FIRST MAURDOR 2013 EVALUATION CAMPAIGN IN SCANNED DOCUMENT IMAGE PROCESSING

Ilya Oparin, Juliette Kahn and Olivier Galibert

LNE, Laboratoire national de métrologie et d'essais
National Metrology and Testing Laboratory, France
firstname.secondname@lne.fr

ABSTRACT

This paper presents the results of the first Maurdor evaluation campaign. This campaign aims at evaluating the complete chain of scanned document image processing. It has a modular structure that includes page segmentation and zone classification, identification of writing type and language, optical character recognition and revealing logical structure of a document. This campaign is based on a unique corpus of 8,000 images of scanned documents annotated at different levels. Presentation of the results of the first campaign is important to assess the state-of-the-art and create common references both for participants in future campaigns and, as the scoring tools are publicly available, for independent tests.

***Index Terms*—** Evaluation, scanned document image processing, page segmentation, OCR

1. INTRODUCTION

Processing of scanned document images is an important issue for information retrieval. The French National Metrology and Testing Laboratory (LNE, Laboratoire national de métrologie et d'essais) conducts the Maurdor (**M**oyens **AU**tomatisés de **R**econnaissance de **DO**cuments éc**R**its) evaluation campaigns in order to support research and help advancing the state-of-the-art in the domain of scanned documents processing.

The specificity of the Maurdor evaluations is that they are based on a complete chain of scanned documents processing in which five separate tasks are implemented. Each task corresponds to a particular function and contributes to the complete processing of scanned document images.

- **Task 1:** Page segmentation and zone classification;
- **Task 2:** Identification of writing type (handwritten or printed);
- **Task 3:** Language identification;
- **Task 4:** Optical character recognition (OCR);
- **Task 5:** Extraction of logical structure;
- **End-to-end:** Complete evaluation of all the tasks from 1 through 5.

A lot of research and evaluations were carried out for most of these tasks: page segmentation and classification [1, 2, 3, 4], printing type recognition [5], language identification [6] or OCR [7, 8]. There is also a number of important competitions held within ICDAR and ICFHR conferences. But the existing evaluations were done in a rather isolated way. The Maurdor evaluations aim at the more challenging task of evaluating the complete processing chain and provide corresponding training, development and test data. At the same time the evaluations are designed in a modular way, participants can choose to tackle only certain tasks and each technology module can also be evaluated independently.

A new metric called ZoneMap has been proposed and implemented to simultaneously evaluate both page segmentation and zone classification [9]. As compared to other metrics, is is characterized by extended functionality (e.g. it takes overlapping zones into account), stable behaviour in different conditions and flexibility.

Scoring metrics for each of the tasks and for the end-to-end evaluation are integrated into the LNE scoring toolkit that is freely distributed under the GPL license. Availability of this toolkit and a large and heterogeneous corpus of annotated document images not only allows assessing the state-of-the-art in processing of scanned documents on different levels but also permits to create a common reference in the domain.

2. MAURDOR CORPUS

The Maurdor evaluation corpus consists of scanned document images annotated on different levels. This corpus was created by ELDA and will be distributed through the ELRA (`www.elra.info`) catalogue under fair licensing conditions once the Maurdor campaigns are finished. Scanned documents belong to one of the following categories:

- **C1:** Blank or completed (by hand) forms;
- **C2:** Printed, but also manually annotated business documents (invoice, bill, catalogue page, etc.);
- **C3:** Private handwritten correspondence, sometimes with printed letterheads;
- **C4:** Printed, but also manually annotated business correspondence (handwritten mail, fax header, etc.);
- **C5:** Other documents such as newspaper articles, plans, schemes, drawings, etc.

Fonts and handwriting are different across documents and documents were digitized according to different methods in

The Maurdor evaluation campaigns (www.maurdor-campaign.org) are part of the Maurdor project, managed by Cassidian and funded by DGA.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

order to obtain images with different characteristics. The documents are either in French, Arabic or English but may occasionally contain text in other languages.

