Institute Document Automization System

1. ***Shinde Sandip, [2].Jape Prajakta, [3].Shinde Prajakta, [4].Nagapalle Pritee,***
2. ***Guidance: Prof. Bhise A.K***
3. ***E-mail: Sandipbshinde9881@gmail.com, 2.prajktabshinde@gmail.com, 3.priteenagapalle@gmail.com***

Final Year Students of Department of Computer Engineering

***Abstract*— This paper presents the achievements of an experimental project called Maurdor (Moyens AUtomatisés de Reconnaissance de Documents ecRits – Automatic Processing of Digital Documents) funded by the French DGA that aims at improving processing technologies for handwritten and typewritten documents in French, English and Arabic. The first part describes the context and objectives of the project. The second part deals with the challenge of creating a realistic corpus of 10,000 annotated documents to support the efficient development and evaluation of processing modules. The third part presents the organisation, metric definition and results of the Maurdor International evaluation campaign. The last part presents the Maurdor demonstrator with a functional and technical perspective.**

I. CONTEXT, CHALLENGES AND GOALS

The automatic processing of numeric documents has been an active research topic for several years now The work presented in this paper is motivated by the strong demand of applications: digital library, documentary heritage, incoming mails, mail sorting … The growing data stream to be processed requires a productivity improvement and some advanced tools to perform the mass analysis of such documents: search engines, named entity extractors or various kinds of tools to support human analysis (graph viewers for relationships analysis, statistical modules, etc.) The data to be processed consists of digital images coming from paper document digitization. The textual information is obtained by the image conversion to text thanks to software techniques called Optical Character Recognition (OCR). The accuracy of the results is variable depending on the script type (hand or typed writing), the digitization and paper quality. Different corpora had been developed to assess and improve the Optical Character Recognition (OCR) systems since twenty years. Some campaigns focused on very basic tasks as numbers or name recognition with line delimitation [1], [2], [3], [4]. The majority of the campaigns focused on one language (English [5], Arabic [2] or French [3],[4]) and one type of writing (handwritten [2], [4] or typed [5]). In the last years, OCR is included in more difficult tasks as Information Retrieval [6], medical image annotation [7] or Speaker identification in broadcast new videos [8] The goal of the Maurdor project is to represent a major technical and technological breakthrough regarding the processing of digitized and faxed documents with particular emphasis on efficient OCR solutions for hand and typed written documents. The project includes studies and development of software modules in the field of automatic processing of written faxed or digitized documents, in addition to the production of a large corpus to sustain such activities. These modules have been defined to realize a complete OCR and data extraction chain. These technological developments are validated through an open and international evaluation campaign. Each module is unitary assessed with the corpus of documents specially made for the project needs. The results of these studies and the implemented modules are integrated in a demonstrator that is delivered in two incremental versions.

The challenges to be undertaken are the following:

_ Increase the overall performance of techniques for automatic processing of multilingual written documents (document segmentation, recognition of handwritten documents, logos and signatures identification, etc.);

_ Develop linguistic resources (corpora, dictionaries) that are required for this work;

_ Organize technical evaluations of document processing modules;

_ Validate the automation of the document processing chain within a demonstrator based on an open, modular and scalable architecture where the results can be displayed, edited and exploited (indexing, information extraction, semantic analysis, etc.) using mainly tools and components off the shelves.

Automatic processing of written documents consists of the six following modules:

_ Locating all information areas in a digitized image and labelling them according to their type, in particular by setting apart written areas (module 1);

_ Identifying whether the text contained in written areas is typed or handwritten (module 2);

_ Identifying the language(s) of the text contained in the various written areas (module 3);

_ Performing writing recognition, i.e. transforming the image of the text contained in written areas into **editable text (module 4);**

_ Determining logical connections between different areas of the document (reading order) and labelling written areas with semantic annotations (i.e. title, object, legend, etc.) (module 5);

_ Document indexing using metadata elements for keyword searching and retrieval (module 6).

II. PRODUCTION OF A REALISTIC CORPUS

The target within the project is to collect at least 10000 documents, out of which 8000 are used and released to the community while 2000 are kept for further testing and comparing the modules. Such documents consist of scans of hardcopy documents, representatives of human daily operations, both in terms of contents, languages, and formats.

