HIP 2013 FamilySearch Competition - Contribution of IRISA

Aurélie Lemaitre, Jean Camillerapp

HIP 2013 FamilySearch Competition Contribution of IRISA

Aurélie Lemaitre
IRISA/Université Rennes 2
Rennes, France
Email: aurelie.lemaitre@irisa.fr

Jean Camillerapp
IRISA
Rennes, France
Email: jean.camillerapp@irisa.fr

***Abstract*—In this paper, we present the method that we have proposed for ICDAR'2013 HIP Workshop FamilySearch Competition. This method is based on the study of the arrangement of local descriptors called Points of Interest (POI). The points of interest are used in this context to realize some word spotting. Then, the word spotting is exploited at two levels in the competition: the localization of regions of interest in the document and the clustering of similar text regions. Due to lack of time, we have submitted a very first version of our method, but we hope to improve it in future work.**

I. Introduction

This work has been realized for ICDAR'2013 HIP Family Search Competition. This competition focuses on Mexican marriage records that are used by the genealogists. In those printed forms, the genealogists are interested in several handwriten fields: month and year of the record, origins of the attendees. Those fields are usually manually transcribed.The goal of the project is to assist the transcription by grouping together the fields that contain the same indication, in different records. The Intuidoc team of IRISA laboratory works on the interactive recognition of document images. Thus, we are familiar with the problem of automated assisted transcription of old documents [1].

The work for the competition can be separated into two tasks. First, we must localize in documents the handwritten fields, called Regions Of Interest (ROI), inside of the records, that contains month, year, and origins of the two attendees. Secondly, for each kind of field, we must gather the ROI that contain the same text. It is not asked to recognize the content of the regions.

For those two steps of analysis, we have based our method on the use of arrangements of local descriptors, called Points Of Interest (POI). The paper is organized as follows. The first section presents the technical concept of POI. Then, we present how we use the POI inside of a grammatical method for the localization of the regions of interest. In section IV, we explain the use of POI for word spotting.

II. Presentation of the concept of POI

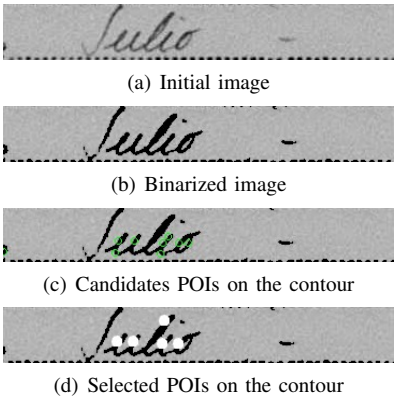
The POI (Points Of Interest) are used in this context to realize some word spotting. In order to present the concept of POI, we will explain three aspects: which pixel is a good POI, how to represent a POI, and how to use a POI.

A. Detection of points of interest in an image

The main objective of the points of interest is to select a small set of points of the image, that present some interesting local variations of luminosity. This selection must be stable: we must select the same points to represent the same object that is present in different images. Moreover, the selected points must be discriminating: in an image, there must be few confusion between local descriptors.

In our work, we first binarize the image. Then, we use the points of the contour, as they are located inside of strong luminosity gradient zones. We arbitrarily choose the points of the left contours as points of interest. This gives some candidates points of interest to represent the zone.

Then, some of the POIs are selected to build a model that represents the zone of image. This selection can be made manually, if we want to define just one model. The selection can also be automatic. In that case, the system selects the POIs that are present on upstrokes and downstrokes of characters. The figure 1 shows an example of 5 points of interest that are extracted to build a model of the word *Julio*.



(a) Initial image

(b) Binarized image

(c) Candidates POIs on the contour

(d) Selected POIs on the contour

Fig. 1. Example of detection of points of interest

B. Choice of local descriptor

For each selected point of interest, we compute a local descriptor. We use the descriptor proposed by Lowe [2]. We use the simple version of its local descriptor. The principle of this descriptor is to compute some statistics on the gradient direction in a small neighbourhood. It uses a 15x15 window, a

8-direction quantization and a calculation in a 3x3 matrix. We then obtain a 72 element vector, that is the final descriptor.

For the comparison of two descriptors, we use an Euclidean distance, as proposed by Lowe [2].

