Table Recognition Evaluation and Combination Methods

Férihane Kboubi
RIADI Laoratory/ENSI,
University of Manouba, Tunisia
Ferihane.kboubi@riadi.rnu.tn

Anja Habacha Chabi
RIADI Laoratory/ENSI,
University of Manouba, Tunisia
Anja.Habacha@ensi.rnu.tn

Mohamed Ben Ahmed
RIADI Laoratory/ENSI,
University of Manouba, Tunisia
mohamed.benahmed@riadi.rnu.tn

Abstract.

In this paper, we propose a new approach of document analysis and recognition (DAR) based on the combination of OCR systems. The proposed approach aims to improve the document recognition by combining the result of several OCR systems according to their performances. We focus our attention, in this paper, on the table combination. We start by presenting the results of the evaluation of OCR system in the table block recognition. Then, we present our table combination method. We are interested in both table structure and table content.

1. Introduction

Recently, several research works was undertaken to improve the accuracy on document recognition. One of the proposed solutions is the combination of multiple optical character recognition systems [1][2]. There are two major approaches of OCR combination, the sequential approach [8] and the parallel approach [6] [7]. Some other works, more recent, are proposed such as, the probabilistic method developed by Bennet and al [9] and the weighted classifier combination methodology developed by Baykut and Erçil [10]. All these methods are defined and applied only to combine textual blocks; none of them is applied to the tables, the graphs or the mathematical formulas.

The majority of the evaluation techniques, proposed in the literature, were focused only on the text recognition performance [5][14][15][16] or the segmentation performance [11][12][13][16]. Few works had considered the problem of evaluating table recognition systems [17]. In [17], Hu et al proposed two evaluation methods. One, for the evaluation of table detection results and the other for the evaluation of table structure recognition results. Their first method is based on the edit distance. The input of this method is the whole page document and not the table block. The elementary object considered in the evaluation is the table block and not the cell. This evaluation method doesn't take into consideration the errors specific to table recognition (such as cell spanning and splitting). Indeed, the error types considered (insertion errors, deletion errors, splitting errors and merging errors) are specific to document segmentation [13]. The second proposed method consists to represent tables (recognition results and ground truth) as a directed acyclic attribute graph. Then, pose a series of probes and correlate the responses of the two graphs. This method has two essential shortcomings. The first is that the performance measure doesn't reflect the committed error types as the proposed measure is expressed in terms of the correct answers number for all the probes. The second limit is how the probes could be generated automatically and efficiently.

In [3], general evaluation of OCR systems was presented. Several criteria were considered, such as: characters, tables and graphics recognition. As regarding to table structure recognition, the only calculated rate was of cell recognition. No interest was taken to identify table errors types and calculate their rates.

In the remainder of this paper we present, in section 2, the table recognition evaluation results. Then, we describe, in section 3, the principle of the proposed table combination method. Finally, we present, in section 4, the result of our combination method.

2. Evaluation of OCR systems

In this section, we identify and classify the table errors and present the proposed performance measures. Table 1 presents the table recognition error types that we defined. Figure 1 shows some examples of table recognition error.

Table 1: Types of table structure recognition errors

Error types of the tables recognition	Notation
Lines split	LS
Horizontal split of cells	HS
Column scission	CS
Vertical scission of cells	VS
Line fusion	LF
Horizontal fusion of cells	HF
Columns fusion	CF
Vertical fusion of cells	VF
Table split	TS
Non-detection of table	NR
Confusion of a table to text	CTT

We tested three groups of OCR systems. Six OCR systems handling Latin characters and operating under Linux: CLARA OCR, GOCR, LOCR, OCRAD, OCRE and OCRShop. Six OCR systems handling Latin characters and operating under Windows: TOCR, Omnipage, FineReader, Character Eyes, Textbridge and TypeReader.