The main characteristics of the corpus are:

_ Multi linguality: documents are drafted in French (50% of the whole set), English (25%) and Arabic (25%) by native/almost-native scribes (writers); _ The scripts are handwritten, typed characters or a mix of both. Images are also part of the documents (photograph, drawings, etc.);

_ 1553 scribes contributed to the corpus to ensure enough variety. Due to the project context, a substantial part of the corpus was produced from scratch instead of compiling existing documents: hence covering all the specifications stated by the project funder, while also avoiding any infringements on Intellectual Property Rights.

A. Specification of the corpus

Due to the project requirements, it was decided to produce a new training and evaluation corpora.
Nevertheless, participants to this evaluation were allowed to use other resources available within the community. For instance, the participants could use the RIMES1 database [9] which consists of letters composed by volunteers, with a free layout, over 1,300 people contributed to the database by writing up to 5 letters. The RIMES database contains about 12,723 pages.
IAM database was also used [10]. The IAM Handwriting Database contains forms of unconstrained handwritten English text, which were scanned at a resolution of 300dpi and saved as PNG images with 256 gray levels. The IAM Handwriting Database 3.0 consists of 657 writers who contributed samples of their handwriting (1539 pages of scanned text). OpenHaRT (available from LDC as [11]) and exploited within the various NIST evaluations series on "the document recognition and translation" series [12], tackles issues related to Open Handwriting Recognition and Translation of collection of annotated naturally-occurring examples of handwriting in multiple languages, genres and domains (mostly developed within the DARPA MADCAT program).
1 Reconnaissance et Indexation de données Manuscrites et de facsimilÉS / Recognition and Indexing of handwritten documents and faxes.) Another important database is APTI [13] that focused on evaluation protocols for Arabic Printed Text Recognition. This database was *"synthetically generated using a lexicon of 113'284 words, 10 Arabic fonts, 10 font sizes and 4 fonttyles"*. During the first phase, the work focused on specifying the nature of the documents. In order to keep the corpus coherent and consistent with the objectives, each scribe was requested to produce at most 10 documents complying with pre-designed models and instructions (called scenarios, to better mimic realistic use-cases). For instance, an "invoice" model comes with a scenario such as "you are invoicing your customer for a number of commodities, please use this template and fill it in handwriting. To achieve this, 1000m models were to be produced with the associated scenarios so as to generate documents complying with the five following gross categories:
_ C1: Printed forms (to be filled in handwriting)(12%)
_ C2: Commercial documents (quotations, orders, invoices, product factsheets, leaflets, newspaper articles, etc.)(40%)
_ C3: Private manuscripts correspondences (25%)
_ C4: Private or professional typed correspondences (20%)
_ C5: Other (diagrams, drawings (freehand, maps)) (3%)
To enhance the realistic aspects, other characteristics were also added as part of the requirements (e.g. documents had to bear logos, signatures, noisy parts, mix of scripts and languages, etc.) During this phase, the ELDA project team selected GEDI (Groundtruthing Environment for Document Images) [14] as annotation platform. Another task conducted during the specification phase was related to the quality assessment of the corpus and aimed at establishing a set of criteria to be met by the scans of the documents as well as their annotations, including thresholds of acceptable error rates for the human annotations.

B. Production of the corpus

In order to create a corpus that complies with the specifications, the team worked out a first set of models based on the compilation of realistic documents (such as real letters, invoices, quotations, forms, leaflets, newspaper articles, etc.). A pilot collection helped validating the whole production procedure. During the project over 1446 models were produced. Following this feasibility phase, ELDA recruited scribes, taking into consideration the different requirements. For instance looking for native/almost-native scribes in Arabic, English, and French imposed that data should be collected in several countries i.e. France, USA, India, Morocco, Lebanon, etc. During the recruitment, each scribe was given clear instructions about the expectations concerning the document characteristics (use of logos, signatures, ink stamps, sending via fax/scan at various resolution quality, etc.). A Web site was also designed and set up to facilitate the registrations, assignment of tasks to scribes (instruction, download of models/scenarios, and upload of documents).
350

After receiving each document, it underwent verification and an acceptance procedure that consisted of checking the scan (digital image) to ensure that it complies with the digitalization requirements, with the corresponding model, scenario, etc. A collection management database was updated to keep track of the production progress. Documents that were accepted were included in one of the sets and sent to annotation.

