A Minimal and Sufficient Way of Introducing External Knowledge for Table Recognition in Archival Documents

Isaac Martinat[1] and Bertrand Coüasnon[2]

[1] IRISA/INSA, Campus universitaire de Beaulieu, F-35042 Rennes Cedex, France
Isaac.Martinat@irisa.fr

[2] IRISA/INRIA, Campus universitaire de Beaulieu, F-35042 Rennes Cedex, France
Bertrand.Couasnon@irisa.fr

Abstract. We present a system that recognizes tables in archival documents. Many works were carried out on table recognition but very few on tables of historical documents. These are difficult to analyze because they are often damaged due to their age and conservation. Therefore we have to introduce knowledge to compensate for missing information and noise in these documents. As there is a very important number of documents of a same type, the cost is not significant to introduce this explicit knowledge. We also want to minimalize the cost to adapt the system for a given document type. The precision of the knowledge given by the user is dependent on the quality of the document. The more the document is damaged, the more the specification has to be precise. We will show in this article how an external minimal knowledge can be sufficient for an efficient recognition system for tables in archival documents.

Keywords: Archival documents, knowledge specification, structured document analysis, table recognition.

1 Introduction

We present a system that recognizes tables in archival documents. Many works were carried out on table recognition [1, 2] but very few on archival tables. These are difficult to analyze because they are often damaged due to their age and conservation. We will only analyze tables with ruling separators between columns and rows. The rulings can be broken, skewed or curved. Another difficulty is that ink bleeds through the paper, thus rulings of flip side can be visible. For these reasons, these tables are very difficult to recognize.

The problem in recognizing archival documents is that these documents have missing information and can contain false information like flip side rulings or stains. Therefore, the user has to give knowledge to compensate these analysis difficulties. However, this knowledge has to be minimal for a fast adaptation between different document types. It has to be simple, so non-document analysis specialists can easily define it. This minimal knowledge must be sufficient to help the system to recognize these difficult documents. Therefore, we have to define a minimal and sufficient knowledge for the archival table recognition.

W. Liu and J. Lladós (Eds.): GREC 2005, LNCS 3926, pp. 206–217, 2006.

In this paper, we will first present the related work on table recognition and on archival document analysis. Furthermore, we will show with the knowledge specification of the DMOS method the necessity for archival documents to give precise knowledge. Section 4 proposes for archival table recognition the necessary knowledge and explains our system uses it. We will finally show our results before to conclude on our work.

2 Related Work

2.1 Table Form Analysis

Many works were carried out on table recognition [1, 2]. We will present only the works on table and form recognition with rulings.

Handley [3] presented a method for table analysis with multi-line cells. This method first extracts from the image word boxes and rulings. Rulings whose end points are closed, are stitched together. Then for each word box, close rulings are researched and a frame is associated for each word box. This method merges word boxes with identical frames. To recognize rows and columns not separated by rulings, it then uses histogram procedure on the two axes. However, this detection is inefficient on curved documents. This method detects only broken rulings with small gaps. The method proposed in [4] detects from a binary image line segments in using erosion and dilation operations. This line segment extraction fills some breaks of form lines. They also used rules to detect bigger gaps, but these gaps are only detected in specific cases. Hori and Doermann [5] reduced the original image. In the reduced image, broken lines can be changed in solid lines but the size of detected gaps depends on factor reduction. The method proposed in [6] analyzes telephone company tables. It can recognize rulings with gaps but user has to give the maximal gap size to group segment lines.

These methods deal with broken lines but only small gaps are filled, or these gaps must be under certain conditions. Archival documents can be very damaged and can contain big gaps. Therefore these methods can not be adapted to archival documents.

2.2 Ancient Document Analysis

Few works were carried out on archival document analysis. The analysis of these documents is difficult because they are quite damaged. These documents have annotations, are torn and ink bleeds through the paper. Therefore a recognition system for archival documents needs knowledge given from the user.