1520-5263/05 $20.00 © 2005 IEEE

Table 2: Errors rates and recognition rate for all OCR softwares

	LS	CS	LF	CF	TS	VS	HS	VF	HF	CTT	NR	TR
Omnipage	41.70	18.31	0.00	0.00	0.00	0.30	0.00	0.26	0.00	6.94	2.78	34.72
Finereader	7.64	0.34	0.00	0.00	1.39	0.26	0.00	0.00	0.13	18.06	2.78	62.50
Typereader	0.87	2.03	2.40	0.34	22.22	1.21	0.13	0.00	0.04	12.50	0.00	45.83
Sakhr	2.40	1.02	0.44	0.00	4.17	0.13	0.13	0.30	1.38	1.39	1.39	66.67
Readiris	74.51	5.65	0.00	0.00	0.00	2.02	1.10	0.00	0.00	33.33	0.00	5.56
Total errors	51.17	13.59	2.76	0.21	4.24	9.34	1.27	2.76	7.64	5.94	1.06	

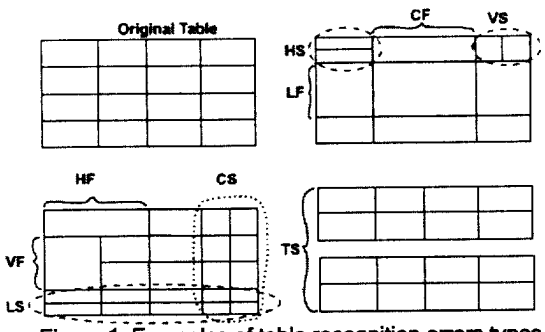


Figure 1. Examples of table recognition errors types

Three OCR softwares handling Arabic characters: AOCR, ReadIris and Sakhr. It is to be noted that Readiris and Sakhr treat Latin and Arabic characters.

In this paper, we calculated, for each system, its rates of all error types. Table 2 recapitulates the obtained results. TR indicates the rate of correctly recognized tables. We noticed that errors rates vary from software to another, for example Omnipage makes many errors of lines scission, its SL errors rate is equal to 45.7%. As for Typereader, it tends to split one table in several tables. Its ST rate is equal to 22.22%. We noticed for some OCR systems that if the framework of the columns is omitted, the table is recognized as being a simple text. The most frequent errors are the SL and SC errors. We were also interested in the evaluation of table content recognition. The obtained results were presented in [4]. We undertook the experiments of evaluation of OCR systems on a variety of documents. The used documents are written in French, English and/or Arabic. The images have a resolution equal to 300 dpi. The test set is made up of 100 images of tables of various structures. For the evaluation, we developed programs allowing the automatic detection of errors, the counting of their number and the deduction of the various recognition and error rates. The principle consists of comparing the obtained result with the ground truth.

3. Our table combination method

Our new approach of document analysis and recognition (DAR) based on OCR combination is composed of four steps. The first one consists of analysing the document, extracting its physical and logical structures, and then splitting it into homogeneous blocks.

The second step consists to recognize each block using the set of the most powerful OCR systems in the recognition of the data type in question (Latin character, Arabic character, table, graphic, etc). The OCR system selection is important. Indeed, taking into account the result of OCR systems, having a low performance, could fall a lot the final combination result. The OCR selection task is based on the type of each block and the characteristics of the OCR software. Consequently, before integrating new software in the combination system, it would be necessary to evaluate its performances concerning the recognition of each type of information. We defined five categories of OCR softwares:

- The ROCTL, which are the most powerful systems in the recognition of Latin characters.
- The ROCTA, which are the most powerful systems in Arabic characters recognition.
- The ROCTab, which are the most powerful systems in the table recognition.
- The ROCG, which are the most powerful systems in the graph recognition.
- The ROCFM, which are the most powerful systems in the mathematical formulaes recognition.