The corpus is annotated at different levels in order to evaluate each of the tasks introduced in Section 1. For Task 1 the annotation includes coordinates of polygons corresponding to the different zones in a document image together with their types (text area, logo, signature etc.). Writing type (print/hand) and language are specified for texts zones for Tasks 2 and 3. Textual transcription is provided for Task 4 to evaluate OCR technologies. For Task 5, information on reading order and semantic roles of different zones is presented in the annotation.

About 8,000 documents have been produced by ELDA for two first Maurdor campaigns. 5,000 documents were used in the first Maurdor evaluation campaign with 3,000 documents in the training set and 1,000 documents in the development and test sets. Participants were allowed to use the development data in training. All the corpora (train, dev and test) are homogeneous according to document categories, number of words and number of text zones per document.

3. EVALUATION TASKS

3.1. Zone Segmentation and Classification

Zone segmentation and classification consists in extracting using closed polygonal-shaped outlines homogeneous semantic areas from the document images. Document zones may have different natures and can be classified into one of the following categories:

- Writing (text) area;
- Photographic image area;
- Line drawing area;
- Graphic area;
- Table area;
- Separator area;
- Damaged/undefined/unspecified area;

A graphic area is further classified into a sub-type, such as logo, diagram, figure, signature, etc.

The aim of this task is to identify the various areas in a document image and specify their position. Different semantic areas may overlap. For instance, a table area may include text and graphic areas (logo, signature, etc.).

The tasks of document page segmentation and classification have existed for several decades and a number of metrics and evaluation schemes were proposed [1, 2, 3, 4, 10]. However, they are not sufficient for evaluating the task of document segmentation and classification in the Maurdor campaigns. A new metric called ZoneMap has been proposed and implemented in the LNE toolkit [9]. It is designed to evaluate both page segmentation and zone classification. Moreover, for the segmentation sub-task it takes into account the superposition of overlapping zones. These characteristics allow to evaluate the task in question in a coherent way using a single ZoneMap metric. Weights assigned to different parameters add additional flexibility to ZoneMap and allow for fine-tuning of the metric in order to reflect the specificity of a particular task to evaluate. ZoneMap is used as a primary metric for zone segmentation and classification. It is compared to the Jaccard score that is used as a secondary metric.

Table 1. Scores ZoneMap and Jaccard for primary systems submitted to the first Maurdor evaluation campaign

System	ZoneMap			Jaccard
	$\alpha_c = 0$	$\alpha_c = 0.5$	$\alpha_c = 1$	
S1	90.0	107.1	124.1	0.150
S2	60.1	75.9	91.8	**0.315**
S3	**31.2**	**57.3**	83.4	0.190
S5	52.2	62.4	**72.7**	0.287

According to the rules of the campaign the participants should be anonymized and the primary submissions of different participants are thus referenced as *S1*, *S2*, etc.

The performance of the primary systems of the different participants in the first Maurdor evaluation campaign is presented in Table 1. ZoneMap results are presented for different decomposition/classification importance ratios. The classification error weight $\alpha_c = 0$ corresponds to the case when no classification error is taken into account (the score is entirely based on decomposition errors) and vice versa for $\alpha_c = 1$. The row $\alpha_c = 0.5$ corresponds to the primary setup according to which errors in zone segmentation and classification are considered equally important. It should be noted that a ZoneMap score can exceed 100% if a submission includes a large number of false alarms.

The ZoneMap score for all systems depending on the decomposition/classification importance ratio (α_c in Table 1) is presented in Figure 1. As the ZoneMap score is a linear combination of segmentation and classification errors, it changes linearly according to the classification error weight α_c. As different systems exhibit different behaviour, an operating point may be chosen depending on the aims of a particular evaluation (0.5 for the Maurdor evaluations). One can see that the submission *S3* is the best in zone decomposition (see the lefmost part of Figure 1 that corresponds to $\alpha_c = 0$) while *S5* is better in zone classification as for higher values of α_c it obtains better scores than *S3*. The possibility of performing such an analysis and focusing on particular operating points is an advantage of ZoneMap over Jaccard as the latter does not take the decomposition error into account.