He et al. [7] used a graphical interface to recognize archive biological cards. Each card contains bibliographic data and other information for one genus-group or one species-group, there are in total about ten text fields and the most of information is typewritten. The user defines boxes with this interface and labels each box. From this one a template is created, then the user can add information. With fuzzy positions, a X-Y cut method is used to analyze cards and a matching algorithm is applied between the template and the analysis. This system is specific for archive biological cards. This method uses positions from the graphical interface and fuzzy positions to analyze documents but it is efficient only on documents of a same type which have not important variations. Esposito et al. [8] designed a document processing system *WISDOM++* that has been used on archival documents (articles, registration card). This system segments the document with a hybrid method, global analysis and local analysis. The result of this analysis can be modified by the user. Training observations are generated from these user operations. With these results, the document is then associated to a class of model documents. The method presented in [9] analyzed lists of Word War II, which do not contain rulings. For a set of documents containing the same logical structure, historians and archivists use a graphical interface to define a *template* where physical entities on a page are associated with logical information. All these methods use physical information from a model generated by a graphical interface or learned on a set of documents corrected by a user. The variations between documents of a same type depend on the matching algorithm between the image and the model. Furthermore it takes time for an user to give the model information.

For the recognition of tables with rulings, Tubbs et al. recognized 1910 U.S. census tables [10] but coordinates for each cell of the tables are given at hand in an input of 1,451 file lines. The drawback of this method is the long time spent by the user to define this description. Furthermore, the coordinate specifications do not allow variations on the documents of the defined type. Nielson et al. [11] recognized tables whose rows and columns are separated by rulings. Projection profiles are used to identify rulings. For each document a mesh is created, and individual meshes are combined to form a template with a single mesh. This method cannot process documents where rulings are skewed or curved. Individual meshes must be almost identical to be combined.

Archival documents are often damaged and recognition systems need an user specification to recognize these. The general systems presented in Sect. 2.1 cannot process these documents because they do not detect broken lines with big gaps. To help the archival document recognition, systems use an user description [10], a graphical interface [7, 9], information of other documents of the same type [11] or user corrections [8]. These works use external knowledge. However, it is often quite long to define and too precise, so these systems do not allow important variations between documents.

A system to recognize archival documents needs an external knowledge, so we propose in this article a minimal knowledge for the recognition of archival tables. We will show how this knowledge is simple, fast to give to the system, independent of physical structure if document is not too damaged and sufficient to recognize very damaged documents.

3 Knowledge Specification with DMOS Method

With the DMOS (Description and Modification of Segmentation) method we can give a description for a document type. DMOS is a generic recognition method for structured documents [12]. This method is made of a grammatical formalism EPF (Enhanced Position Formalism) and an associated parser which is able to change the parsed structure during the parsing. With the DMOS method we can build a system for a kind of document by defining a description of the document with an EPF grammar. This grammar is then compiled to produce a recognition system. We will show how the knowledge is represented in EPF formalism and the necessity for archival documents to have a very precise description.

3.1 Knowledge Representation in EPF Formalism

With the DMOS method and the EPF formalism, a system is created much faster than to develop completly a new recognition system. EPF can be seen as an adding of several operators to mono-dimensional grammars like the principal one, the *Position operator (AT)*. For example, `A && AT(pos) && B` means A, and at position **pos** in relation to A, we find B.

The DMOS method is generic because the EPF formalism allows to define very different kinds of structured documents. This method was tested, for example, on musical scores, on mathematical formulae, on table structures [12] and on archival documents [13].

3.2 General and Specific Systems in EPF

A general system was built in EPF formalism to analyze all kinds of table-forms [14]. This system can recognize the hierarchical organization of a table made with rulings, whatever the number/size of columns/rows and the depth of the hierarchy contents in it. However we [14] showed that this general system was not able to be applied for archival documents. These documents are damaged and gaps in rulings are too large, which makes it impossible for a general system to decide if there is a gap or a normal absence of ruling. Therefore, a much more precise description is necessary to recognize these. A system was built for military forms of the 19th Century. A grammar describing these forms and the relative positions of the cells was written in EPF. It has been tested on 164,479 forms and 98.73% were well recognized with correct cell positions. There was no bad recognition. Another system was built for naturalization decrees [13]. These documents are on two columns separated by spaces. These systems are efficient on archival documents. However, even if these descriptions in EPF are faster to write than to develop a specific system, they are still quite long to define and accessible only for document analysis specialists.

