A combined strategy of analysis for the localization of heterogeneous form fields in ancient pre-printed records

Aurélie Lemaitre, Jean Camillerapp, Cérès Carton, Bertrand B. Coüasnon

▶ **To cite this version:**

Aurélie Lemaitre, Jean Camillerapp, Cérès Carton, Bertrand B. Coüasnon. A combined strategy of analysis for the localization of heterogeneous form fields in ancient pre-printed records. International Journal on Document Analysis and Recognition, 2018, 21(4) (269-282), 10.1007/s10032-018-0309-y . hal-01858192

HAL Id: hal-01858192
https://inria.hal.science/hal-01858192v1

Submitted on 20 Aug 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

International Journal on Document Analysis and Recognition (IJDAR)

A combined strategy of analysis for the localization of heterogeneous form fields in ancient pre-printed records

Aurélie Lemaitre · Jean Camillerapp · Cérès Carton · Bertrand Coüasnon

Received: 9 June 2017 / Accepted: 17 July 2018

Abstract This paper deals with the location of handwritten fields in old pre-printed registers. The images present the difficulties of old and damaged documents, and we also have to face the difficulty of extracting the text due to the great interaction between handwritten and printed writing. In addition, in many collections, the structure of the forms varies according to the origin of the documents. This work is applied to a database of Mexican marriage records, which has been published for a competition in the workshop HIP 2013 and is publicly available. In this paper we show the interest and limitations of the empirical method which has been submitted for the competition. We then present a method that combines a logical description of the contents of the documents, with the result of an automatic analysis on the physical properties of the collection. The particularity of this analysis is that it does not require any ground truth. We show that this combined strategy can locate 97.2% of handwritten fields. The proposed approach is generalizable and could be applied to other databases.

Keywords Historical documents · Field localization · Heterogeneous layout · Rule based system · Word spotting · Unsupervised clustering

1 Introduction

Nowadays, most of archive services have led to digitization of amounts of documents. More and more systems are proposed for handwriting recognition in ancient documents. However, it is not always necessary to recognize the entire content of a document page. Indeed, depending on the context of application, it is sometimes only necessary to localize a few specific fields that will be essential for document exploitation.

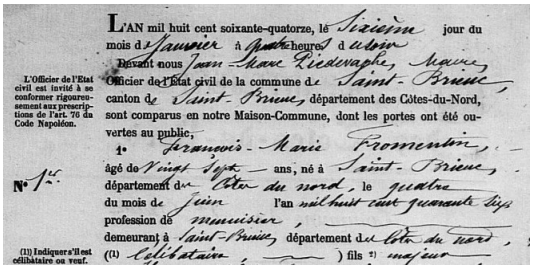
This paper focuses on ancient pre-printed form documents. Indeed, it is very common in archive documents to have some pre-printed forms that were filled in by employees, like census records, marriage, birth and death records, and many other kinds of registers. These documents are of great interest to historians or genealogists. In these documents, an automatic recognition of the full page is not always necessary. Thus, the analysis can focus on specific fields that are relevant for a given context of application.

The form field localization has been widely studied in contemporary documents. Thus, in well printed documents, it is now quite easy to apply a given template of the pre-printed form in order to automatically extract the handwritten characters that were filled in. However, concerning ancient pre-printed forms, several difficulties occur for the analysis, that make the topic still a research challenge.

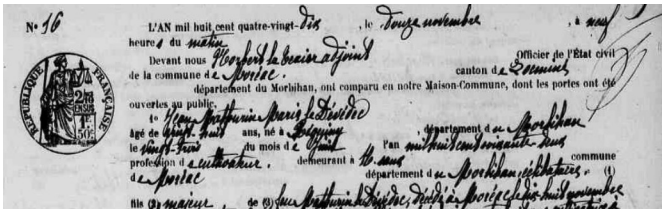
Firstly, we have to face with all the well-known degradation of ancient documents: damaged documents, bleed trough ink, pale ink... Secondly, the forms present a specific challenge: the great interaction between pre-printed text and handwritings. Thus, when the handwritings overlap printed text, it may decrease the performances of both printed text recognition and handwriting recognition. Lastly, the difficulty of ancient form is the large variety of physical layouts. For example, Fig. 1 presents two registers of the same type (French marriage records) but having a different layout: see for example the position of words "canton de", at the beginning of the fifth line on Fig. 1(a), or at the end of the fourth line on Fig. 1(b). In our work, we focus on forms

Univ Rennes - CNRS - IRISA
F-35000 Rennes
E-mail: aurelie.lemaitre@irisa.fr

having the same logical content inside of a collection, like on Fig. 1: the two forms present similar keywords, in the same order, even having different layouts.



(a)



(b)

Fig. 1 Two forms with a same logical content (French marriage records from 1874 and 1890) but a different layout: keywords for form fields are differently shifted.

The paper is organized as follows: it begins by a presentation of the chosen database and of its difficulties. Then, in section 3 we present related work. The global approach is presented in section 4, followed by the the pre-processing used in section 5. The main contribution of the paper consists in section 6 that compares strategies of analysis and demonstrates the interest of a combined strategy. The results presented in section 7 validate this comparison.

2 Database and objectives

As an example of application, we use a public database that has been proposed for the Family Search HIP 2013 Competition, in relation to ICDAR 2013 [3]. This database represents a problem from real life, of which results could be helpful to genealogist, in a context of assisted transcription of contents.

2.1 Position with HIP 2013 competition

The competition proposes to study some Mexican marriage records from the 20th century (Fig. 2). They are pre-printed forms, containing handwritten text. The goal of the competition is to localize four regions of

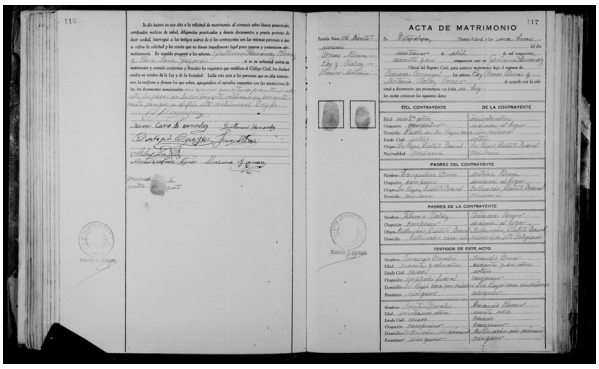


Fig. 2 Example of a Mexican marriage record

interest in the documents: the month of the record, the year of the record, and the origins of groom and bride. Those four regions were chosen because of their interest for genealogists. The competition also evaluated a clustering on handwritten words: the objective was to group the text fields having the same content.

Thus, the competition merges two scientific challenges: the localization of handwritten fields in heterogeneous forms, and the clustering of handwritten words. We think that both tasks are interesting, but the first task of field localization requires an appropriate method. That is why this paper only focuses on the *localization of handwritten fields*, and more particularly on the localization of two fields: the month and the year of the record (Fig. 3).