Another important feature of ZoneMap is its behavorial stability in its whole range of values. As it was shown in [9], a ZoneMap score does not change when the hypothesis changes in a balanced way while a Jaccard score not only increases when the hypothesis is changed in a balanced way but also the amount of the increase varies depending on the original system performance. Thus measuring the impact of small changes during system development is much more difficult with the Jaccard score than with ZoneMap.

Table 2 represents the three most confused zone types for each system with a ratio of the global error associated to each confusion. This information may be important to better understand the nature of the classification errors of the different areas detected in a scanned document. Most confusions are similar across different systems. For example graphic re-

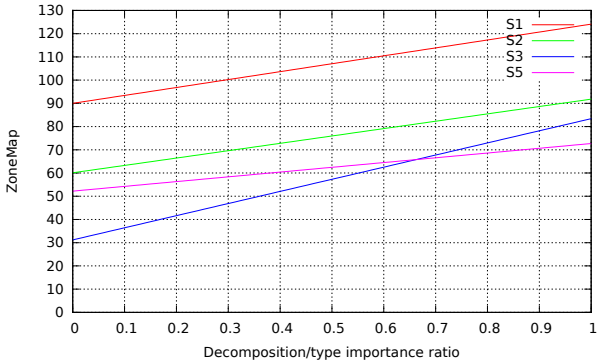

Fig. 1. Zonemap score for all systems depending on the decomposition/typing importance ratio

Table 2. Confusions of different zone types

System	Miss	Reference	Hypothesis	Error	Correct
S1	51.5%	graphic	table	13.4%	19.3%
		graphic	text	3.7%	
		text	table	2.6%	
S2	31.4%	graphic	text	6.2%	46.3%
		graphic	table	6.1%	
		image	graphic	5.4%	
S3	22.2%	image	graphic	6.3%	57.1%
		graphic	text	5.6%	
		text	graphic	3.3%	
S5	46.6%	graphic	text	15.2%	29.7%
		image	text	4.1%	
		graphic	table	3.3%	

gions are often misclassified as text or table regions as they appear in the top three lists for most of the systems. This points out to the fact that the zones types that are difficult to distinguish seem to be system independent. Column *Miss* in Table 2 gives the percentage of zones in the reference with no type assigned. This corresponds to zones that were not detected at page segmentation level and thus to segmentation errors (without taking false alarms into account). *Correct* corresponds to the percentage of correctly detected zone types.

3.2. Identification of writing type

Identification of writing type consists in determining the type of writing used in text areas: handwritten or printed. The identification of writing type is evaluated by means of precision and recall. In the general case precision coincides with recall as the information on the total number of text zones is available to the participants. Table 3 presents results in general and also according to different printing types in terms of precision (*P*), recall (*R*) and F-measure (*F-m*). One can see that performance slightly varies according to different printing types across two submissions.

Table 3. Task 2 results for different printing types

System	Printed			Handwritten			Global
	P	R	F-m	P	R	F-m	Precision
S2	92.4	**95.6**	**94.0**	**83.1**	73.3	77.9	**90.6**
S5	**94.0**	92.6	93.3	78.9	**82.3**	**80.6**	90.0

Table 4. Task 2 precision for different languages

System	English	Arabic	French	Other
S2	**86.4**	89.3	**93.3**	90.2
S5	85.9	**93.0**	90.8	**96.2**

Table 5. Task 2 precision for different document categories

System	C1	C2	C3	C4	C5
S2	92.6	**88.7**	**92.3**	**91.1**	**86.5**
S5	**93.1**	88.4	90.5	89.3	81.7

Results are also presented separately for different languages and document categories in Tables 4 and 5.

3.3. Language identification

Language identification consists in determining the language used in each text area. Languages to be identified are French (*FR*), English (*EN*) and Arabic (*AR*). Metrics used for this task are the same as for Task 2 and the results are presented in Tables 6, 7 and 8.