4 Knowledge for Archival Table Recognition and Recognition System

Our goal is to propose a specific system for archival tables. For archival documents, the user has to give knowledge to compensate for missing information in these. DMOS is an efficient method but descriptions in EPF can be still long to define. Furthermore, it is difficult to define a precise knowledge for damaged documents.

The proposed system is specific but it can deal with a large variety of tables and with a fast adaptation. The specified knowledge can be given by a non-specialist user. Thus it must be easy to specify, minimal but sufficient to help the system. We will show the necessary knowledge for archival tables and how our system uses it.

4.1 Necessary Knowledge Formalization

We have a very important number of documents to process. For example we have a dataset of about 130,000 census tables from 1817 to 1968. These censuses were carried out on 24 different years and often different from a year to another. Therefore, we have an important document quantity of the same type (about 5,400 images) so that the time used to give this short specification is not significant. We can ask the user to spend little time to define an external knowledge if the latter is useful for a large quantity of documents.

We want to adapt quickly the recognition system to a large variety of tables. The knowledge introduced by the user has to be simple, so an archivist can give this specification. Therefore document analysis parameters (gap size between two line segments to form a line, ruling minimal size . . .) cannot be used for this purpose. We want to minimize the specification given by the user but the system needs enough precise specification to be efficient. This knowledge must be minimal and sufficient to help the system for the document recognition. Thus user can give specification in relation to document quality, if document is good quality few informations are necessary but more precise informations are necessary for very damaged documents.

For a table, the minimal knowledge can be the number of rows and the number of columns. In Fig. 1, we show on the left example that this information is sufficient to help the system to recognize a synthetic document which misses information. In this example, a system not adapted to archival documents will recognize only two rows. However with the user specification that the number of rows is three, system can detect the line segment for a row separator ruling.

For a grade table of 25 students, the user gives the following specification using the number of rows and columns or the name of each column:

```
[ rowNumber 25 , colNumber 3 ] Or
[ rowNumber 25, col "last name", col "first name", col "grade" ]
```

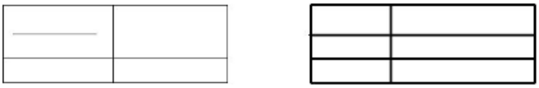

Fig. 1. left: synthetic image illustrating missing information, right: structure with 3 rows and 2 columns to recognize

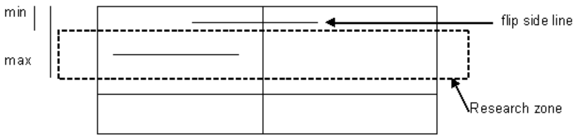

Fig. 2. Example with the same structure as previous example to illustrate the knowledge introduction more specific to detect ambiguous cases

For more damaged documents, the previous knowledge can be insufficient. A more precise specification must be given by the user to process more difficult documents. On another example (Fig. 2) with the same user specification as previously we show that the row number is not sufficient. If, on the document, ink bleeds through the paper, a false line segment is detected because of a visible flip side ruling. The following detected line segment is a row separator, but it has an equal length to the false line segment. Therefore, the system cannot decide which line segment is a row separator. However, if the user specified a minimal size of rows large enough to avoid the false line segment, the system will research the row separator ruling in a research zone that does not contain the false line segment.