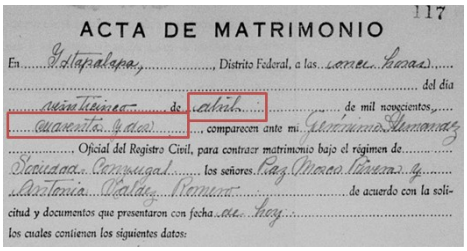


Fig. 3 Studied task: localization of month and year handwritten fields

2.2 Database and ground-truth

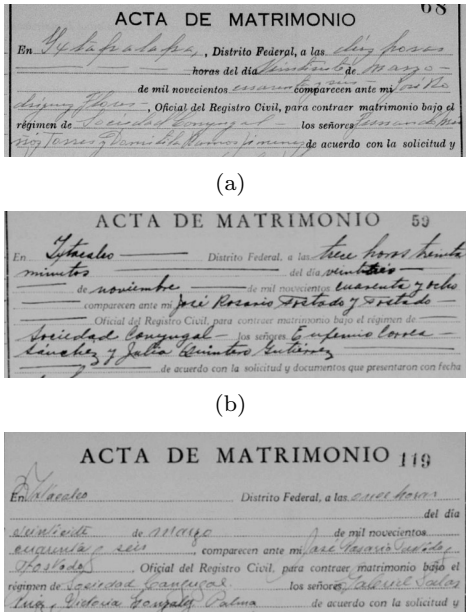
The database of HIP competition [3] contains two sets of images: 10,490 images for training and 20,000 images for the competition. The big size of this database makes it particularly interesting.

The ground-truth that was given for the competition is made of the text transcription of the regions of interest for each document (month, year and origins). This ground-truth is available for the 10,490 pages of

the training set. This ground-truth also exists for the evaluation of 10,000 out of the 20,000 images of the evaluation set, but it was not provided to the participants. However, there is not available ground-truth concerning the position of the searched fields, inside of the documents. This situation corresponds to a real-case context when no ground-truth is available for field localization.

2.3 Difficulties of the database

The documents present several scientific challenges, illustrated on Fig. 4.



(a)

(b)

(c)

Fig. 4 Difficulties of the corpus: ancient documents with varying contrast in ink, interaction between handwritings and printed text, variety of pre-print forms (see the various positions of word "del dia" on the second line).

Firstly, we have to take into account the difficulties that are related to ancient and damaged documents. The documents are not always printed with the same quality of ink, and sometimes the printed text is very pale (Fig. 4(b)). Moreover, the conditions of digitization are not constant. Indeed, we have noticed that even if the resolution level is supposed to be 300 dpi for each image, in reality, the distance to the camera during the acquisition makes some variations in the dimensions of the final image. Thus, an analysis of physical properties (position, size) may require a normalization process of the images, without knowing the exact resolution.

Secondly, these documents present a strong interaction between the pre-printed form and the handwritten text. Thus, the extraction and the recognition of printed keywords are made much more difficult due to the overlapping with handwritten characters (Fig. 4(a)).

At last, we have to deal with the heterogeneous aspect of the database. Indeed, the collection is extracted from various registers and from various cities of Mexico. Depending on the year and on the city, the pre-print form does not have exactly the same aspect: font of title, dimensions of form fields. Figure 4 presents three forms with different pre-prints. See for example the position of the keyword "del dia" (day) on the second line, and the size of its associated blank field to write the date just after: there is a large variation between the three examples that prevents from directly applying a template for field localization. Moreover, as the images are provided in a deliberately random order, it is not possible to exploit information coming from the previous or the following image from the same register, in order to improve the recognition of one image. We have no idea of the extent of the possible variations of registers. Our process must be generic enough to absorb all the possible variations.

The challenges for the study of this database are:

- the difficulty to extract printed keywords, due to the interaction with handwritten text,
- the variability of the pre-print forms, and the lack of knowledge about the extent of this variability.

Moreover, as we are in a real context, no ground-truth is available for supervised automatic learning.

In this paper, we apply our work on this specific database, but it is important to see that the two difficulties listed above are common for any pre-printed ancient records. Thus, the presented approach could be applied to other pre-printed ancient records.

3 Related works

We present here how related works address the two difficulties: the various layouts of pre-printed forms and the interaction between printed and handwritten text.

3.1 Deal with various layouts

The handwritten field extraction in printed document is a common problem in recent documents, such as bank cheques. Bank cheque processing can be seen as an easy task of form analysis. Thus, Jayadevan *et al.* cite in their survey [11] various approaches used for the step of extraction of handwritten fields. When the structure is well known, like for example American bank cheques,

some rule based systems are applied to find long horizontal lines, dashes, slashes, using relative sizes, shapes and positions. In many other countries, a database of well-known templates is used to automatically extract the areas of interest. This is the most efficient strategy but it requires an exhaustive knowledge of all the possible templates. Liang *et al.* [16] also suggest learning a model for each template: they propose a method based on graph matching to link the physical and the logical structure of the document. The physical analysis of the document is based on bounding box position, size and font size. They conclude that their method is well adapted for simple contents but it requires optimization to deal with noise and variant data.

In ancient documents, the template cannot always be clearly detected, it is often more irregular [24]. Some techniques must be used to select the good template. Nielson and Barrett [20] apply this principle. They segment ancient table forms into regions of interest. Their work is based on the detection of table lines, in order to cut the form into isolated cells. However, due to the ancient nature of the documents, they must face with degradation of the documents and the lines are not always present in all the documents. Then, they propose to merge the detected lines in several documents having the same layout in order to obtain a common template by consensus. This is a typical application in which the degradation of the document, and the presence of interactions between handwritten text and form lines are solved by the analysis of a wide set of data in order to learn the physical layout. However, this requires the presence of a stable physical layout.

In the same idea, several methods propose to exploit information coming from the analysis of other documents in the same collection. This approach is used by Coüasnon *et. al* [6] for the tabular layout recognition of military forms. It is also exploited by Mas *et al.* [18] for census record transcription, while mastering the role of human in the loop [10] [18]. This is a very interesting strategy, but it requires that the documents are provided in a regular order, which is not our case in the studied database.

When the layout varies too much, the systems can be based on the extraction of small physical clues, in order to isolate the layout entities. This is the approach proposed by Garz *et al.* in [9] for layout analysis of handwritten historical manuscripts. They use SIFT features to describe the visual aspect of layout entities such as main body text, headings or initials. The context is similar for the heterogeneous Maurdor database, in which it is required to separate printed, handwritten text, forms, graphics and tables, without knowledge on the layout variety. For that purpose, Barlas *et al.* [2] also propose a local analysis based on connected components. Then, they build code-books that enable connected components classification by training a MLP. This system can be used when no *a priori* knowledge is available neither on the physical organization nor on the logical content of the document, but is does not enable to build a high level layout organization.