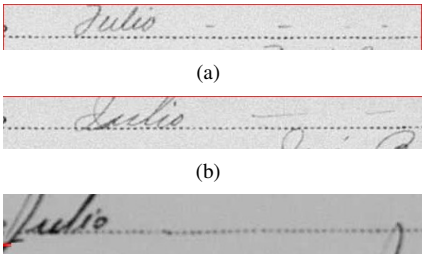
C. Localization of a model in an image

A model is represented by a set of points of interest, their coordinates and the associated descriptor. The localization of a model consists in finding some points in the image that match with the points of the model.

The matching between two points is correct when the distance between the descriptors is smaller than a given threshold. It is a photometric matching.

The model is found in the image only if all the points of the model are found if the image. The principle of matching is the following: we match the first point of the model in the image. Then, wee look for every other matching point of the model, in restricted areas. It is a geometric matching.

For example, with the five points of interest that are selected on figure 1, we build a model. We try to find this model on other images. The figure 2 shows some examples of images in which this model is found.



(a) (b) (c)

Fig. 2. Example of images in which the model of figure 1 is localized

We will now detail how we have used the points of interest in the two steps of our analysis process: the localization of regions of interest, and the clustering of words.

III. Localisation of Regions Of Interest

The first step of our analysis process consist in localizing the regions of interest that are required by the competition: month of record, year of record and origin of the two attendees. For that purpose, we use a grammatical method for document structure recognition, DMOS-P. In this section, we first present the existing DMOS-P method, before detailing how we used is in the context of the competition.

A. DMOS-P method

The DMOS-P (Description and MOdification of Segmentation with Perceptive vision) method is a grammatical method for document structure recognition [3] [4]. It is based on a grammatical formalism, EPF (Enhanced Position Formalism) that enables a syntactical, semantic and symbolic description of the content of the document. Thus, for each new kind of document to recognize, it is only necessary to describe its content with EPF language. Then, the associated parser is automatically produced by a compilation step. The method is qualified of "perceptive" has it enables to build a cooperation between several points of view of the documents: several resolution levels or various kinds of primitives.

This method has been applied on many kinds of documents: musical scores, tabular forms, archive documents, handwritten mails.... It has been widely validated and applied at a large scale (about 800,000 documents).

In the context of FamilySearch competition, we just had to write a specific grammar for the description of marriage record pages. Then, the associated parser was automatically produced by a compilation step. We detail the grammatical description used in the following sections.

B. Input primitives

As we mentioned above, the DMOS-P method can combine several points of view of the document, by using as an input various kinds of primitives. The primitives are used as terminals for the grammatical description. We use two kinds of terminals: the line segments and some models that are localized thanks to points of interest.

The line segments are extracted with a method based on Kalman filtering. The line segments are detected on an image at low resolution: the dimensions of the initial images are divided by 8. This enable to keep only in the image the most important line segment. The figure 3 shows an example of line segments that are extracted and given as input primitives.

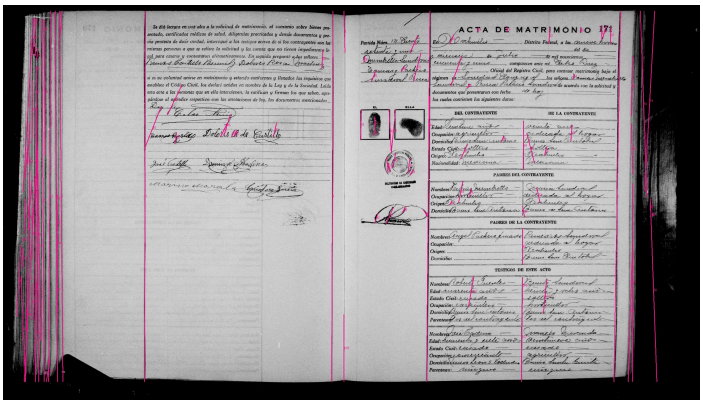


Fig. 3. Example of input primitives: line segments extracted on a image at low resolution

The second important input of the grammatical description are some zones that are localized with the use of specific models, based on POI (points of interest). Thus, we have defined five sets of models, that represent some keywords that are necessary to localize the data in the documents (figure 4). The five models are:

- the letter *A* from *ACTA DE MATRIMONIO*,
- the letters *de mil n* from *de mil novecientos*,
- the letters *pare* from *comparecen*,
- the word *de*,
- the letters *Ori* from *Origen*.

Those five sets of models are applied on the initial image, to try to localize some similar fields. Consequently, as input of our grammatical description, we have some small zones where those labels have been localized. The mechanism used is the one presented in section II-C. The figure 4 shows some examples of zones that are given as input primitives of the grammatical description.