For columns and rows, minimal and maximal global sizes can be given by the user. These sizes are used for every row or column. Sizes can be given in pixels or if the document density is known, sizes can also be given in centimeters or in inches. An user can give the following specification with global sizes:

```
[ rowMin 20, rowMax 150,colMin (cm 1.0), colMax (cm 8.0),
  rowNumber 25, col "last name", col "first name", col "grade" ]
```

When documents are very damaged, if these global sizes are not sufficient, the user can give specific sizes for each column/row or for a specific column/row. Column and row sizes are more constrained but they can have some variations between documents. In this example, a grade is given in digits, so the user can give a small size for this column with this following specification:

```
[ rowNumber 25, col "last name", col "first name",
  colsize (inch 1.2) (inch 2.3) "grade" ]
```

The user gives a specifications in relation to the quality of the document. He will give only the necessary knowledge. When archival documents are not too damaged, only the numbers of rows and columns are necessary. On the other hand, when documents are very damaged, the user can specify more precise knowledge to help the system make the right choice when it recognizes a document.

4.2 System Defintion

To build a document analysis system, we need to choose constraints. This choice is difficult because if we choose too many constraints, documents will be undersegmented. However, if we choose too few constraints, documents will be oversegmented. For example, to detect a broken ruling, we have to choose the gap size between two line segments to decide if they belong to the same ruling. If the size is too small, few broken rulings will be detected, but if the size is too big, false rulings could be detected. The knowledge given by the user helps the system to decide which ruling has to be detected.

Our recognition is made of three steps, the first one is the detection of table borders, the second one is the column detection from right border to left border and the last one is the row detection from top border to bottom border. The two last steps use the user specification and they allow to adapt constraints for the recognition.

4.3 Use of External Knowledge

We have shown the advantages of the DMOS method: its efficiency and its associated EPF formalism. The EPF formalism allows a document analysis specialist to define quite quickly a recognition system. Therefore we used this method to define the proposed system for archival tables. The latter takes in argument a knowledge easy to define for a non-specialist of document analysis, for example an archivist.

Number of Rows and Columns. The system tries to detect the number of rows N and columns according to the user specification. As for rows, the system from the top border detects the row separators. Gap size is fixed to a small value, thus the system can not oversegment the table. However, if the bottom border is detected and the number of detected rows is less than N, the document is undersegmented, some rows were not detected. Gap size is then increased and the length ratio is decreased to allow the system to detect more broken lines. Length ratio is the ratio between the top border and the detected line. The system tries again the recognition until N rows are detected or if constraints are too weak, i.e the gap size is too big or the length ratio is too small. This method is written easily with several rules in the EPF formalism, we removed some arguments to simplify the writing. We will show the other arguments to explain how sizes are used by the system.

```
findRows Gap LengthR TopLine N ::=
  not (findRowSep Gap TopLine N ListDetectedLines) &&
  ''(NewGap is Gap + IncrGap, LengthR2 is LengthR - DecrRatio) &&
  findRows NewGap LengthR2 TopLine N.
```

`findRowSep` searchs N row separators from the top border. This rule fails when the bottom border is detected and the number of recognized rows is less than N. To check that a false row was not detected, the last line of the list of detected lines must be the bottom border. Otherwise, the system stops and informs the user. It is defined for columns as for rows.

Sizes. For the recognition of rows and columns, the system uses sizes given by the user when sizes were specified. If the user did not give sizes, minimal size is 0 and maximal size is the image size (width for columns and height for rows).

```
findColSep GlobalMin GlobalMax RightLine [Col|ListeCols] ::=
    getSize Col GlobalMin GlobalMax Min Max &&
    findLineV Gap Min Max RightLine LeftLine &&
    ''( sameSize RightLine LeftLine LengthRatio ) &&
    findColSep GlobalMin GlobalMax RightLine ListeCols.
```

`getSize` returns the *Min* and *Max* sizes in relation to the user specification. `findLineV` finds a broken vertical line nearest to the left of RightLine at a distance between GlobalMin and GlobalMax and `sameSize` is true if the Ratio of *LeftLine* and *RightLine* is greater than LengthRatio. By recursivity, the system detects then the other vertical lines.

If the user specifies sizes for a specific column, these sizes are used to search the next vertical line.

```
findLineV Gap Min Max RightLine LeftLine ::=
   AT (nearLeft Gap Min Max RightLine) &&
   brokenLineV Gap LeftLine && ''(parallel RightLine LeftLine).
```

With the *AT* operator, we defined the research zone to find *LeftLine* from *RightLine*, *Min* and *Max* values define the zone width, and zone height is defined with Gap value.