Concerning forms, as synthesized by Ye *et. al.* [27], we can consider two main types of forms, according to the rigidity of their structure: the *rigid form*, in which the physical information such as positions and sizes of the fields remain stable; and the *flexible form* where the items may appear in different locations while preserving certain important logical structures. Kooli *et al.* [12] work on entity recognition in flexible forms. They demonstrate the interest to combine two sources of data: the knowledge of the entities stored into a database, and a structural modeling of entities inside of the document. The physical description of the organization of fields over several text-lines enables to improve the entity recognition. This is an intersering idea, but this work is dedicated to printed business documents.

In the FamilySearch database, the layout can be considered as *flexible form*, as the layout varies but not the logical keywords. For HIP'2013 competition on this database, two other methods were submitted. Adam *et al.* propose in [1] a very flexible registration based on qualitative positioning and document logic, called geometrical regular expression, Gregex. This system is required to deal with local shift of pre-print information. In the second method, presented in [22], the field localization is also made by a rule based system. In this specific context, as no information is given neither on the number of available layouts, nor on the collection order, we are convinced that the most efficient approach is to use the logical stability of the models. This requires extracting printed keywords inside of handwrittings.

3.2 Deal with printed and handwritten content

The separation of handwritten and printed text in forms is particularly complex in the case of ancient documents. A favorable case is one where the physical layout is extremely well known. That is the case for example with census records presented in [25][21]. In these works, some threshold parameters are specifically dedicated to the layout of the table forms. This enable to easily extract the physical layout, and the work can then concentrate on handwritten extraction. This is also the case in the work proposed by Richarz *et al.* [23]. They focus on transcription of handwritten historical weather reports. In these documents, the tabular layout is very stable. Even if it may interfere with

handwritings, it can be easily detected using well known templates. However, this implies a big stability of the physical layout of the collection.

Ye *et al.* [27] insist on the difficulty brought by the interactions between handwritings and printed text in forms. They explain that the logical structure, or even the physical structure is often well known, but the problems comes from the fact that handwritten data touch or cross the pre-printed form frames and text. Thus, they propose a method for text cleaning and enhancing. The main limit of their approach is that it requires the knowledge of a blank form that has exactly the same structure as the analyzed document.

In the field of historical registers, Stewart *et al.* [26] propose to separate the different layers of a document (printed text, dotted line, handwritings) using a fully convolutional neural network. Their method provides a segmentation at pixel level, but it also requires a laborious labeling phase at pixel level to train the system. A training phase is also required for all the recent approaches based on deep learning, like object localization or structure recognition [19]. This requires to label ground-truth data.

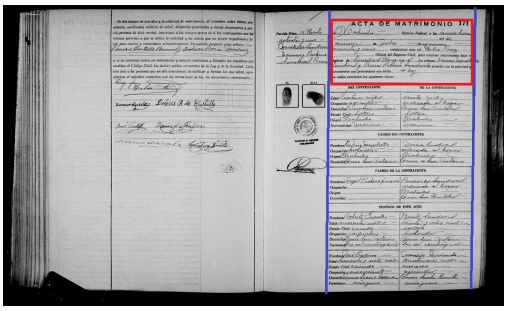
Concerning the FamilySearch database, Adam *et al.* mention in [1] the difficulty to localize printed text, due to the bad quality of the ink and the interaction with handwritings. Thus, they propose to use several text anchors, in order to make the registration process more precise and more resistant to noise. They independently apply several local binarizations to increase the success rate. The approach proposed in [22] is also based on salient keywords, such as "de".

Those approaches demonstrate that when the interaction between printed text and handwritings is too strong, the usual word-spotting or text classification methods are not sufficient to localize handwritten fields. It is necessary to combine several text anchors, or even to know or learn the layout model, using the layout repetition in the collection. In the context of marriage records, we do not know the number of possible layouts, no ground-truth is available for a supervised training, and the images are not provided in a regular order, so it is not possible to exploit neighbours in the collection. We propose to learn the various possible layouts using an unsupervised analysis, by studying the repetitions on a big amount of data. As this training will necessarily be imperfect, we will improve the detection using the expression of logical rules on the content.

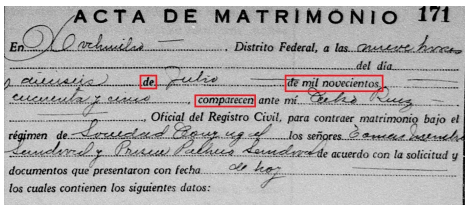
4 Overview of the approach

In order to easily express knowledge on the logical organization of the document, we choose to use a rule based system for document analysis (like the one presented in section 7.1). The global strategy is illustrated on Fig. 5.

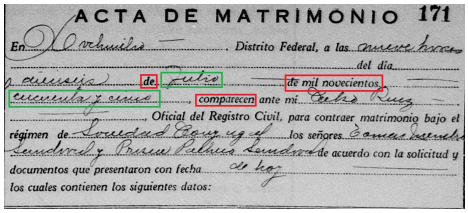
1. In the whole page, find the position of the right column of the act in the whole page, and focus on the upper part of the column (Fig. 5(a)).
2. Look for the printed keywords that delimits the searched fields: "de" and "de mil novecientos" for the month; "de mil novecientos" and "comparecen" for the year (Fig. 5(b)).
3. Build the handwritten text field between the printed keywords: this step requires an appropriate strategy of analysis (Fig. 5(c)).



(a) Localization of the right column of the act (in blue), and the upper part of this column (in red)



(b) Localization of the printed keywords (in red)



(c) Extraction of handwritten fields (in green) between printed keywords

Fig. 5 Global mechanism of handwritten field localization.

The two first step of analysis enable to find the printed keywords on the top right of the documents (Fig. 5(b)).

They are considered as a pre-processing, detailed in the following section. Then, the difficulty is to apply a generic enough strategy to determine the position of handwritten fields (Fig. 5(c)). This third step is challenging as it must face with the two difficulties previously identified: the difficulties to extract keywords in a confusing environment and the large variety of pre-prints. We will compare in section 6 four proposed strategies of analysis to solve these difficulties.

5 Pre-processing: find printed keywords

5.1 Localization of the right column

For the localization of the right column, we apply a rule based method. As input, we provide some line segments that are extracted inside of the documents thanks to a method based on Kalman filtering [15], applied to the original grey-level image. It provides many candidates, presented on Fig. 6. Then, a rule based method selects the most convenient line segments, which correspond to the edges of the right column (Fig. 5(a)).

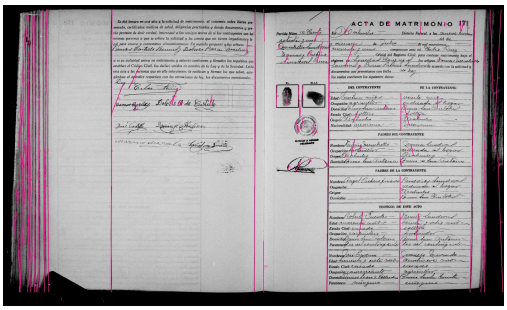


Fig. 6 Input primitives: vertical line segments that are used to detect the position of the right column (figure 5(a)).

