Combination of deep neural networks and logical rules for record segmentation in historical handwritten registers using few examples

Solène Tarride, Aurélie Lemaitre, Bertrand B. Coüasnon, Sophie Tardivel

Noname manuscript No.
(will be inserted by the editor)

Combination of deep neural networks and logical rules for record segmentation in historical handwritten registers using few examples

Solène Tarride · Aurélie Lemaitre · Bertrand Coüasnon · Sophie Tardivel

Received: date / Accepted: date

Abstract This work focuses on the layout analysis of historical handwritten registers, in which local religious ceremonies were recorded. The aim of this work is to delimit each record in these registers. To this end, two approaches are proposed. Firstly, object detection networks are explored, as three state-of-the-art architectures are compared. Further experiments are then conducted on Mask R-CNN, as it yields the best performance. Secondly, we introduce and investigate Deep Syntax, a hybrid system that takes advantages of recurrent patterns to delimit each record, by combining u-shaped networks and logical rules. Finally, these two approaches are evaluated on 3708 French records (16-18th centuries), as well as on the Esposalles public database, containing 253 Spanish records (17th century). While both systems perform well on homogeneous documents, we observe a significant drop in performance with Mask R-CNN on heterogeneous documents, especially when trained on a non-representative subset. By contrast, Deep Syntax relies on steady patterns, and is therefore able to process a wider range of documents with less training data. Not only Deep Syntax produces 15% more match configurations and reduces the ZoneMap surface error metric by 30% when both systems are trained on 120 images, but it also outperforms Mask R-CNN when trained on a database three times smaller. As Deep Syntax generalizes better, we believe it can be used in the context of massive document processing, as collecting and annotating a sufficiently large and representative set of training data is not always achievable.

Keywords Historical handwritten documents · Deep neural networks · Hybrid systems · Layout analysis

Doptim - Univ Rennes - CNRS - IRISA
F-35000 Rennes
E-mail: solene.tarride@irisa.fr

1 Introduction

French parish registers are handwritten books from the 16th century onward. Information about local religious ceremonies, mainly baptisms, marriages and burials, were recorded by the priests. These records were initially written to prevent bigamy and consanguineous marriages. Parish registers are structured in acts - or records - that are paragraphs describing a specific ceremony. The records are independent of each other, and written in chronological order. A page from a parish register is shown in Fig 1. These documents are especially useful to genealogists because they contain local information on births, marriages and deaths. As a result, they are used to find ancestors and reconstruct family links. An illustration of the recurrent information that can be found in such records is provided in Fig. 2. Moreover, parish registers are the only reliable source of demographic data for French people born be-

(a) Original image **(b)** Record segmentation

Fig. 1: Page from a French parish register (a) and its record segmentation (b). Figure best viewed in color.

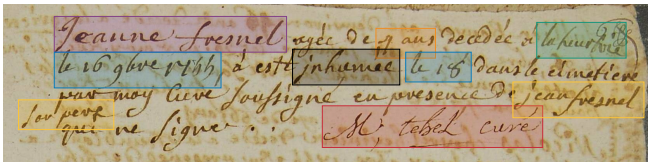
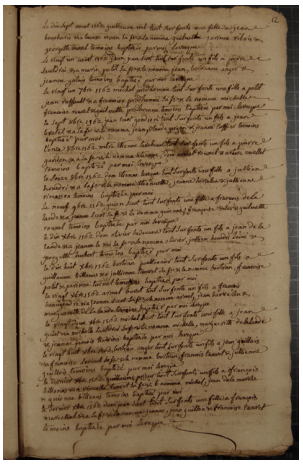


Fig. 2: Zoom on a burial record. Recurrent information is highlighted: name, age, place, date, ceremony, witness, and signature. Figure best viewed in color.

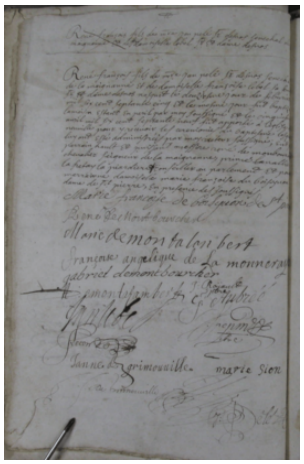
fore the French Revolution, as mandatory civil registration was established after the Revolution, in the late 18th century.