C. Annotation of the corpus

Documents were clustered according to their models and complexity (about 10 docs) and each cluster assigned to one annotator. Once this first annotation completed, a second annotator carried a cross-verification. Revisions were immediately implemented if the annotators agreed. In case of disagreement, a senior annotator opinion was requested. An annotation manual with adopted conventions tackles aspects such as tagging of semantic zones, transcription of textual sections, logical relations between zones, extraction of specific information i.e. Meta-data, etc.and associate to each identified element a tag such as:
_ Writing (text) area;
_ Photographic image area;
_ Line drawing area;
_ Graphic area and subtypes (logo, diagram or figure, stamp, signature, comb field or sequence of identical boxes, free text field with frame or guide line, line drawing, etc.);
_ Table area;
_ Separator area;
_ Noise area;

_ Unspecified or undefined area (rejection)
Further to this annotation step, annotated documents were sent to the validation team for quality check.

D. Validation and quality check

The validation task consisted firstly of an exhaustive and automatic control and secondly of a human control of a random sample. The automatic control was based on a number of scripts capable of detecting missing items and inconsistencies. For instance, a zone that is labelled as graphic zone must have an attributed function (a logo, chart/diagram, signature, stamp, form-field, etc.). Other scripts targeted metadata elements (e.g. a fax document could have a sender and a receiver and if such fields are empty, it has to be double checked). At the end of this procedure, the whole set of documents for which scripts have revealed errors or inconsistencies are sent back to the annotation team for revision. The control team also drew a typology of errors for which new scripts were developed. This process was conducted iteratively to ensure that all errors, automatically detected, were corrected. At that stage, each set of about 500 documents underwent a human quality control based on an exhaustive analysis of a sample of a randomly selected documents covering at least 5% of the agreed upon labels and meta-data elements. If for one of the parameters the error rate (1% to 3% depending on the error type) was above the agreed threshold, a new revision of the whole set was launched for that type of error and the process was re-iterated till all error rates were below their respective threshold. The whole corpus should be made available to the community through the ELRA [15] catalogue under fair licensing conditions, to be announced after the project end.

II. THEMAURDOR EVALUATION CAMPAIGN

The collected corpus was used to carry out two evaluation campaigns. The first one exploited 3000 docs for training, 1000 for development and 1000 for testing. The second evaluation campaign exploited the additional sets: extra 3000 docs (train, validation and test). These campaigns were open to any person, institution or company. Their goal was to quantify the ability of existing systems to extract relevant information in scanned documents.
The following was provided to participants:
_ Consistent data for training, development and test;
_ Automatic scoring tools;
_ Common rules for assessment of different steps essential for scanned documents processing.

A. The six evaluation tasks

The aim of the first task is to identify various zones in a document and specify their position (module 1). Area classification consisted in outlining an image region and attributing it with a type that describes its nature (labelling). Area segmentation is carried out using closed polygonalshaped outlines. Different semantic areas may overlap. For instance, a table area may overlap a set of other areas [16], including text areas, and a graph area (logo, signature, etc.) that may appear in the background of a text area. The second task is the identification of the writing type (module 2). It consists in determining the type of writing used in text zones: handwritten, printed or unspecified (rejection).
The third task is a language identification task. It consists in determining the language(s) used in each text zone (module 3). Languages to be identified are French, English and Arabic. Other languages used in the document should be classified as "Other language". Task 4 consists in transcribing the content of each text area (module 4). Task 5 is on extraction of logical structure (module 5) and consists in determining logical connections between semantic areas (for instance, the connection between an image and the text area in the caption associated with it) and, where applicable, readings order for various areas (for example, a column sequence in an article).
At last, task 6 is based on keyword spotting scenario performed for the end-to-end processing chain.