The same system can belong at the same time to several classes. If we have an OCR software which is powerful in the recognition of text in Latin characters and tables, and we want to recognize a document which contains at the same time a block of text in Latin characters and a table block, then this system will be activated twice: the first time to recognize the text and the second to recognize the table. The activation of the OCR softwares is done in a parallel way. For each block all the qualified softwares to its treatment are activated in parallel.

The third step consists to combine all OCR results concerning a given block. It is logical before starting this step to convert the results of all OCR systems into the same format.

In the last step, each fusion result must be formatted according to the description of the segment to which it corresponds and must be integrated in the final result in its adequate position.

In the remainder of this section, we focused our attention on the combination of table recognition results. We start by describing the proposed combination method. Then we present the experimentally obtained results.

3.1. The fusion of the tables blocks

Currently, several OCR systems are able to recognize table structures and offer many output formats. In order to

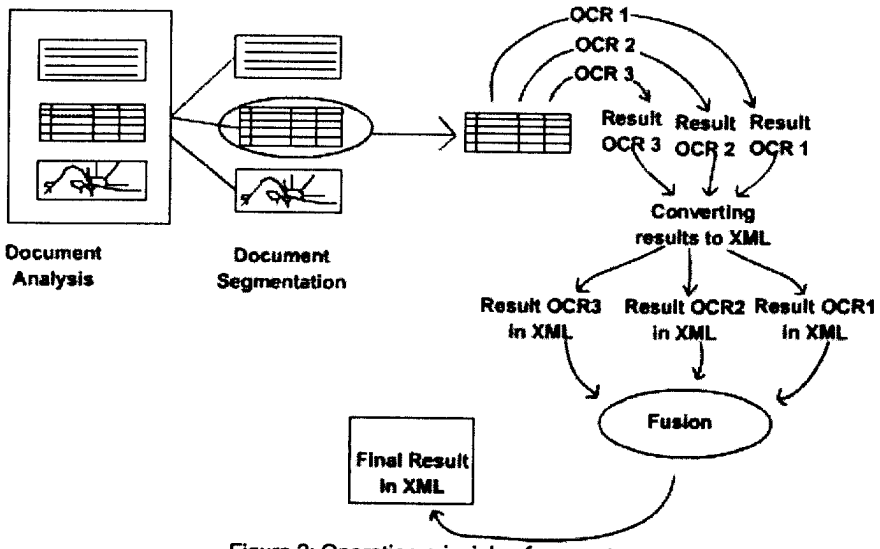


Figure 2: Operation principle of our system

combine OCR results we must use the same format for all the outputs. The selected format must allow the representation of all information relative to both structure and content of tables. In this section, we first propose a general representation of table blocks based on XML. Next, we present our table combination method. Figure 2 illustrates the operation principle of our system. For the fusion, we are interested as well in the structure as in the contents of the tables.

3.1.1. Converting OCR results to XML. We assume that a table is composed of a set of cells. Each cell belongs to one or several line(s) and column(s). Each cell has four position attributes. FL: the number of the first line to which the cell belongs, LL: the number of the last line to which the cell belongs, FC: the number of the first column to which the cell belongs and LC: the number of the last column to which the cell belongs. A cell is called spanning cell when the values of its FL and LL attributes are different or the values of its FC and LC attributes are different. The conversion is done automatically using the tool we developed to extract the table structure from the OCR results. Figure 3 shows a sample of converted table file. To gain space we simplified the DTD and the XML code presented in this Figure.

3.1.2. Fusion of the structure. For the fusion of the table recognition results we start by analysing them. Further to this analysis we eliminate the results where there is total or partial confusion of the table to text. Then, we classify *"Under reserves"* the results containing a table splitting. Lastly, we mark as *"Principal"* each result that contains exactly only one table. From the not eliminated results we extract several information:

- For each result, we determine the number of tables,
- For each table we calculate the number of lines and columns.
- For each line we determine the number of cells.
- For each cell we determine its content and its position (FL, LL, FC and LC attributes).