Fig. 4. Example of input primitives: zones corresponding to a matching with one of the five models described by POI. *A* in pink, *de mil n* in green, *pare* in yellow, *de* in blue, *Ori* in red.

1) Variations of models: Befor building a grammatical description, we had to identify the possible configurations of documents. It appears that the pre-print that are present in the competition are not all similar. For example, sometimes, the year is at the beginning of a line, sometimes at the end, sometimes half of the year is on a line, and the other half under.

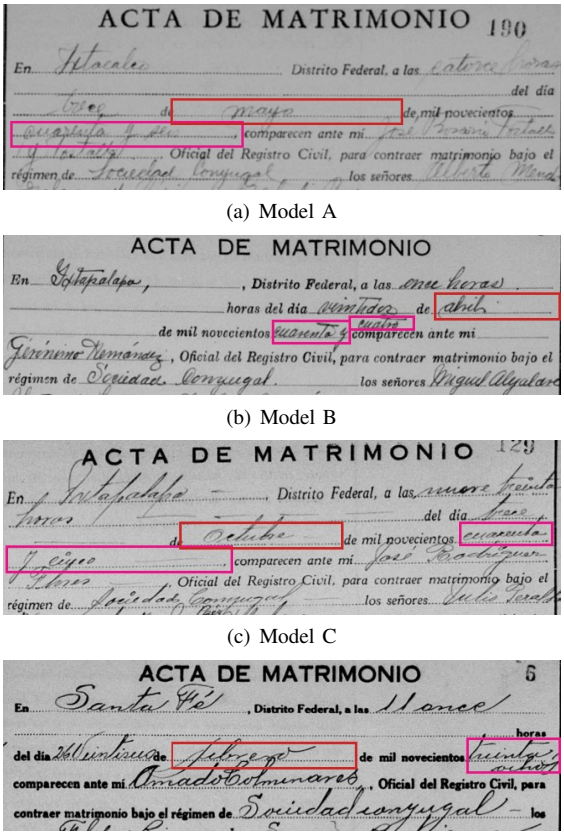
In order to treat the problem, we have identified four big families of formulae: A, B, C, D, that distinguish the different configurations of the position of year and month in the document. For example, the year region of interest is sometimes on the 3rd line, on the 4th lines, on both lines and even between two text lines. The table I synthesizes those models, that are illustrated on figure 5. Consequently, we had to adapt our grammatical description to the four categories of documents. We hope that we have identified all the existing categories of document, and as we will see in the last section, the documents have a wide variety even inside of a model.

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

TABLE I. FOUR MODELS OF REGISTERS THAT WE HAVE IDENTIFIED, WITH DIFFERENT CONFIGURATIONS OF TEXT POSITION

C. Grammatical description of models

Our grammatical description aims at combining the input primitives in order to produce the localization of the Regions



(a) Model A

(b) Model B

(c) Model C

(d) Model D

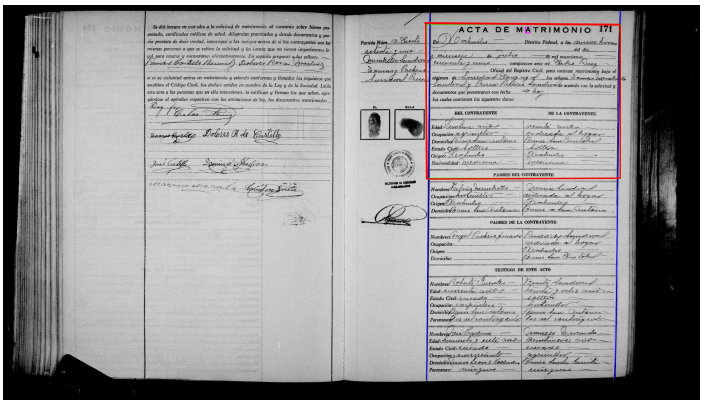
Fig. 5. Example of the four models that we have identified, with different configurations of text position, described in table I

Of Interest. It is based on the four models presented above.