Table 6. Task 3 precision for different printing types

System	Printed	Handwritten	Global
S4	35.5	49.0	38.9
S5	**64.5**	**61.7**	**63.8**

Table 7. Task 3 precision for different document categories

System	C1	C2	C3	C4	C5
S4	42.7	34.4	49.8	47.3	27.2
S5	**71.7**	**53.2**	**62.3**	**60.5**	**65.2**

Table 8. Task 3 precision for different languages

System	Precision			Recall			F-m		
	EN	AR	FR	EN	AR	FR	EN	AR	FR
S2	**41.7**	28.7	58.8	**27.5**	69.7	30.9	**33.2**	40.7	40.5
S5	-	**75.5**	**60.4**	0.0	**73.6**	**94.0**	-	**74.5**	**73.5**

3.4. Optical character recognition

This task consists in transcribing contents of each text area and is measured by means of the Word Error Rate (WER) and the Character Error Rate (CER). It should be noted that no line segmentation was provided. The participants thus faced a challenging real-life task of handling entire portions of text that can contain one or several paragraphs rather than performing conventional OCR on presegmented lines.

For the best overall system, the influence of the line segmentation on the WER was measured by t.test [11]. The WERs per zone were calculated. The significance of the difference of WER for the zones comprising a single line and the WER for the zones comprising more than one line was estimated. For French (handwritten and printed), the WERs for single line and multiline zones were found to be significantly different ($p < 0.01$ and $p < 0.001$ respectively), while this was not the case for English handwritten zones ($p = 0.53$).

Tables 9 and 10 present the CER and WER results in general and separately for different writing types and languages. For the participant *S6* that submitted system outputs only for

Table 9. Detailed CER results for Task 4

System	Printed				Handwritten				Global
	AR	FR	EN	All	AR	FR	EN	All	
S1	**39.8**	17.1	25.5	22.7	**31.9**	41.1	32.5	**37.1**	**24.4**
S3	54.7	**11.4**	**13.2**	**16.8**	-	75.5	84.6	84.0	24.8
S4	98.2	79.3	88.0	84.6	112.6	102.0	124.7	108.9	87.5
S5	65.4	27.7	29.2	32.4	82.3	67.7	85.8	74.9	37.5
S6_1	-	-	-	-	-	24.1	22.0	45.1	-
S6_2	-	-	-	-	-	**20.8**	**20.0**	42.9	-

Table 10. Detailed WER results for Task 4

Syst	Printed				Handwritten				Global
	AR	FR	EN	All	AR	FR	EN	All	
S1	**58.3**	31.0	39.2	37.1	**58.0**	71.7	59.1	65.5	40.6
S3	91.3	**21.0**	**20.8**	**28.9**	-	98.6	103.4	99.9	**37.8**
S4	161.3	141.5	160.4	150.6	149.3	173.8	201.8	172.8	153.4
S5	112.4	63.9	66.7	70.4	101.0	98.1	119.2	103.0	74.5
S6_1	-	-	-	-	-	39.4	41.6	56.2	-
S6_2	-	-	-	-	-	**34.5**	**38.0**	**52.8**	-

the recognition of handwritten text in French and English, we present not only results with its primary system (*S6_1*) but also those with a secondary system (*S6_2*) as these numbers are important to define the state-of-the-art for this task.

3.5. Extraction of logical structure

Extracting the logical structure of the document consists in determining semantic groupings of text zones (for instance, the connection between an image and the text area in the caption associated with it), a reading order for various areas (for example, a sequence of columns in a newspaper) and a semantic classification of the role of text zones.

Three scores are computed, one for each aspect (group, order, type). They follow the same structure: first a per-zone score in $[0..1]$ is extablished, where 1 is perfect. For grouping the sets of zones in the group are compared between reference and hypothesis, using singletons when not in a group. The F-measure of the coverage ratios is used as a metric. For reading order the previous zone identify (if any) is compared, yielding 1 in case of equality and 0 otherwise. Finally for the semantic classification the lists of types are compared and the F-measure between precision and recall is measured.

Once all the per-zone scores are computed a first mean is computed per-document to get the triplet of document scores, which are then averaged together to obtain the raw collection scores (S_r). A problem with such a score is that the annotations are extremely sparse, giving high scores even when doing nothing. In order to normalize the result the scores of an empty hypothesis (S_0) are also computed, and the final score S is created in a linear fashion by setting the "all wrong" score at -100, the "empty hypothesis" at 0 and the "perfect" at 100:

$$S = 100 \times \begin{cases} \frac{S_r - S_0}{S_0} & \text{if } S_r < S_0 \\ \frac{S_r - S_0}{1 - S_0} & \text{if } S_r \geq S_0 \end{cases}$$

Table 11. Global results for Task 5

System	Type	Order	Group
S2	10	2	26
S3	**59**	**22**	27
S5	42	17	**37**

For each aspect the final score in then positive if a system adds more correct information than errors.