The EPF formalism has allowed us to define quickly a system using external knowledge. This description show how the system uses the knowledge given by the user.

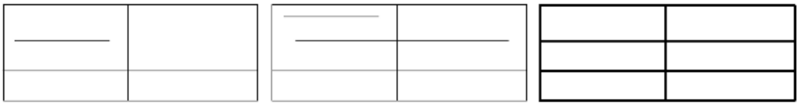

Fig. 3. left: synthetic image illustrating missing information, middle: synthetic image illustrating false ruling, right: structure with 3 rows and 2 columns to recognize

4.4 System Efficiency

We show in Fig. 3 how the constraint adaptation makes our system efficient. With a weak constant constraint of minimal ruling size, the system would recognize the middle example with four rows instead of three rows. With a strong enough constant constraint, the left example would be recognized with only two rows. Whereas our system, on the left example, begins the recognition with strong constraints and does not detect three rows so it will try again with less constraints until it recognizes the structure specified (`[rowNumber 3, colNumber 2]`). On the middle example, our system begins the recognition with strong constraints, the shortest ruling is not recognized as a row separator so the system will correctly recognize the structure. Therefore, it is very important to begin the recognition with strong constraints and to reattempt with less strong constraints only if the document is not well recognized.

5 Results

We show on some documents how the user specification helps the recognition system.