5.2 Localization of the printed keywords

Once the right column has been detected, the second step consists in the detection of the printed keywords which are required for the localization of the handwritten text. At a minimum, three keywords are required: "de", "de mil novecientos" and "comparecen" (Fig. 5(c)). Note that some of the keywords like "de" are present several times in the zone of interest of the document.

In order to have a bigger expressiveness in our rules, we also look for some neighbor keywords. Thus, checking the presence of several keywords in the good order can bring interesting information on the context. So we detect 8 keywords in the documents: "Distrito", "Federal", "del dia", "de", "de mil novecientos", "comparecen", "Oficial", "Registro" (Fig. 7(a)).

For the extraction of the printed keywords, we first use an existing commercial OCR. But due to the bad performance of this OCR, we apply a dedicated method based on word spotting. We detail these two approaches.

5.2.1 With commercial OCR

We apply the commercial OCR Abbyy Fine Reader in the zone of interest of the documents, and look for the interesting keywords in the transcribed text. We have no ground-truth of the exact transcription of the text zone, but we know that the eight keywords should be extracted at least one time in each document.

In order to roughly estimate the quality of the extraction, we apply the OCR on 7,000 documents. For each keyword, we evaluate the number of documents in which no occurrence of the keyword is found. It does not mean that the detected keywords are found at the good position, but when no keyword is found, we cannot infer any. The results are shown in the first column of Table 1.

This first study of the quality of the OCR shows that in sometimes up to 50% of the documents, the interesting keywords are not found at all. This is due to the high interaction between handwritten text and printed text. The OCR often considers our zone of interest as an image and does not propose any transcription of the textual content. In this condition, it appears that we have to find another solution for the keyword extraction, based on word spotting.

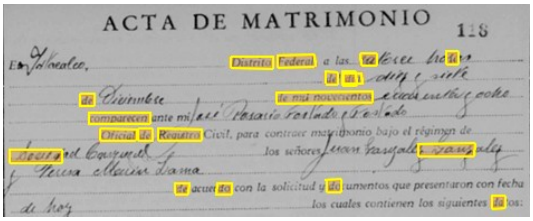
Word	OCR Abbyy Fine Reader	Word spotting based on POI
Distrito	38.6%	1.5%
Federal	32.8%	1.6%
del dia	34.2%	0.8%
de	42.0%	0.1%
de mil novecientos	58.1%	4.0%
comparecen	55.6%	0.1%
Oficial	51.8%	0.2%
Registro	46.8%	7.5%

Table 1 Percentage of images (over 7,000 test images) in which no occurrence of the searched word is found, whereas it should be: the keywords are too often missed with OCR (lower is better).

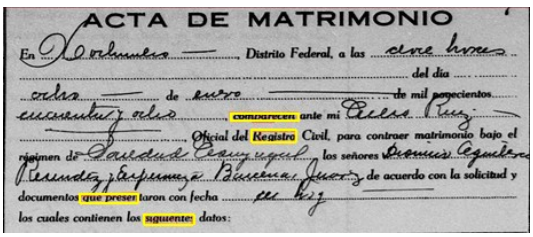
5.2.2 With word spotting

We use a method based on the arrangement of local descriptors (points of interest), adapted from the method of SIFT proposed by Lowe [17]. The main objective of the Points Of Interest (POI) is to select a small set of points of the image, which present some interesting local variations of luminosity. Some details on the use of POI are given in [13]. We manually build 2 or 3 models of arrangements of POI for each of the 8 keywords that we want to search in the zone of interest. We then use these models for a step of word spotting in order to detect the 8 keywords in each document. The last column of Table 1 gives comparative results with OCR. This method based on word-spotting seems to obtain better results than a commercial OCR, even if we cannot evaluate the false alarm computed by the method. However, as keywords are found in most images, data is available for further processing.

(a) Ideal case: 8 printed keywords are detected



(b) False positive keywords detected inside of handwritten text



(c) Some keywords are omitted due to the interaction with handwritten text: for example "de mil novecientos"

Fig. 7 Example of the printed keywords detected with POI (in yellow): the detection is not perfect

A qualitative analysis shows that even with this method, the extraction of keywords remains a difficult task (Fig. 7). In the ideal case, 8 keywords should be detected. Sometimes, many false positive keywords are detected, even inside of handwritten text. On the opposite, the interaction between printed and handwritten text may prevent from finding some printed keywords.

6 Strategies for handwritten field localization

The extracted printed keywords are supposed to be used for the localization of the handwritten fields. The field for month is between the keywords "de" and "de mil novecientos" and the field for year is between the keywords "de mil novecientos" and "comparecen".

However, we have to face two problems:

- the keywords are not always correctly detected. What can we do when a keyword is not found ?
- for building the handwritten fields, what does it mean to be located "between" two keywords ? We have to deal with a variety of pre-prints which implies that sometimes a handwritten field is located as the end of the line, sometimes over two lines, sometimes on the line under the previous keyword.

Consequently, handwritten field localization is not so obvious, and this is the main goal of our contribution in this paper. We study four strategies, and their contributions to the two difficulties presented above.

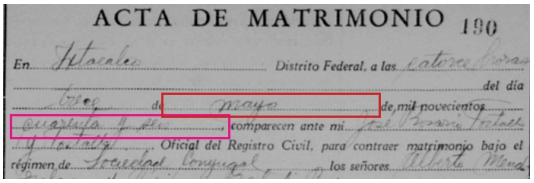
6.1 Empirical analysis of the corpus

As we are convinced by the interest of syntactical methods for the description of document layout, we propose a first method, called *empirical* that is made of a *grammatical description of physical layout* of documents. This first version has been proposed for the Family Search competition in 2013 [13].

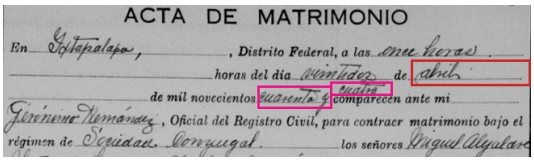
We had a manual look at a small part of the collection (about 100 examples). We manually identified four big families of pre-prints: A, B, C, D. They distinguish the different configurations of the position of year and month in the document. Table 2 synthesizes these models that are illustrated on Fig. 8. For example, the year is sometimes on the 3rd line, on the 4th line, on both lines and even between two text lines. This analysis of the corpus is totally empirical, as we do not know if all the pre-print forms have been identified by our manual analysis and can match with one of the four models A, B, C or D.