B. Metrics

The Zone Map metric is used to evaluate the module 1. It is the generalization of the metrics P set [17] and Det Eval [18]. Zone Map allows taking into account superposition of zones as it can appear, for example, in tables or crossingouts. The so-called Jac card metric takes into account the surface measured in black pixels, but not the decomposition of zones. It is calculated based on surface assigned to a class in the reference and the surface assigned to this class in the hypothesis. For each zone class i the Jac card index Ji was defined as

$$J_i = \frac{H_i \cap R_i}{H_i \cup R_i}$$

The document score Jdoc was defined as:

$$J_i = \frac{H_i \cap R_i}{H_i \cup R_i}$$

The document score Jdoc was defined as:

$$J_{doc} = \frac{\sum_{i=0}^{N} ((H_i \cup R_i).J_i)}{\sum_{i=0}^{N} (H_i \cup R_i)} = \frac{\sum_{i=0}^{N} (H_i \cap R_i)}{\sum_{i=0}^{N} (H_i \cup R_i)}$$

Modules 2 and 3 are evaluated by means of precision. The module 4 is evaluated at two different levels. The Word Error Rate (WER) is used at the word level and the Character Error Rate (CER) is used at the character level. The metric for the task 5 is based on the zone score. Each zone was characterized by three features that can be void if no logical structure is assigned to the zone: _ Semantic subtype (header, text body, etc.);
_ The area that precedes, in reading order, the one in question;
_ All areas (E) in the same non-ordered group of the area in question (E). Each of these characteristics gave a score between 0 and 1. For the first two, 1 point was counted for each correct answer. For the last one the harmonic mean (F-mesure) of the precision and recall was calculated after adding

the zone in question to the hypothesis. Each zone was attributed a zone score corresponding to the mean of three scores introduced above. The mean was calculated at the document level and then at the level of the set of documents that corresponded to the raw score Sb. Sb was then normalized with respect to S0 which is the score obtained by a hypothesis with all three characteristics void (i.e. a system provides no information on the logical structure). The final score S was as follows:

$$\text{If } S_b \le S_0, S = 100\frac{S_b - S_0}{S_0}, \text{ otherwise } S = 100\frac{S_b - S_0}{1 - S_0}$$

The final score was thus between -100 and 100. If it was positive it means that a system adds more correct information than errors.

C. Results of the first campaign

Four organizations took part in Task 1. The ZoneMap scores ranges from 57.3% to 107.1%. A ZoneMap score greater than 100 means that the system produced a lot of false alarms. The Jaccard scores ranged from 0.173 to 0.409. The two metrics are complementary. The difference between the two metrics is explained by the fact that Jaccard don't take into account the split and merge situations. All the results are presented in the table below:

System	Run	ZoneMap (%)	Jaccard
1	1	107.1	0.150
	2	90.7	0.169
	3	91.5	0.157
	4	91.6	0.162
2	1	**57.3**	0.190
3	1	75.9	0.315
	2	72.8	**0.382**
4	1	62.4	0.287
	2	62.3	0.286

Two organizations took part in Task 2. For global results, Precision ranged from 38.9% to 63.8% according to the system. The following table presents the results according to the writing type.

System	Précision (%)		
	All	Typed	Handwritten
1	38.9	35.5	49.0
2	**63.8**	**64.5**	**61.7**

Two organizations took part in Task 3. Precision ranged from 89.9% to 90.4%. The following table presents the results according to the language.

	ALL	ENG	ARB	FRA	Other
1	89.9%	**86.4**	89.3	**93.3**	90.2
2	**90.4%**	85.9	**93.0**	90.8	**96.2**

Five organizations took part in Task 4. One participant submitted only for handwritten and latin zones. All the results are presented in tables below. WER varies significantly depending on the writing type and the language.

	System	Run	ARB	FRA	ENG	latin	all
Typed	1	1	**58.3**	31.0	39.2	34.4	37.1
	2	1	91.3	**21.0**	**20.8**	**20.9**	**28.9**
	3	1	161.3	141.5	160.4	149.4	150.6
	4	1	112.4	63.9	66.7	65.0	70.4
	5	1	-	-	-	-	-
		2	-	-	-	-	-
		3	-	-	-	-	-
		4	-	-	-	-	-

	System	Run	ARB	FRA	ENG	latin	all
Handwritten	1	1	58.0	58.0	58.0	58.0	58.0
	2	1	-	-	-	-	-
	3	1	149.3	149.3	149.3	149.3	149.3
	4	1	101.0	101.0	101.0	101.0	101.0
	5	1	-	39.4	41.6	39.9	56.2
		2	-	38.5	41.7	39.3	55.7
		3	-	36.3	40.5	37.4	54.3
		4	-	**34.5**	**38.0**	**35.4**	**52.8**

Three organizations took part in Task 5. All the results are presented in table below.