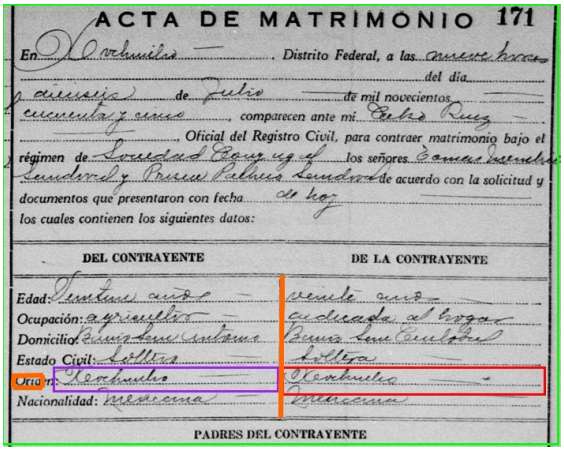
1) Steps of analysis: The analysis follows the grammatical rules:

1) Find the beginning zone of the record (figure 6(a))
   a) Find two vertical line segments in the right part of the image, that delimit the interesting column
   b) Find a model of *A* letter to localize the title
   c) Consider the upper part of the column for the remaining analysis
2) Find the origin zones (figure 6(b))
   a) At left part of the column, find a model of word *Origen*
   b) At the right of word origen, find a vertical line segment that separates the two columns
   c) Compute the two regions of interest taking into account the positions of those elements (origen word, line segments)
3) Find the month and year zones (figure 6(c))
   a) At upper part of the column, find a model of word *de mil novecientos*
   b) Before this word in the text, find a model of word *de*
   c) After this word in the text, find a model of word *comparecen*

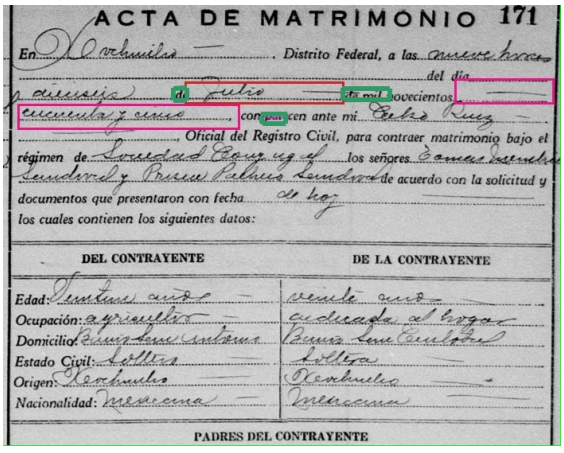
d) Compute the month ROI between *de* and *de mil novecientos*
e) Compute the year ROI between *de mil novecientos* and *comparecen*



(a) Delimitation of the interest zone, at the beginning of the record, thanks to two vertical segments (in blue) and a letter *A* from the title



(b) Localization of the origin ROI, thanks to the word *origen* and a vertical line segment



(c) Localization of the year and month ROI, thanks to the words *de*, *de mil novecientos* and *comparecen*

Fig. 6. Steps of analysis for the localisation of regions of interest

IV. Word spotting and production of final result

Once we have localized all the regions of interest, the goal is to gather the regions that contains the same text. We use the POI (points of interest) to characterize each region. Thus, we automatically build a model with POIs (as presented in section II-A) for each region of interest.

In the learning phase, we build all the models for each region. We associate the ground-truth value to each model. Then, we try to assign, for each image, the nearest model. This process enables to detect the models that are used by another image. In order to decrease the combinatory, we keep, at that step, only the models that are recognized by another image.

In the competition phase, we distinguish two cases. We consider that the month and years are closed class vocabulary, whereas we consider that the vocabulary is open-class for origins.

For the clustering of years and months in the competition, we try to assign each region of interest to one of the models that has been extracted in the learning database.

For the clustering of origins, we try to assign each region of interest either to one of the models extracted into another origin of the competition dataset.

V. First results

As we received a first database of 700 images, we present the results that we have obtained on this database. The table II presents the F-m rate, by comparison with the base scores that are obtained in a merge-all or shatter-all strategy. Those results shows that we managed to overclass by 36% the basic weighted score with our method.

	Month	Year	Origin 1	Origin 2	Weighted score
Base scores	18.8%	49.8%	43.1%	44.6 %	30%
Our results	68.5%	91.4 %	58.2 %	58.2 %	66.2%

TABLE II. Our first results (F-measure) obtained on the first 700 image dataset

We estimate that the regions of interest are quite well extracted, but the main limit is due to the clustering method. However, we will probably obtain less good results on the competitions, due to the difficulties that we met on the 10,000 images learning dataset, and mainly because we did not have time enough to overcome those difficulties.

VI. Discussion

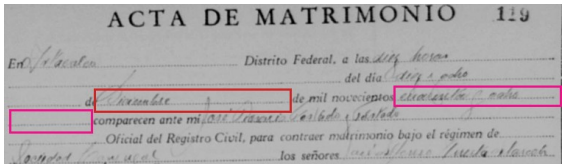
In this paper, we have presented the global method that we have used for Family Search competition. Due to lack of time, we have submitted only a very first version of our results, but we would be very interested in keeping improving that work to obtain better results. We would like to mention to aspects: the difficulties that we met and a remark about the metric.