Model	Year position	Month position
A	beginning of 4th line	middle of 3rd line
B	middle of 3rd line	end of 2nd line
C	end of 3rd line, and beginning of 4th line	middle of 3rd line
D	end of 3rd line	end of 3rd line

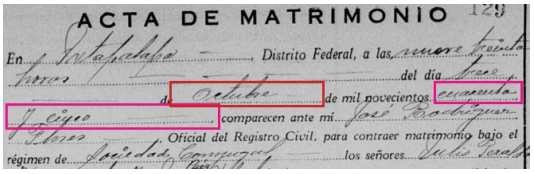
Table 2 Four models of registers, manually identified, with different configurations of text position



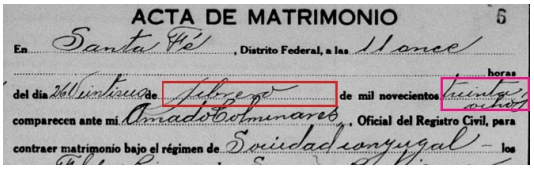
(a) Model A



(b) Model B



(c) Model C



(d) Model D

Fig. 8 Example of the four identified models, with different configurations of text position, described in Table 2 (month in red, year in pink)

These four models (A, B, C, D) are used to build a rule-based system. For the recognition step, each document is supposed to match with one model, by application of rules like algorithm 1. Each specification of a position, like "on the right" is manually tuned and requires several parameters.

This way to build a recognition system may seem very rough. However, faced with a simple problem, it is not always necessary to develop complex solutions. This method has been used for the Family Search competition in 2013. We will show in section 7.3 that it enables to localize 94.2% of fields (Figs. 11(b), 12(b)).

Algorithm 1 Rule to find the position of the year

```
Look for "de mil novecientos"
if "de mil novecientos" is on the middle then
    build year on the same line (model B)
else if "de mil novecientos" is on the right then
    build year on the line under (model A)
else if "de mil novecientos" is on the right, but not
too much then
    Look for "comparecen"
    if "comparecen" is at the beginning of a line then
        build year on the same line as "de mil novecien-
        tos" (model C)
    else
        build year on two lines (model D)
    end if
end if
```

Let us discuss the properties of this method, in relation to the two difficulties raised in introduction.

First, this empirical method has a small ability to deal with missed keywords. Due to the physical rule based description of four hard models, when a keyword is missed, its position might be inferred. For example, the position of the month can be roughly built even if the keyword "de" is not found (Fig. 13(b)). However, when the keyword "de mil novecientos" is not found, it is a most important problem as this keyword is used to choose the appropriate model, as shown in algorithm 1.

The empirical method is not able to deal with variable pre-prints. Thus, as said previously, we are not sure that we have identified all the pre-print forms during the manual analysis of the collection. So, position settings that were manually tuned might be inappropriate with a new form template.

6.2 Logical description of the content

In order to make more generic the previous grammatical description, we proposed another one called *logical* that is a *grammatical description of the logical content* of documents. This grammatical description aims to exploit the logical stability of the content of the documents and to avoid manually setting of the position parameters.

The analysis is based on the logical organization of the text. We look for the succession of keywords, without taking care on their physical position. The description of a document is presented on Fig. 9.

This description considers that the textual content is made of a succession of elements: printed keywords and text zones. Two of the text zones are the zones of interest which are required as a result: the month and

```
textAct ::=
     keyword "del dia" T1 &&
          AT(afterText T1) &&
     zone "day" T2 &&
          AT(afterText T2) &&
     keyword "de" T3 &&
          AT(afterText T3) &&
     zoneOfInterest "month" T4 &&
          AT(afterText T4) &&
     keyword "de_mil_novecientos" T4 &&
          AT(afterText T5) &&
     zoneOfInterest "year" T6 &&
          AT(afterText T6) &&
     keyword "comparecen" .
```

Fig. 9 Logical description of the textual content

the year. This logical description enables to build the searched zones by only studying the input positions of the found keywords. Note that here we use the keyword "del dia" in order to give more context to the analysis.

The physical constraints are isolated inside of the position operator `afterText`. Indeed, this position may take several values, depending on the position of the previous word: the following text can be either on the right, or on the line under. We also accept that the `zoneOfInterest` may be split over two lines. No manual parameters are required, unlike to the empirical method.

We will see in section 7.3 that this method obtains 88.4% recognition rate (Fig. 11(c)).

As this method is based on the logical description of the textual content, it can easily deal with the large variety of pre-prints. Thus, it does not need any specific knowledge to absorb a new kind of pre-print form, assuming that this form follows the logical organization of printed keywords. The description proposed on Fig. 9 can work regardless on the physical organization of contents (Fig. 12(c)).

However, the zones of interest are entirely built on the positions of the found keywords. So, when a keyword is missed, it is neither possible to infer it, nor to generate the zone of interest (Fig. 13(c)).

6.3 Automatic analysis of the corpus

To deal with the case of undetected keywords, we propose to extend the empirical method by creating a *rule based system for all the models of forms* present in the training set. Thus, the recognition will consist in applying the best template. However, as said previously, we do not have any knowledge on the number of possible pre-print forms, and there is no available ground-truth for learning. So this third method realizes an *automatic analysis of the corpus* in order to identify the different available pre-print layouts.

6.3.1 EWO method

We use the Eyes Wide Open (EWO) method [4]. EWO is a system to provide an exhausted view on a corpus in order to automatically extract properties on the content. The extracted properties are used as input of a step of clustering, based on EAC (Evidence Accumulation Clustering [7]). This clustering does not require a prori knowledge neither on the number of clusters nor on their size, and it does not need any free parameter. The unsupervised clustering automatically defines some classes, based on criteria chosen by a user. Note that EAC clustering allows to detect classes even with few elements. EWO then generates rules that are based on the physical properties of each class, for the recognition of documents. More details on EWO method are given in [4].

In this work, EWO is used for two stages of analysis:

- the construction of a pseudo ground-truth, with automatic clustering, validated by the user;
- the unsupervised analysis of the corpus to compute the possible models of documents.

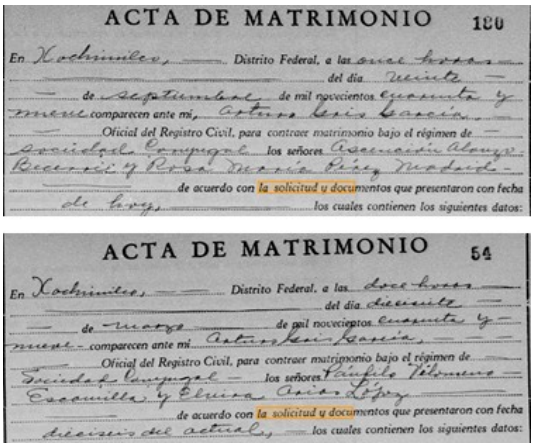
6.3.2 Obtaining a pseudo ground-truth

In our context, the goal is to automatically detect the number and the kinds of pre-prints in the training set.

