Precision (%)	Recall (%)	Silence (%)
System A+	77.58	70.74	8.86
System B+	86.22	86.22	0.00

TABLE VI. Writing type + Language identification:Results on the documents of the second campaign per language

	Arabic			English			French		
System	P (%)	R (%)	S (%)	P (%)	R (%)	S (%)	P (%)	R (%)	S (%)
System A+	58.42	96.03	2.34	92.17	50.20	10.89	86.45	70.30	10.20
System B+	70.10	91.80	0.00	86.32	72.50	0.00	93.30	89.86	0.00

VII. Discussion and future work

This paper has presented a system for writing type and language detection in heterogeneous and complex documents. Writing type is identified thanks to an original set of physical codebooks classified by a MLP. The language identification is split into two steps : the script alphabet detection and the language identification step. The first step is made with a similar codebook/MLP system and the second relies on statistical analysis of bi-grams in an OCR output. The results obtained on the MAURDOR dataset compare favorably our systems to the other participants. Even without reject, the writing type identification is 93.34% accurate and the best language identification system relies on two-stages bi-gram analysis and achieves a precision rate of 86.78%.

Although efficient, our writing type identification system can be improved adding a preprocessing step in order to correct the inverse video and improve the quality of the contour fragments. Our system also seems weaker on the identification of the handwritten writing type. The small lengths of fragments used in the codebooks ($l = 8, 10$ pixels) seems more efficient on the printed writing type identification. Another way of improvement could be to build expert codebooks : one codebook dedicated to the printed writing type with small fragments and a second codebook specialized on the handwritten writing type with longer fragments more adapted to the shape of the handwritten writing.

In the language identification system, we use an OCR at character level, that is the hardest way for text transcription. An alternative approach can process the OCR with both French and English language models and compares recognition scores to choose the correct language. A similar scheme can also be used as a correction step, choosing the best language model of the two or three more likely languages according to the bi-gram distributions. Finally, our systems need to be evaluated on databases with more script alphabets and more languages.

Acknowledgment

The authors would like to thank the DGA (French Defense procurement agency) for financing the MAURDOR campaigns and AIRBUS Defence and Space for the coordination.

References

[1] Alex Graves, Marcus Liwicki, S. Fernandez, Roman Bertolami, Horst Bunke, Jrgen Schmidhuber: A Novel Connectionist System for Unconstrained Handwriting Recognition. IEEE Trans. Pattern Anal. Mach. Intell. 31(5): 855-868 (2009)

[2] P.Barlas, S. Adam, C. Chatelain, and T. Paquet, ”A typed and handwritten text block segmentation system for heterogeneous and complex documents”, to be published in Document Analysis Systems, 2014.

[3] T. Kasar, P. Barlas, S. Adam, C. Chatelain, and T. Paquet, ”Learning to Detect Tables in Scanned Document Images using Line Information,” in ICDAR, 2013.

[4] E. Augustin, M. Carr, E. Grosicki, J. M. Brodin, E. Geoffrois and F. Preteux, *RIMES evaluation campaign for handwritten mail processing*, In Proceedings of the Workshop on Frontiers in Handwriting Recognition, 2006, 231–235

[5] M. Bulacu and L. Schomaker, *A comparison of clustering methods for writer identification and verification*, In Proceedings of the Eighth International Conference on Document Analysis and Recognition,2005, 1275–1279

[6] L. Schomaker, K Franke and M Bulacu, *Using codebooks of fragmented connected-component contours in forensic and historic writer identification*, Pattern Recognition Letters, Volume 28, 2007, 719–727

[7] G. Ghiasi and R.W. Daly, *An efficient method for offline text independent writer identification*, Pattern Recognition (ICPR),IEEE,2010,1245–1248

[8] J. Iivarinen and A. Visa, *Shape Recognition of Irregular Objects*, Intelligent Robots and Computer Vision XV: Algorithms, Techniques, Active Vision, and Materials Handling, Proc. SPIE 2904,1996,25–32

[9] MAURDOR campaign website, *http://www.maurdor-campaign.org/*

[10] J. Ohya, A. Shio and S. Akamatsu, *Recognizing characters in scene images*, IEEE Trans. Pattern Anal. Mach. Intell., 1994 , 214–224

[11] H. Hase, T. Shinokawa, M. Yoneda, C.Y. Suen, *Character string extraction from color documents*,Pattern Recognition,Volume 34, 2001, 1349–1365

[12] V. Wu, R. Manmatha, E.M. Riseman, *TextFinderan automatic system to detect and recognize text in images*, IEEE Trans. Pattern Anal. Mach. Intell., 1999, 1224–1229

[13] B. Gatos, N. Stamatopoulos, G. Louloudis, *ICDAR 2009 Handwriting Segmentation Contest*, ICDAR, 2009, 1393–1397

[14] W. B. Cavnar, J. M. Trenkle, *N-Gram-Based Text Categorization*, SDAIR, 1994, 161–175

[15] G. Grefenstette, *Comparing two language identification schemes*, JADT, 1995

[16] T. Dunning, *Statistical Identification of Language*, Techreport, 1994

[17] L. Grothe, E. W. De Luca, and A. Nrnberger, *A comparative study on language identification methods*, LREC, 2008

[18] R. Řehůřek, M. Kolkus, *Language identification on the web: extending the dictionary method*, CICLing, 2009, 357–368

[19] J. Kumar, R. Prasad, H. Cao, W. Abd-Almageed, D. Doermann, P. Natarajan, *Shape codebook based handwritten and machine printed text zone extraction*, IS&T/SPIE Electronic Imaging, 2011, 787406–787406

[20] J. Hochberg, K. Bowers, M. Cannon, P. Kelly, *Script and language identification for handwritten document images*, IJDAR, 1999, 45-52

[21] Y. Zheng, H. Li, D. Doermann. *Machine Printed Text and Handwriting Identification in Noisy Document Images*, IEEE Transactions on Pattern Analysis and Machine Intelligence, March 2004, 337–353

[22] K. Ait Mohand, T. Paquet and N. Ragot, *Combining structure and parameter adaptation of HMMs for printed text recognition*, to be published in IEEE PAMI, 2014