Recently, there has been a raising interest in scanning the archives, as it eases access to the documents while avoiding their degradation. In France, most parish registers are now accessible online. In spite of this, the search for ancestors remains time-consuming and laborious as it is necessary to search the archives to find relevant records. As a result, there is a need for automatic methods able to analyze the contents of these documents. Structure analysis is a key step in this process. Automatic delimitation of each record has an immediate and practical advantage for genealogists, as it eases the reading process and allows them to save only the records that are relevant to their research. But it is also the first step towards text recognition and word spotting. Once each record is detected, it is possible to train models to spot relevant keywords and extract valuable knowledge, such as names, places and dates. This could substantially ease the search for ancestors.

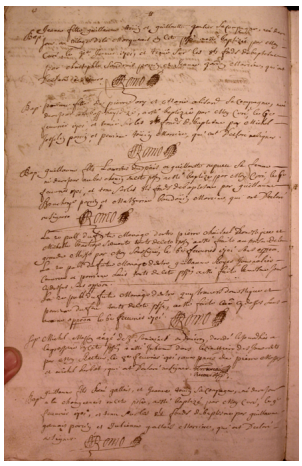
However, these documents are difficult to process. Firstly, parish registers are poorly-structured since the records were written one under the other, with no clear separation. In some documents, there are patterns indicating the localization of the records: signatures, vertical spacing, horizontal lines, marginal annotations... However, they are not consistent within the corpus as they depend on the writer. The handwriting is often compact with very few vertical spacing between successive records. Moreover, the writing style differs from one page to another, but is uniform within a page since two successive records were likely written by the same priest at the same date. It is also frequent to have an overlap between two successive records, mainly due to overlapping text and signatures. On top of that, some documents are heavily degraded with notably ink stains, ink fading, bleed-through and torn or cut pages. But the main challenge of this work is the variability of parish registers, as they come from churches from all over France from three centuries. As such, they were written by different priests, each with different writing style and phrasing. In addition, they are unevenly pho-



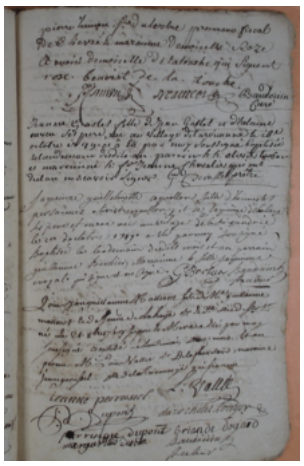
(a) Page with 15 records from 1562.



(b) Page with 1 record from 1675



(c) Page with 7 records from 1701.



(d) Page with 4 records from 1749.

Fig. 3: Samples of pages from multiple French parish churches at different periods.

tographed since different cameras were used for scanning, as well as different poses, uneven illumination and contrast variation. This variability is illustrated in Fig. 3. Collecting and annotating documents from each church and time period is not achievable in practice. Thus, the aim of this work is to design a method that can handle a wide variety of documents while being trained on a small, non-representative subset.

In the next section, some approaches that tackle document layout analysis of historical documents are presented. Then, we introduce two strategies applied to record detection. In section 3, several architectures of object detection networks are compared, and further experiments are performed on Mask R-CNN that yield the best performance. In section 4, we introduce Deep Syntax, a hybrid method based on neural networks and logical rules. Finally, these two approaches are compared on two databases in section 6. Results show that

Deep Syntax is able to handle more heterogeneous documents while being trained on a few documents. It outperforms object detection networks, even when trained on a database three times smaller. As a result, it is more adapted to the context of massive document processing with few training examples.

2 Related works

Document layout analysis is the process of identifying regions of interest in document images, such as text-blocks, tables or graphics. It is a key step for automatic document understanding. In this section, an overview of methods that have been developed to solve this task is presented. It is worth noting that we focus on methods performing an image-based analysis, although other methods also rely on Optical Character Recognition (OCR) to delimit zones based on local text cohesion [9, 27].

This study mostly focuses on methods applied to handwritten or degraded documents. However, some interesting methods applied on printed text are also considered. There are four main types of strategy used for document layout analysis:

- **Bottom-up** strategies start from the smallest elements of the images (pixels, connected components...) and group them based on similarity.
- **Top-down** strategies start from the whole page and partition it into homogeneous zones.
- **Hybrid** strategies combine the two previous types of approaches.
- **Neural networks** learn to recognize different layouts from annotated examples.

2.1 Bottom-up or data-driven strategies

Bottom-up algorithms agglomerate the smallest components of a document such as pixels, connected components, words or text-lines to create homogeneous regions. They are able to determine the structure of a wide variety of documents without any prior knowledge. However, they are sensitive to noise, and can essentially be applied to documents with clearly delimited areas (i.e. newspaper columns, text/image separation). There are three main bottom-up strategies.