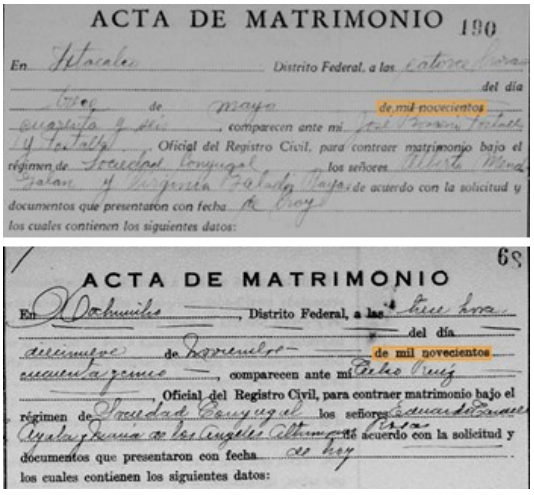
As input, EWO takes the keywords extracted by the Points Of Interest (section 5.2.2). The analysis is based on 8 keywords (Fig. 7(a)). As shown in Fig. 7, the obtained result is not ideal: some keywords are missed, some false positive are found. However, the large amount of images (7,000) provides sufficient redundancies for keywords analysis. EWO method then automatically produces some clusters of keywords, based on their position, their label and their dimension. This clustering is made using Evidence Accumulation Clustering [7].

As output, for each keyword, EWO proposes clusters of similar keywords at the same position. EWO system then requests a user interaction: for each cluster of keywords, a few examples are presented to the user, who has to validate or not the whole cluster (Fig. 10).

After the user interaction, we obtain a pseudo ground-truth of detected keywords. Thus, the user has validated entire clusters with only few examples of similar images. The ground-truth is maybe not totally reliable but the large amount of data makes it sufficient



(a) Cluster that will be rejected by the user



(b) Cluster that will be accepted by the user

Fig. 10 Interaction with the user for the selection of reliable clusters. Example of two clusters presented to the user for the keyword "de mil novecientos". By seeing only two representative samples, the user can choose to accept or reject the extracted keywords.

to be used as an input for a learning process. This pseudo ground-truth is not complete: when keywords were not detected by POI method, EWO does not produce ground-truth.

6.3.3 Computing possible models

The learning process aims at detecting all the kinds of pre-prints in the training set. We consider that a pre-print is characterized by the position of the eight founding keywords. We focus on the documents in which the eight keywords have been validated in the pseudo ground-truth, which represents 5,406 documents over the initial database of 7,000 pages. We study the relative position of the eight keywords: this constitutes the signature of the document. Then EWO method is used to build clusters of documents having the same signature, again with EAC clustering.

As a result, the EWO method automatically detects 11 kinds of pre-prints in the training set. For a comparison, let us remember that we assumed in the empirical version (section 6.1) that there were only 4 kinds of pre-prints. We can also mention the fact that the data is unbalanced for the 11 classes: the most common class appears for 27% of documents, the less common appears only 0.5%. This explains why a manual leaf through the collection may not meet all the models.

6.3.4 Field localization

Once the 11 templates of documents are known, we have to build a rule based description of these documents. This is automatically produced by EWO method that is able to infer position operators [4]. Those positions are similar to the ones that were manually tuned in the empirical method (section 6.1).

During the document analysis, we then just have to apply the best fitting rule, out of the 11, depending on the position of the extracted keywords. The best fitting rule is computed taking into account the best combination of keywords matching by overlapping with models. This system is able to produce a result even if not all the keywords are found. More details on the use of EWO for this database are given in [4].

6.3.5 Discussion

We will see in section 7.3 that this learning based method obtains a recognition rate of 96.8%.

This method is based on an automatic learning of all the kinds of pre-prints. Consequently, it is able to deal with many pre-prints, assuming that these pre-prints are present in the training set. This is the case for the examples of Fig. 11(d) and Fig. 13(d). But, when a template is not available in the training set, like example of Fig. 12(d), the results can not be properly inferred.

Concerning the ability to deal with missed keywords: as this method is based on the application of the best template on the document, if a keyword is missed, it can be inferred by its expected position on the template. This is the case for Fig. 13(d).

6.4 Mixed method

The *learning based* method presented above obtains very good results, even if some keywords are missed, but it requires that the kind of pre-print form is present in the training set. On the opposite, the *logical* method

presented in section 6.2 is able to deal with any pre-print form, assuming that enough keywords are found. That is why we propose the fourth approach, called *mixed* method, which combines both logical and learning based methods.

The analysis of a document follows the presented strategy.

If there are enough found keywords to build the month and the year field, we apply the logical description of the content for that purpose. Enough keyword means that at least the three keywords around the searched fields must be found ("de", "de mil novecientos", and "comparecen"), plus 2 others keywords over the 5 remaining, in order to confirm the logical position of elements. In that case, the system takes the result of logical method presented in section 6.2.

Otherwise, the system uses the rules of the learning based method, as presented in section 6.3.

By automatically selecting the most appropriate method, this mixed approach enables a fusion of knowledge between a logical description of the content, and an automatic analysis of the collection. It enables to deal with the two difficulties of the database: the possible miss-detection of keywords and the large variety of pre-prints. Indeed, if a pre-print has not been detected by the learning based method, it can be treated by the logical description (see example Figs. 11, 12). If too many keywords are missing, they will be inferred by the matching of a model resulting from the learning method (see example Fig. 13). The mixed method enables to deal with reject cases of each approach (learning based or logical) with the other one.

7 Experiments

We will now detail how we have implemented and evaluated the four strategies presented in the previous section.

7.1 Context of implementation

We have implemented our rule based strategies using an existing method DMOS-PI [5][14]. DMOS-PI is a grammatical method for the recognition of structured documents. It is based on a bidimensionnal grammatical formalism, EPF (Enhanced Position Formalism), which enables a physical, syntactical, and semantic description of the content of the document. For each kind of document to recognize, once the recognition rules have been expressed, the associated parser is automatically produced by a compilation step.

This method has been validated for the analysis of various kinds of documents: tabular, archive documents, mathematical formulae, heterogeneous documents... and at a large scale (more than 700,000 images). It is particularly adapted to our application as we want to compare some rule based strategies. Nevertheless, the strategies presented in this paper could have been implemented in another language.

For each strategy of analysis, we create a description of the documents that includes: the pre-processing for the localization of the right column, the integration of extracted keywords with POIs, the rules for the description of each strategy.

7.2 Ground-truth and metric

We evaluate the ability of each strategy to find the correct position of the handwritten fields, inside of the pre-print forms. We created a ground-truth, containing the position of the two handwritten fields (month and year) for a test subset of 2,000 images. Each field may be composed of one or two bounding boxes when the field is straddling two lines. This ground-truth database can be downloaded for free in our website[1].

Concerning the metric, we need to evaluate the spatial correspondence of the zones produced by our methods with the ground truth zones. We choose the metric proposed by Garris in [8] which evaluates the surface of zone which has been correctly detected. This method requires applying a threshold, depending on the field of application. Due to our context of application, we choose the following thresholds:

- a **field is totally recognized** if at least 95% of its width and 75% of its height matches,
- a **field is partially recognized** if 1) it is not totally recognized, 2) at least 80% of its width matches and 75% of its height,
- a **false positive** is a field that does not correspond to a ground-truth,
- a **document is recognized** if all its fields are totally or partially recognized.