For fusion we consider only the *"Principal"* results, the results *"under reserves"* are used only if there is no *"Principal"* results. The principle of the table fusion algorithm is the following:

- Determine the number of lines to retain: start by seeking and retaining the majority number of lines. If there is irresoluteness then retain the minimum number of lines. Eliminate all the results, which do not have the same number of lines as the selected number.
- Determine the number of columns to retain: start by seeking the majority number of columns. If there is a case of irresoluteness, then retain the minimum number of columns. Eliminate the results, which have a different number of columns.
- Merge the structures of the remaining tables. To do this, we start by running, in parallel, through all the tables cell by cell:
  - *Step 1:* For a given level, if all cells have an identical structure (same values for FL, LL, FC and LC attributes) then retain this structure and merge the contents of all these cells.
  - *Step 2:* Else, seek the including cell on this level, then check if there is or not a possibility of cell splitting error. There is a possibility of scission if for all tables, the number of empty cells corresponding and nonequal to the including cell is higher or equal to the number of cells corresponding and nonequal to the including cell - 1. The example of Figure 4 illustrates this principle. The including cell is the C [2,5][1,1] of Finereader. The set E of cells corresponding to it in the result of Omnipage, is {C [2,2] [1,1], C [3,3] [1,1], C [4,4][1,1], C [5,5][1,1]}. The number of empty cells

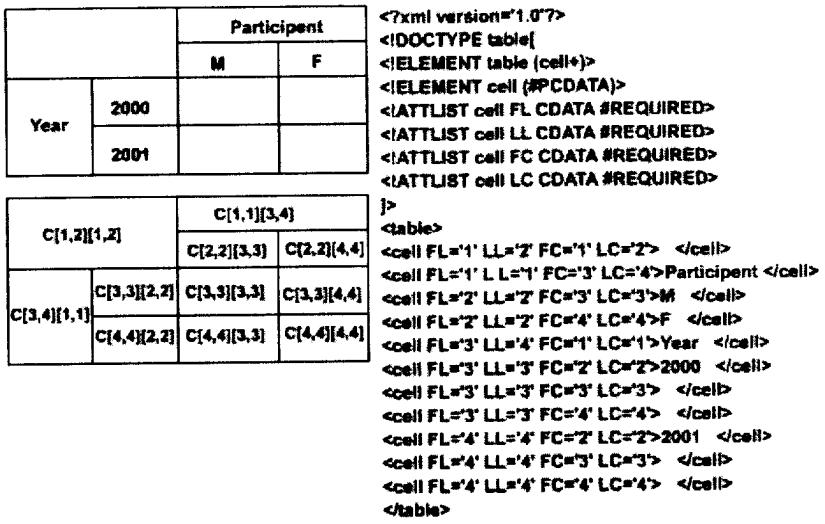
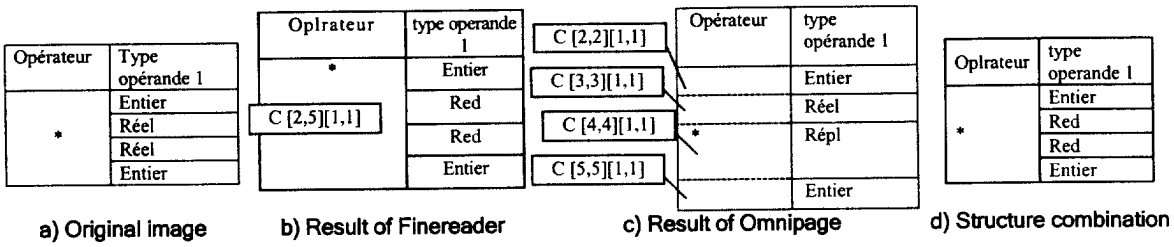


Figure 3. Table XML source code



a) Original image b) Result of Finereader c) Result of Omnipage d) Structure combination

Figure 4: Example of a situation of cell scission

belonging to E is equal to the total number of cells belonging to E-1. Then, there is a situation of scission; the structure selected is that of the C [2,5][1,1].