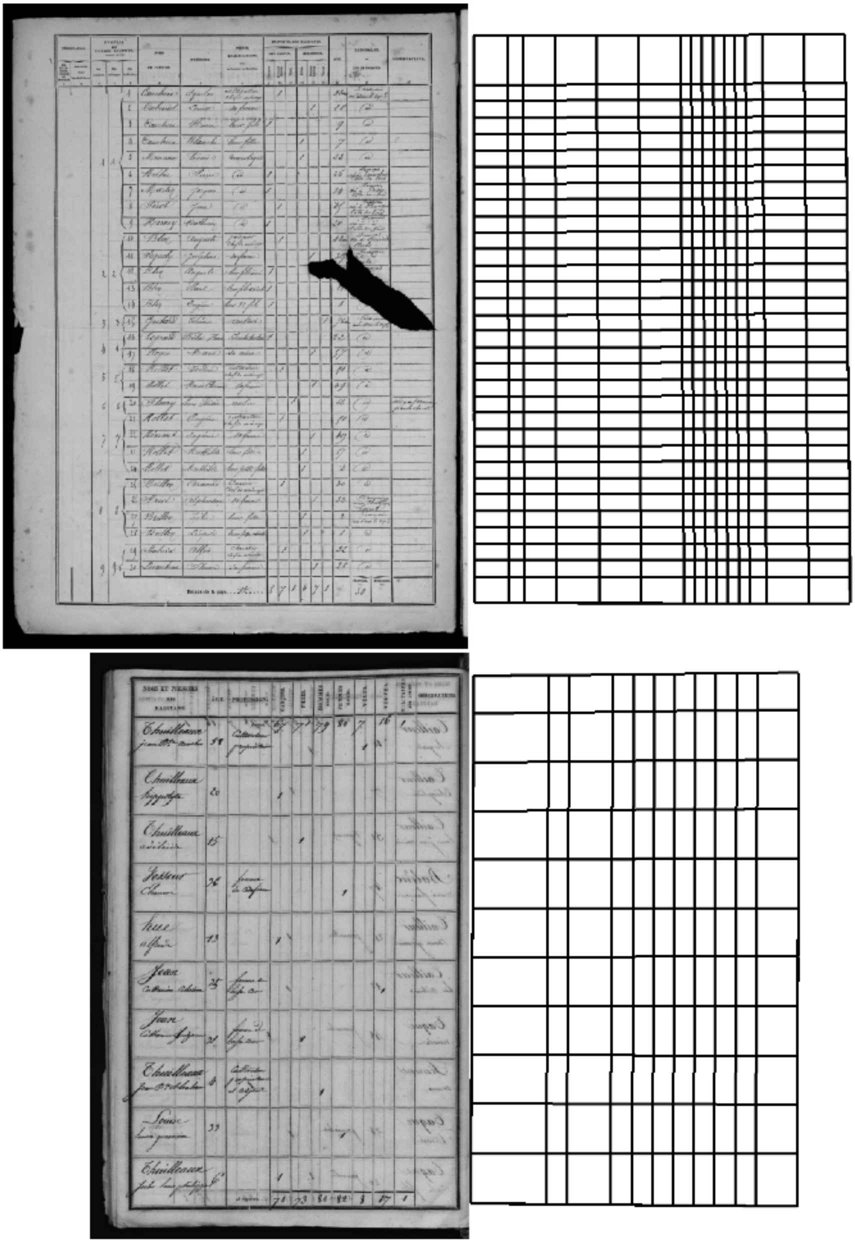

Fig. 4. Example on archival documents, on right the recognized structure

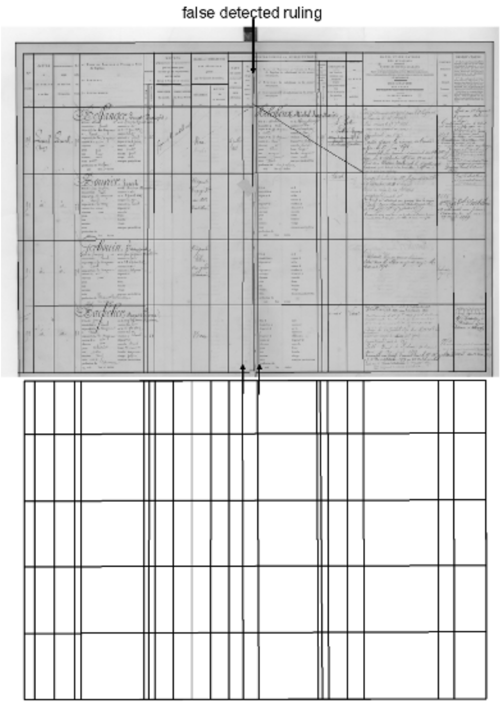

Fig. 5. Example on a two page document where with a specific column size a false detected ruling is avoided, on bottom the recognized structure

The document in the top of Fig. 4 is very damaged. The following specification: `[ rowMin 80, rowNumber 32, colNumber 15]` allows to recognize this document. The column number is sufficient to detect columns even if the paper is very torn. At the first step, the last columns are not detected, the gaps in column separators are too big. Thus the system tries again several times the column recognition by increasing the gap size value until the right number of columns is detected. As for rows, we need to give a minimal size to avoid detecting flip side rulings. Therefore, the system will research in zones which do not contain them.

In the document on the bottom of Fig. 4, vertical flip side rulings are visible. To avoid detecting these rulings we have to give general column sizes and specific sizes with the following specification:

```
[ rowNumber 11, colMin 100, colMax 500,
colsize  200 500 "names", col "age", colsize 200 500 "profession",
col "boys", col "girls", col "bridegroom", col "bride",
col "widower", col "widow",col "military", col "observations" ].
```

Figure 5 shows a result on an archival document of two pages. A false ruling can be detected with the separation of these two pages. Therefore, to avoid this problem, the user can specify the minimal size for the column containing the false ruling. Therefore when this column is detected, the system from the right column separator searchs the left column separator to a distance greater than the distance between the right separator and the false ruling. In this case, the user cannot give a general minimal size for all columns because on this document, there are very small columns that would not be detected.

On our first tests, we tested our system on 62 tables with the same specification and we checked 1922 cells. Only 2 adjacent cells were not well detected. Handwriting was present on the row separator and a false segment was detected from this handwriting. Table recognition evaluation is not easy so we need much more time to check results on a much more important number of documents which have been recognized. We have demonstrated on these results how the minimal knowledge that we proposed is easy to define and useful for the recognition system.