Mathematical morphology These algorithms rely on filtering techniques to reveal areas of interest. In practice, filters can be applied on printed documents to remove graphics or noise [25]. They can also be used to detect text-lines in printed and handwritten documents, assuming that the text is not too slanted [12, 42, 63].

Clustering These approaches try to agglomerate elements based on specific sets of features. Some algorithms rely on texture features to find homogeneous zones, and a few of them have been applied to historical documents. For instance, they can be used to separate textual and graphical regions [45,46]. Journet et al. [38] identify main areas in historical documents without any prior knowledge by extracting five local texture characteristics at different resolutions. These approaches have also been applied to printed documents in the context of text-block detection. In [30], the authors rely on features derived from the geometry of the document and perform hierarchical graph coloring to retrieve the structure of postal mails. In [39] text-lines are grouped based on alignment, distance and graphical features like font, thickness and color to form homogeneous zones. It is also common to gradually merge connected components to obtain text-blocks in printed documents [4,37] Clustering methods are also applied to find text-lines using generic features, such as orientation features [40, 71]. In [14], the authors perform partitioning of connected components at different resolutions on each color layer. Geometrical features such as distance, area, and density are also commonly used to extract text-lines [23, 70]

Classification These algorithms classify structural elements (pixels, letters, text-lines...) from a set of learned features. Some of them have been successively applied to separate handwritten annotations from printed text by using connected component and patch level features [53, 54], shape context features [22] or more traditional features [8]. In [36], structure detection of degraded newspaper archives is achieved by localizing titles, text-lines, background, separators and noise using a Conditional Random Field.
These methods have also been applied to handwritten documents. In [26], a structure analysis is performed using relative location features. In [32], the authors manage to detect initials, headings and text areas in historical documents. Other works focus on historical manuscripts to solve structure recognition [15, 65] as well as text-line extraction [6, 16, 28].

2.2 Top-Down strategies

Top-down strategies start from the whole page and partition it into smaller homogeneous zones. These strategies are useful to delimit well-defined and invariant structures, as they require a prior knowledge to guide the analysis. They are generally very fast but are not suitable for documents with complex or varying structures.

These methods mostly rely on document structure assumption, projection profiles and function analysis.

Document structure assumption These strategies are based on a strong prior knowledge of a document layout, that is invariant and codified (e.g. forms, letters). As a result, they are not very flexible since they are not immediately applicable to other layouts. A document structure description tool was proposed in 2006 by Coüasnon: DMOS [20]. Document structure is described using logical rules to achieve segmentation and classification of areas of interest. Although DMOS can also rely on bottom-up analysis, the grammatical part is essentially top-down. This method has been successfully applied to many layouts [41], such as correspondence letters, administrative documents, historical newspapers or musical scores. Although, some analysis is based on bottom-up analysis A probabilistic method was introduced by Shafait et al. [62]. For each document, the method returns the most probable zones, using a prior user-defined breakdown. Finally, Alvaro et al proposed a 2D Stochastic Grammar using two sets of features: Gabor and RLF.

Projection profiles This strategy consists of identifying physical boundaries between areas of interest. These methods are very effective on documents containing only text, but are difficult to apply to complex documents (e.g.presenting degradation, complex structure or graphic elements...). For printed documents, it is common to take advantage of the regular gaps between text-lines to find text-block boundaries [4]. For instance, the Viterbi algorithm is used in [52] to find text-lines by locating optimal succession of text and gap areas. In [17], the authors propose to analyze white spaces to separate columns in old newspapers. This strategy can also be applied to extract curved text-lines, such as in [51], where text-line orientation is determined using the Wigner-Ville distribution on the projection histogram profile.

Function analysis This type of method is based on the optimization of a function specifically designed to solve a given problem. Bukhari et al. propose a method based on active contours to delimit text-lines from the top and bottom [11]. It has been successfully applied to printed documents featuring extremely curved text-lines. The method proposed by Ryu et al. [60] tackles text-line segmentation in handwritten documents based on an energy function designed in such a way that its minimization yields text-lines. Yin et al. [69] propose to estimate the number of text-lines using a fuzzy filter and then apply a variational Bayesian method to segment text-lines. Weliwitage et al. [67] use an optimization technique to minimize text pixels cut by the frontier between text lines on distorted handwritten documents. Function analysis can also be used for layout analysis: in [21], the authors propose to apply a Gaussian mixture to the different regions of the page to obtain the logical distribution of the handwritten document.