- *Step 3:* If the possibility of scission is checked, then retain the structure of the including cell and merge its contents with that of all the cells corresponding to it. Then, eliminate from all the other tables, all the cells corresponding and nonequal to the including cell.
- *Step 4:* Else, (i.e. possibility of scission is not checked for the structure of the including cell), eliminate this including cell and apply for the remaining cells the same treatment starting from step 2.
- *Step 5:* If no possibility of scission exists, then there is a situation of cell fusion. Therefore, it is necessary to retain the structure of the included cell and to merge its contents with all cells of the other tables that have the same structure as it.

3.1.3. Fusion of the table contents. The powerful softwares in the recognition of the text included in the tables can be eliminated at the fusion of the structure because they made errors of structure recognition. Thus, the softwares used for the fusion of the structures can be not very powerful in the characters recognition. If such a case arises, we propose to use the results of the first category of OCR softwares to improve the quality of text recognition. Thus, it would be necessary to put in correspondence the badly recognized structures with the selected one. For this, we use the edition distance measure to put in correspondence the cells of the various tables. In what follows we present the principle of the proposed algorithm for mapping two tables having different numbers of lines:

- Among the retained results, select the reference one, which is provided by the most powerful software in the character recognition.
- Let *Tsup* and *Tinf* respectively indicating the table having the high number of lines and the table having the lowest number of lines.
- Run through the two tables line by line. While we did not reach the end of *Tinf* and *Tsup*:
  - *Step 1:* For each line of *Tinf*, run through the cells,
  - *Step 2:* For each cell of the current line of *Tinf*, calculate the edition distance (let *EDmin* this distance) between its contents and that of the cell corresponding to it in *Tsup*,
  - *Step 3:* Then concatenate the content of the current cell of *Tsup* with that of the cells corresponding to it in the following line of *Tsup*,
  - *Step 4:* If the edition distance between the result of the concatenation and the current cell of *Tinf* is lower than *EDmin* then there is a possibility of line fusion. So we

Table 3: The evaluation results of the table structure fusion algorithm: recognition and error rates

	LS	CS	LF	CF	TS	VS	HS	VF	HF	CTT	NR	TR
Combination method	1.53	0.00	0.22	0.00	0.00	0.09	0.00	0.00	0.35	0.00	0.00	90.28

must assign the new value of edition distance to *EDmin* and loop back to step 3. We stop when the current edition distance is higher than *EDmin*. For the current level of line, the number of line to be merged is according to the number of concatenations carried out and is determined by majority vote.

4. Experimentation.

We undertook the evaluation experiments of our system on the same test set described in section 2. For the evaluation of the table fusion algorithm we used a set of OCR software regrouping: Omnipage, Finereader, Typereader and Sakhr. We eliminated Readiris because of its weak rates of table structure recognition. The obtained evaluation results are recapitulated in tables 3. We noticed, for the fusion algorithm, a remarkable reduction in the error count, and an appreciable increase in the number of completely recognized tables. Indeed, our method could recognize perfectly the structure of 90.28% of the tables of the test base, whereas the highest rate obtained by the OCR softwares is about 66.7% (Table 2). We noticed that the error rates of the combination method are definitely lower than those of OCR systems, some of them (CS, CF, TS, HS, VF, CTT and NR) became null after the combination.

5. Conclusion and future works

In this paper, we presented table recognition evaluation and combination methods. In our work we considered table structure and content. We accepted as input table blocks. With our combination method we obtained an enhancement of about 23.6% for the recognition rate compared to the best-tested OCR system. For the evaluation as well as for the combination we considered the cell as the elementary object. So for the structure of the table we focused our attention on the position of each cell in the table. We did not labeled cells according to their content (value or indication cell) because none of the tested OCR systems gives this type of information. We will consider this task in future works. We plan also, to test our combination method on a larger test set.