System Type Order Group

System	Type	Order	Group
1	**59**	**22**	27
2	42	17	**37**
3	10	2	26

F. Management of the processing chain

The Maurdor demonstrator allows the user to define a particular processing scenario. The processing chain can be edited and the user can chose to plug one or another of the available processing modules to assume a selected step. It is also possible to insert alternative instructions within logical conditions into the processing chain definition in order to perform one or another module depending on the properties of the document to be processed. To help the user choice, the demonstrator displays performance indicators that have been measured during previous run sessions. At last, a running chain can be monitored to check the progress (count of processed documents, average time for a document processing, etc.), to pause or to stop the current execution.

III. CONCLUSION

The Maurdor project is a thorough analysis of state of the art OCR modules on a challenging corpus. The variety of document models, scenarios and scribes renders a realistic overview of documents an OCR could encounter. The second evaluation campaign is taking place in December 2013 on a larger subset of documents from the corpus. The different modules previously developed are being improved and new competitors are presenting their systems. The demonstrator will be completed by the end of February 2014. The final results will be presented during the DAS workshop.

REFERENCES

__ N. Kharma, M. Ahmed, and R. Ward, "A new comprehensive
database of handwritten Arabic words, numbers, and signatures used
for OCR testing," in *1999 IEEE Canadian Conference on Electrical
and Computer Engineering*, 1999, vol. 2, pp. 766–768 vol.2
__ NIST, NIST 2013 "Open Handwriting Recognition and Translation
Evaluation Plan",
http://www.nist.gov/itl/iad/mig/upload/OpenHaRT2013_EvalPlan_v1
-7.pdf, 2013
__ E. Grosicki, M.Carré, JM. Brodin, E. Geoffrois – RIMES evaluation
campaign for handwritten mail processing - In *Proc. of the Int. Conf.
on Frontiers in Handwriting Recognition*., 2009
__ E. Grosicki, H. El-Abed French handwriting recognition competition**.**
In *Proc. of the Int. Conf. on Document Analysis and Recognition*.
ICDAR Conference, Beijing 2011
__ C. A. Mello and R. D. Lins, "Image segmentation of historical
documents," *Vis. Mex. City Mex.*, vol. 30, 2000
__ E. Voorhees and D. K. Harman, *TREC: Experiment and evaluation in
information retrieval*, vol. 63. MIT press Cambridge, 2005
__ T. Deselaers, H. Müller, P. Clough, H. Ney, and T. M. Lehmann,
"The CLEF 2005 automatic medical image annotation task," *Int. J.
Comput. Vis.*, vol. 74, no. 1, pp. 51–58, 2007.
__ O. Galibert and J. Kahn, "The First Official REPERE Evaluation,"
First Workshop Speech Lang. Audio Multimed. SLAM 2013, 2013
__ http://www.rimes-database.fr/doku.php
___ http://www.iam.unibe.ch/fki/databases/iam-handwriting-database
___ http://catalog.ldc.upenn.edu/LDC2013T09
___ http://www.nist.gov/itl/iad/mig/hart2013.cfm
___ http://diuf.unifr.ch/diva/APTI/
___
http://lampsrv02.umiacs.umd.edu/projdb/project.php?id=53
___ http://www.elra.info/
___ T Kasar, P.Barlas, S.Adam, C.Chatelain and T.Paquet – "Learning to
Detect Tables in Scanned Document Images using Line Information"
– ICDAR Conference, Washington 2013
___ S. Mao and T. Kanungo. Architecture of PSET: a page segmentation
evaluation toolkit, International Journal of Document Analysis and
Recognition (IJDAR), 4(3):205-217, 2002.
___ Ch. Wolf and J-M. Jolion. Object count/Area Graphs for the
Evaluation of Object Detection and Segmentation Algorithms,
International Journal on Document Analysis and Recognition
(IJDAR), 8(4):280-296, 2006.
___ http://weblab-project.org
___ http://www.liferay.com

354