6 Conclusion

We have shown in this article how archival tables are very difficult to process because they can be very damaged. An external knowledge is necessary to help the recognition system to analyze these. This knowledge allows the system to recognize a structure which misses information and containing false information. To adapt this system quickly and to facilitate the introduction of this knowledge, we defined a minimal one. We have also presented how this minimal knowledge is sufficient and how that is easy for a user to give this specification. We presented on some results how our system is able to recognize very difficult documents.

Our future work is to design a much more general language, simple to use and sufficient to recognize all kinds of archival documents with tabular structures. We will seek to define a minimal and sufficient knowledge for more complicated tables: tables whose rows and columns can be separated by spaces, tables with recursive structure and forms.

Acknowledgments

This work has been done in cooperation with the *Archives départementales des Yvelines* in France, with the support of the *Conseil Général des Yvelines.*

References

1. Lopresti, D.P., Nagy, G.: A tabular survey of automated table processing. In: Selected Papers from the Third International Workshop on Graphics Recognition: Recent Advances and Perspectives. Volume 1941 of LNCS., Springer (2000) 93–120
2. Zanibbi, R., Blostein, D., Cordy, J.R.: A survey of table recognition. International Journal of Document Analysis and Recognition (IJDAR) **7**(1) (2004) 1–16
3. Handley, J.C.: Table analysis for multiline cell identification. In: Proceedings of SPIE – Volume 4307 Document Recognition and Retrieval VIII. (2000) 34–43

4. Xingyuan, L., Gao, W., Doermann, D., Oh, W.G.: A robust method for unknown forms analysis. In: 5th International Conference on Document Analysis and Recognition (ICDAR 1999), Bangalore, India (1999) 531–534
5. Hori, O., Doermann, D.S.: Robust table-form structure analysis based on box-driven reasoning. In: 3th International Conference on Document Analysis and Recognition (ICDAR 1995), Montreal, Canada (1995) 218–221
6. Chhabra, A.K., Misra, V., Arias, J.F.: Detection of horizontal lines in noisy run length encoded images: The fast method. In: Selected Papers from the First International Workshop on Graphics Recognition, Methods and Applications. Volume 1072 of LNCS., Springer (1996) 35–48
7. He, J., Downton, A.C.: User-assisted archive document image analysis for digital library construction. In: 7th International Conference on Document Analysis and Recognition (ICDAR 2003), Edinburgh, UK (2003) 498–502
8. Esposito, F., Malerba, D., Semeraro, G., Ferilli, S., Altamura, O., Basile, T.M.A., Berardi, M., Ceci, M., Mauro, N.D.: Machine learning methods for automatically processing historical documents: From paper acquisition to xml transformation. In: 1st International Workshop on Document Image Analysis for Libraries (DIAL 2004), Palo Alto, CA, USA (2004) 328–335
9. Antonacopoulos, A., Karatzas, D.: Document image analysis for world war 2 personal records. In: 1st International Workshop on Document Image Analysis for Libraries (DIAL 2004), Palo Alto, CA, USA (2004) 336–341
10. Tubbs, K., Embley, D.: Recognizing records from the extracted cells of microfilm tables. In: ACM Symposium on Document Engineering. (2002) 149–156
11. Nielson, H., Barrett, W.: Consensus-based table form recognition. In: 7th International Conference on Document Analysis and Recognition (ICDAR 2003), Edinburgh, UK (2003) 906–910
12. Coüasnon, B.: Dmos: A generic document recognition method, application to an automatic generator of musical scores, mathematical formulae and table structures recognition systems. In: 6th International Conference on Document Analysis and Recognition (ICDAR 2001), Seattle, WA, USA (2001) 215–220
13. Coüasnon, B., Camillerapp, J., Leplumey, I.: Making handwritten archives documents accessible to public with a generic system of document image analysis. In: 1st International Workshop on Document Image Analysis for Libraries (DIAL 2004), Palo Alto, CA, USA (2004) 270–277
14. Coüasnon, B.: Dmos, a generic document recognition method: Application to table structure analysis in a general and in a specific way. International Journal of Document Analysis and Recognition (IJDAR) (To be published)