A. Difficulties and future work

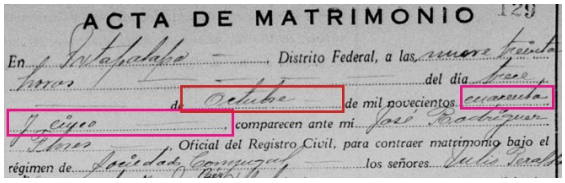
We have met several difficulties in that competition. Some are usually met in the study of archive documents, such as pale ink or damaged paper. Some are specific to those kinds of documents. We have identify several solutions for

the following problems, but we did not have time enough to introduce them in our system.

First, there are many kinds of pre-print formulae in the provided dataset. We tried to classify them into four models, but the variation inside of the models is strong. For example, the figure7 shows 3 variations of what we called model C : sometimes the year is written on line 3, sometimes on line 4, sometimes on both lines 3 and 4. This has an impact on our results.



(a) Year on 3rd line



(b) Year on both 3rd and 4th line

(c) Year on 4th line

Fig. 7. Variations of formulae inside of the C model

The second problem is on the origin field: when the name of the origin is the same for the two attendees, the origin is often written only once, in the middle of the two fields (figure 8). We should detect this case, but is it not yet active in the submitted version.

Fig. 8. Difficulty with origins: the origins of both attendees is written in the middle of the two regions of interest

Another difficulty that we plan to overcome in a future work is the line segments or marks that have been written to fill in the blanks parts of the formulae. Those marks are present inside of the regions of interest, and disturbs the word spotting. For example, they are line segments on figure 7(b), at the end of the line. There are dash lines on figure 8. We are planning to remove those disturbing lines.

The last problem we want to study is the choice of the points of interest to build the models. Indeed, in the current method, the poinst are not stable enough. Once again, it is necessary to obtain more time to study this aspect of the work.

B. About the metric

The last point we would like to mention is about the metric. The b-cubed score seems well adapted to judge a competition. However, we are not sure it is well convenient to evaluate results that aims at helping a manual annotation. Indeed, we think that for manual annotation, it is important to have a good precision: the human transcriber should not have to re-segment the proposed clusters. The recall seems less important.

With the proposed metric that computes a F-measure between recall an precision, we are tempted to build a strategy that gives a good recall, even if it decreases the precision. That gives a better final result, but it does not seems satisfactory for a purpose of helping manual annotation.

Sometimes, the system knows he cannot take a decision (for example, because the region of interest is not detected). In that case, it may be interesting that the metric takes into account a rejection class, so that the precision measure is more accurate.

VII. Conclusion

We proposed an appraoch based on the use of Points Of Interest (POI), for both localization and clustering of words. The POIs seems very adapted for the localization of regions of interest. Thus, we obtained a good localization rate on the first 700 images datasets. With few adaptations, we can detetect most of the regions on the 10,000 training dataset.

Concerning the clustering with closed class vocabulary (for month and year regions), the POIs are adapted, assuming that the automatic models are correctly choosen. This is done in two steps: the POIs that are selected for the models of words are supposed to be the most discriminant, are they are the one on upstrokes and downstrokes. Then, by application on the training set, we select only the interesting models, that is to say the ones that do not cause confusion or mistakes. Concerning the clustering of origins, that are open-class labels, the use of POIs might be discussed.

References

[1] L. Guichard, J. Chazalon, and B. Coasnon, "Exploiting Collection Level for Improving Assisted Handwritten Words Transcription of Historical Documents," in *International Conference on Document Analysis and Recognition (ICDAR'2011)*, 2011, pp. 875–879.

[2] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 91–110, Nov. 2004. [Online]. Available: http://dx.doi.org/10.1023/B:VISI.0000029664.99615.94

[3] B. Coüasnon, "DMOS, a generic document recognition method: Application to table structure analysis in a general and in a specific way," *International Journal on Document Analysis and Recognition, IJDAR*, vol. 8(2), pp. 111–122, 2006.

[4] A. Lemaitre, J. Camillerapp, and B. Coüasnon, "Multiresolution cooperation improves document structure recognition," *International Journal on Document Analysis and Recognition (IJDAR)*, vol. 11, no. 2, pp. 97–109, November 2008.