3.6. End-to-End

The end-to-end task consists in evaluating the complete chain of document image processing, i.e. performing tasks 1 to 5 consecutively. The evaluation is based on the detection of the presence of words considered as queries to a search engine.

First a list of words of interest is defined. Once this list is obtained, the detection quality is measured by calculating the presence of these words in the documents. In order to take into account the general logic of the system the information on zone classification from Task 1 and the information on the logical structure from Task 5 is added to the list of words as pseudo-words. Thus it is possible to evaluate the searches like "all the documents with logos and legends".

Two metrics are used to evaluate this task. The first one is the classical cosine distance used in information retrieval. It gives the score between 0 and 1 based on word occurrence in documents. The second one is a metric of utility in which every found word is counted as +1, an extra word as -1 and a not-found word as 0. The sum of scores is divided by the number of words to find. The minimal score is fixed at -1 so that a particularly badly scored document cannot dominate the global results. The document score is thus between -1 and 1. Only *S2* participated in the first end-to-end evaluation and obtained cosine and utility scores of 0.4593 and 0.0909 respectively.

4. CONCLUSIONS

The results of the first Maurdor campaign were presented in this paper. To our knowledge, this is the first campaign that aims at evaluating a complete chain of scanned document image processing. This evaluation has a modular structure with separate evaluations of the subtasks.

A new metric called ZoneMap was developed and validated in order to evaluate page segmentation and zone classification in scanned document images. This matric, together with the metrics implemented for other tasks within this campaign is available in the LNE *maurdor-eval* toolkit, distributed under the GPL license. A unique corpus of scanned document images annotated at different levels was also developed. We expect that this corpus and the evaluation toolkit will be used in research community to evaluate different aspects of automatic processing of scanned documents as a common reference. With this aim in mind we presented the results of the first campaign in order to introduce the possibilities Maurdor data and tools enable to evaluate image processing technologies and to assess the state-of-the-art.

5. REFERENCES

[1] S. Mao and T. Kanungo, "Empirical performance evaluation methodology and its application to page segmentation algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, pp. 242–256, 2001.

[2] B.A. Yanikoglu and L. Vincent, "Pink Panther: A complete environment for ground-truthing and benchmarking document page segmentation," *Pattern Recognition*, vol. 31, pp. 1191–1204, 1998.

[3] T. Pavlidis and J. Zhou, "Page segmentation and classification," *CVGIP: Graphical Models and Image Processing*, vol. 54, no. 6, pp. 484 – 496, 1992.

[4] R. Haralick Y. Wang and I. Phillips, "Zone content classification and its performance evaluation," in *Pattern Recognition*, 2001, pp. 540–544.

[5] H. Li Y. Zheng and D. Doermann, "Machine printed text and handwriting identification in noisy document images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, pp. 337–353, 2004.

[6] T. Baldwin and M. Lui, "Language identification: The long and the short of the matter," in *Proc. of ACL'10*. 2010, pp. 229–237, Association for Computational Linguistics.

[7] H. Bunke and P. Wang, *Handbook of character recognition and document image analysis*, World Scientific, 1997.

[8] "OpenHaRT evaluation campaigns," `http://www.nist.gov/itl/iad/mig/hart.cfm/`, [Online; accessed Feb 2014].

[9] J. Kahn, O. Galibert, and I. Oparin, "The ZoneMap metric for page segmentation and area classification in scanned documents," in *Submitted to the XXI IEEE International Conference on Image Processing ICIP'14*, 2014.

[10] J. Kanai, S. V. Rice, T. A. Nartker, and G. Nagy., "Automated evaluation of OCR zoning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, pp. 86–90, 1995.

[11] E. S. Pearson, "Student as a statistician," *Biometrika*, vol. 30, pp. 210–250, 1939.