7.3 Results

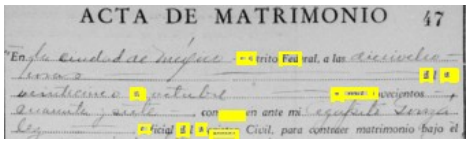
We evaluate the four strategies of analysis using the 2,000 manually annotated documents of the database, and the metric presented in the previous section. Some examples of recognition are proposed in Figs. 11, 12, 13.

Table 4 summarizes the qualitative estimation of the capacities of each method. The quantitative results, presented in Table 3, correspond to this estimation.

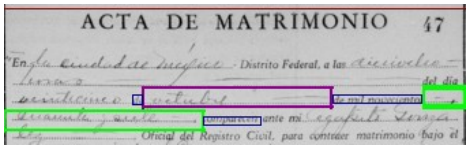
[1] https://www-intuidoc.irisa.fr/hip_db/

Method	Number of defined models	Total field recognition	Partial field recognition	Total+partial field recognition	False positive	Document recognition
Empirical	4 manual	90.2%	4.0%	94.2%	5.8%	80.2%
Logical	none	88.4%	0%	88.4%	9.8%	76.6%
Learning	11 automatic	90.1%	6.7%	96.8%	4.7%	91.2%
Mixed	11 automatic + none	**92.6%**	4.6%	**97.2%**	3.8%	**92.2%**

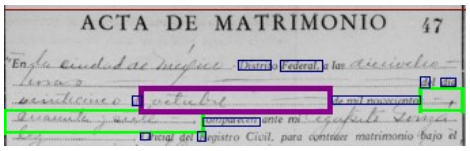
Table 3 Comparative results of the four approaches on a test set of 2,000 images. The mixed approach that combines the 11 learned models and the logical description obtains the better results



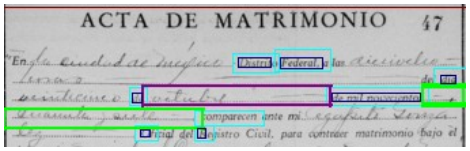
(a) Extracted keywords



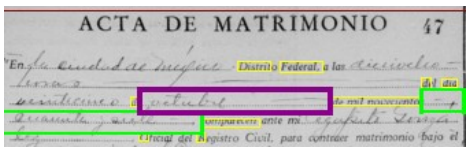
(b) Empirical method: correct month (in purple) and year (in green)



(c) Logical method: correct month (in purple) and year (in green)



(d) Learning based method: correct month (in purple) and year (in green)



(e) Mixed method: correct month (in purple) and year (in green)

Fig. 11 Example 1: the correct keywords are found, and it is a well known model. The empirical method succeeds with model A. The logical method succeeds as the keywords are present. The learning based method well matches a known model. As enough keywords are present, the logical method succeeds based on the logical version.

(a) Extracted keywords

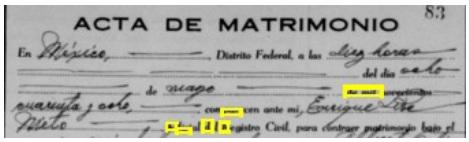
(b) Empirical method: correct month (in purple) and year (in green)

(c) Logical method: correct month (in purple) and year (in green)

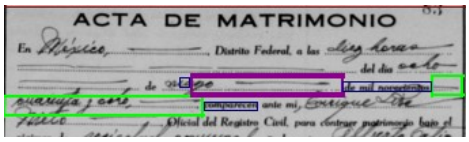
(d) Learning based method: missed month and year

(e) Mixed method: correct month (in purple) and year (in green)

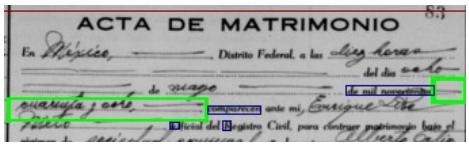
Fig. 12 Example 2: the correct keywords are found, but it is a rare layout. It corresponds well to model D of empirical method, which succeeds. The logical method succeeds with 6/8 found keywords. However, the model has not been learned by unsupervised analysis of the corpus: it matches with the nearest model that is indeed very far. The mixed method succeeds with logical version.



(a) Extracted keywords



(b) Empirical method: imprecise month (in purple), correct year (in green)



(c) Logical method: missed month, correct year (in green)

(d) Learning based method: correct month (in purple) and year (in green)

(e) Mixed method: correct month (in purple) and year (in green)

Fig. 13 Example 3: some keywords are missed at the beginning of the act. The empirical method infers the "de" keyword, using parameters of Model A with an imprecise position. The logical method cannot take a decision concerning the position of the month. The learning based method uses the few detected keywords to match with the best model that enables to correctly infer the position. The mixed method uses the learning based version.

Method	Ability to process missed keywords	Ability to process variable pre-prints
Empirical	+	-
Logical	-	++
Learning	++	+
Mixed	++	++

Table 4 Qualitative evaluation of the properties of the 4 strategies

The second column of Table 3 recalls the number of model descriptions that are used by each method. The *empirical* method, based on 4 manual models, obtains a document recognition rate of 80.2%. This is due to the fact that the 4 manually chosen models do not cover the whole database. The *logical* method is not dedicated to specific physical layouts. But, it is sensitive to the bad extraction of keywords, which explains its weak score of 76.6% at document level. The *learning based* method enabled to automatically extract 11 models, with unbalanced classes. It obtains 91.2% recognition rate at document level. The remaining errors come from models present in the training set but not with a good enough quality for automatic clustering.

The best results are obtained with the mixed strategy, which combines the logical description of the content and an automatic analysis of the collection. With this mixed method, we partially or totally localize the fields at 97.2%, with recognition at document level of 92.2%. Indeed, this method can use the logical description to deal with the images that are rejected by the learning based method. The remaining 2.8% errors are related to poorly detected printed keywords. In this case, we chose to generate the most appropriate fields, but it would be possible to ask the rule system to reject the image under a given threshold of matching. For the HIP Family Search competition, the empirical method has been used [13] (it was then the only available method). We want to remind here that it is not possible to compare with the other approaches that were submitted for the competition, as the field localization is an intermediate task that was not evaluated by the competition metric.

8 Conclusion

In a real context of non-annotated documents, we proposed to use rule-based systems for field localization in marriage forms. The constraints are the difficulty to extract keywords due to strong interaction between handwritten and printed text; and the great variety of layouts in the collection, with the presence of rare models.

At first, we developed an *empirical* system with parameters set by hand. This rough and simple approach enables to localize 94.2% of fields, which demonstrates the interest of using a rule based system. It has been used for HIP competition. However, it is not satisfying to tune parameters by hand.

We then proposed a *logical* approach that totally relies on extracted keywords, in a logical order. This approach does not contain specific parameters but may fail when all the keywords are not correctly detected.