References

[1] B. A. Yanikoglu, L. Vincent, *PINK PANTHER: A complete environment for ground-truthing and benchmarking document page segmentation,* Pattern Recognition, September 1998, Vol. 31, No. 9, pp. 1191-1204.

[2] Kittler, J. M. Hatef, R. P. W. Duinand J.Matas, "*On Combining Classifiers*", IEEE Trans. on Pattern Analysis and Machine Intelligence, 1998, Vol.20, No.3, pp.226-240.

[3] Férihane Kboubi, Anja Habacha Chabi, Mohamed Ben Ahmed, *A New Strategy of OCR Combination*, 8th World Multi-Conference on Systemics, Cybernetics and Informatics, SCI 2004, Orlando, Florida.

[4] Férihane Kboubi, *Une nouvelle approche d'analyse et de reconnaissance de documents basée sur la combinaison de logiciels de ROC multilingues*, Master Thesis, ENSI, Tunisia, September 2004.

[5] A. Belaïd, L. Pierron. *A generic approach for OCR performance evaluation*. Electronic Imaging. San Jose,. 2002. 5p.

[6] T. K. Ho and J. J. Hull, S. N. Srihari, *Combination of Decisions by Multiple Classifiers*, Structured Document Image Analysis, Ed. H. S. Baird and H. Bunke and K. Yamamoto, Springer- Verlag, Heidelberg, 1992, pp. 188-202.

[7] T. K. Ho, J. J. Hull, S. N. Srihari, *On Multiple Classifier Systems for Pattern Recognition*, Proc.11th Int Conference on Pattern Recognition, Netherlands, 1992, pp. 84-87.

[8] S. Behnke, M. Pfister, R. Rojas, *A Study on the Combination of Classifiers for Handwritten Digit Recognition*, Proc of Neural Networks in Applications, Third International Workshop 1998, Magdeburg, Germany, 39-46,

[9]P. N. Bennett, S. T. Dumais, E. Horvitz, *Probabilistic Combination of Text Classifiers Using Reliability Indicators: Models and Results*, Proc 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Finland, August 2002

[10] A. Baykut, A. Erçil: *Towards Automated Classifier Combination for Pattern Recognition*. Multiple Classifier Systems 2003, pp 94-105

[11] B. A. Yanikoglu, L. Vincent, *PINK PANTHER: A complete environment for ground-truthing and benchmarking document page segmentation,* Pattern Recognition, 1998, Vol. 31, No. 9, pp. 1191-1204.

[12] J. Kanai, S. V. Rice, T. A. Nartker, G. Nagy. *Automated Evaluation of OCR Zoning*. IEEE Transaction on Pattern Analysis and Machine intelligence,Vol 17, No 1, 1995. 86-90.

[13] S. Mao, T. Kanungo. *Empirical Performance Evaluation Methodology and its Application to Page Segmentation Algorithms*. IEEE Transaction on Pattern Analysis And Machine Intelligence, Vol. 23, No. 3, Mars 2001, pp 242-256.

[14] T. Kanungo, G. A. Marton, O. Bulbul, *Paired Model Evaluation of OCR Algorithms, UMD*--TR3972, December 1998.

[15] K. Swam, A. Brodeen, *FALCon: Evaluation of OCR and Machine Translation Paradigms*, US Army Research Laboratory, 1999.

[16] T.Kanai, T.A.Nartker, S.V.Rice, *Performance Metrics for Document understanding systems*, In Proc. 2nd Intl. Conf. on Document Analysis and Recognition, Tsukuba Science City, Japan, October 1993, pp 424-427.

[17] J. Hu, R. Kashi, D. Lopresti, G. Wilfong, *Evaluating the performance of table processing algorithms*, International Journal on Document Analysis and Recognition, Vol. 4, No. 3, March 2002, pp. 140-153.