Consequently, we proposed a third *learning* based system. It uses specific position parameters, which are learned. No ground-truth was available so we used the Eyes Wide Opens (EWO) method to build a pseudo ground-truth by an unsupervised analysis of a big amount of data, with a validation of the clusters by the user. This pseudo ground-truth was used to automatically produce a rule based system, with learned parameters that describes each possible layout of the forms in the learning database. This method obtains 96.8% recognition. Its strength is to work without ground-truth, as it uses the study of redundancies on a big amount of data. However, the inferred rules cannot deal with new layouts.

The paper demonstrates the interest of the fourth strategy, called *mixed*. This is a combined strategy that proposes a fusion of information between two kinds of data: the logical knowledge on the organization of the form, and the results of an automatic analysis on the physical organization of data, on a big amount of document, without ground-truth. This strategy enables to deal with the two difficulties of the database: the difficulty to extract physical information, due to the high interaction between printed and handwritten text, and the variation of the physical layout, with a stable logical organization. The strength of the mixed strategy is that the logical method can deal with the reject cases of the learning based method and vice versa. The quantitative results confirm the interest of this method, as 97.2% of fields are totally or partially recognized, on a test set of 2,000 images.

This method has been applied to a publicly available competition database, FamilySearch Mexican marriage records. However, it can be applied on other collections of ancient pre-prints having the same properties: difficulty to extract physical templates, lack of knowledge on the number of possible different layouts. Indeed, no ground-truth is required with the proposed approach, so it can be applied to other similar real cases.

References

1. Adam, P., Knibbe, M., Bernard, A.L., Mtaireau, P.Y.: ICDAR 2013 HIP Workshop FamilySearch Competition A2ia Submission. In: Historical Image Processing (HIP) (2013)
2. Barlas, P., Adam, S., Chatelain, C., Paquet, T.: A typed and handwritten text block segmentation system for heterogeneous and complex documents. In: DAS'14 (2014)
3. Cannaday, A.B., Gehring, J.: ICDAR 2015 HIP Workshop FamilySearch Competition Capstone Summary. In: Historical Image Processing (HIP) (2013)
4. Carton, C., Lemaitre, A., Coüasnon, B.: Eyes Wide Open: an interactive learning method for the design of rule-based systems. IJDAR **63**, 411 – 411 (2017)
5. Coüasnon, B.: DMOS, a generic document recognition method: Application to table structure analysis in a general and in a specific way. IJDAR **8(2)**, 111–122 (2006)
6. Coüasnon, B., Camillerapp, J., Leplumey, I.: Making handwritten archives documents accessible to public with a generic system of document image analysis. In: Int. Conf. on Document Image Analysis for Libraries (DIAL), pp. 270–277 (2004)
7. Fred, A.L., Jain, A.K.: Data clustering using evidence accumulation. In: Int. Conf. on Pattern recognition (ICPR), vol. 4, pp. 276–280 (2002)
8. Garris, M.D.: Evaluating spatial correspondence of zones in document recognition systems. In: Int. Conf. on Image Processing (ICIP), pp. 304–307 (1995)
9. Garz, A., Sablatnig, R., Diem, M.: Layout analysis for historical manuscripts using sift features. In: Int. Conf. on Document Analysis and Recognition (ICDAR), pp. 508–512 (2011)
10. Guichard, L., Chazalon, J., Coüasnon, B.: Exploiting Collection Level for Improving Assisted Handwritten Words Transcription of Historical Documents. In: Int. Conf. on Document Analysis and Recognition (ICDAR), pp. 875 –879 (2011)
11. Jayadevan, R., Kolhe, S.R., Patil, P.M., Pal, U.: Automatic processing of handwritten bank cheque images: a survey. IJDAR **15**(4), 267–296 (2012)
12. Kooli, N., Belad, A.: Semantic label and structure model based approach for entity recognition in database context. In: Int. Conf. on Document Analysis and Recognition (ICDAR), pp. 301–305 (2015)
13. Lemaitre, A., Camillerapp, J.: HIP 2013 FamilySearch Competition - Contribution of IRISA. In: Historical Image Processing (HIP) (2013)
14. Lemaitre, A., Camillerapp, J., Coüasnon, B.: Multiresolution cooperation improves document structure recognition. IJDAR **11**(2), 97–109 (2008)
15. Leplumey, I., Camillerapp, J., Queguiner, C.: Kalman filter contributions towards document segmentation. In: Int. Conf. on Document Analysis and Recognition (ICDAR), pp. 765–769 (1995)
16. Liang, J., Doermann, D.: Logical labeling of document images using layout graph matching with adaptive learning. In: Document Analysis Systems (DAS), pp. 224–235 (2002)
17. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. Journal on Computer Vision **60**(2), 91–110 (2004)
18. Mas, J., Fornés, A., Lladós, J.: An interactive transcription system of census records using word-spotting based information transfer. In: Document Analysis Systems (DAS), pp. 54–59 (2016)
19. Moysset, B., Kermorvant, C., Wolf, C.: Full-Page Text Recognition: Learning Where to Start and When to Stop. In: Int. Conf. on Document Analysis and Recognition (ICDAR), pp. 871–876. Kyoto, Japan (2017)
20. Nielson, H.E., Barrett, W.A.: Consensus-based table form recognition of low-quality historical documents. IJDAR **8**(2-3), 183–200 (2006)
21. Nion, T., Menasri, F., Louradour, J., Sibade, C., Retornaz, T., Metaireau, P., Kermorvant, C.: Handwritten information extraction from historical census documents. In: Int. Conf. on Document Analysis and Recognition (ICDAR), pp. 822–826 (2013)
22. Pham, T.A., Alaei, A.: ICDAR 2013 HIP Workshop Family Search Competition: A multi-scale image analysis approach for historical document image classification. In: Historical Image Processing (HIP) (2013)

23. Richarz, J., Vajda, S., Fink, G.A.: Towards semi-supervised transcription of handwritten historical weather reports. In: Document Analysis Systems (DAS), pp. 180–184 (2012)
24. Romero, V., Fornés, A., Serrano, N., Sánchez, J., Toselli, A.H., Frinken, V., Vidal, E., Lladós, J.: The ESPOS-ALLES database: An ancient marriage license corpus for off-line handwriting recognition. Pattern Recognition **46**(6), 1658–1669 (2013)
25. Sibade, C., Retornaz, T., Nion, T., Lerallut, R., Kermorvant, C.: Automatic indexing of french handwritten census registers for probate geneaology. In: Historical Image Processing (HIP), pp. 51–58 (2011)
26. Stewart, S., Barrett, B.: Document image page segmentation and character recognition as semantic segmentation. In: Historical Document Processing (HIP), pp. 101–106 (2017)
27. Ye, X., Cheriet, M., Suen, C.Y.: A generic method of cleaning and enhancing handwritten data from business forms. IJDAR **4**(2), 84–96 (2001)