2.3 Hybrid strategies

These methods rely on a combination of bottom-up and top-down strategies. They are efficient since they combine advantages of both types of methods. However, they take time to implement because many parameters must be optimized for each type of document.

In [13], the authors use projection profile to find text-lines and achieve text-block detection by taking advantage of the rigid structure of their collection of historical documents. Wei et al [66] study a hybrid selection of textual characteristics to tackle the task of layout analysis on historical documents. In [7], connected components are aggregated before vertical and horizontal white spaces are detected to produce a mask of areas of interest. Asi et al. [5] manage to simplify the layout of historical documents by locating, segmenting, and dewarping text lines with severe curvature. These strategies have also been applied to segment text-lines. Clausner et al. [18] propose to combine a connected component analysis (bottom-up) with logical rules (top-down) to obtain text-lines. Their study shows that their hybrid approach outperforms the purely bottom-up or top-down approach.

2.4 Neural network-based strategies

These methods are the most recent and have taken the lead in most competitions in the field since 2015.

Moysset et al. [48] were among the first to develop a sequential RNN-LSTM model that allows to obtain lines, text and paragraphs on various documents in the MAURDOR database. Recently, neural networks, and particularly fully convolutional networks, have gained popularity and have proven to be particularly efficient for text-line extraction and semantic segmentation in historical documents [24, 47]. Grüning et al. have proposed ARU-Net [33], a U-net with recurrent layers and coupled to an attention network, which achieves the best results for the detection of text lines in the cBAD database. Renton et al [58] have also proposed a fully convolutional network with dilated convolutions and have obtained competitive results for the detection of text-lines. Finally, Oliveira et al. have proposed dhSegment [50], a U-Net of which the contracting path con-

sists of a ResNet-50 trained on ImageNet. The authors demonstrated its genericity by successfully solving five semantic segmentation tasks on historical documents: page extraction, text line extraction, structure detection, decoration detection and photo detection. Albertini et al. [3] have used the DeepDIVA framework [2] to obtain high quality semantic segmentation before extracting text-lines. Alaasam et al. [1] have used siamese networks at the patch level for semantic segmentation of challenging historical Arabic manuscripts.

Deep neural networks have also been increasingly applied for block detection and classification. Several methods rely on object detection networks to locate and classify text-blocks, tables, equations and figures in complex printed documents [49,61,68]. These networks can handle different layouts of printed documents, but require many training examples - more than 1,000 documents in these studies. To the best of our knowledge, the only attempt at applying object detection networks on historical documents was done by Prusty et al. [55]. They have trained Mask R-CNN on 120 to 350 documents to find instances of different page objects, such as text-lines and page boundaries, in historical Indic manuscripts.

2.5 Discussion and outline of the paper

In this section, we discuss the applicability of these strategies to parish registers within the context of massive document processing.

Many approaches are not applicable to parish register structure, as their record structure does not appear clearly. The layout is tight, with no clear separation between two successive records (e.g. white space, separator line). Text-lines can be skewed, and some words can partially overlap with the text belonging to the previous record. Moreover, these documents are old, and therefore some pages are degraded. For these reasons, bottom-up strategies do not appear to be the best methods to be applied to these documents. Furthermore, the layout of parish registers is not rigid. The page layout depends on the priest writing the register: the number of records varies from one page to another, some records are on two pages, some pages do not contain any records. A few indicators help to delimit the records, such as margin annotations, vertical and horizontal spacing. However, they are not consistent within the corpus. For these reasons, applying top-down methods relying on projection profiles and function analysis might not be suited.

Yet, we believe that some strategies described in this literature review can be successfully applied to parish registers. First, numerous studies have shown the efficiency of deep neural networks for layout analysis tasks. These methods have the capacity to automatically extract relevant features for a wide range of layouts. U-shaped networks are successfully applied to text-line detection [33] [50], page extraction, layout analysis and ornament extraction [50] in historical documents. However, they are not able to deal with overlapping regions of a same class, and thus they cannot be applied for record segmentation. Whereas, object detection networks are able to retrieve overlapping instances of a same class. Several methods rely on object detection networks to localize and classify text-blocks, tables, equations and figures in printed documents [61] [68]. Consequently, we argue that deep neural networks seem applicable to these documents. The main limit of these methods is that they require a lot of representative training records to learn relevant features. Considering the large variability of parish registers, collecting and annotating such a database would require too much time and effort. In this context, object detection neural networks might not be suitable. Another approach to consider would be document structure assumption. Although the page layout is not rigid, most records present similar features, such as recurrent patterns and keywords. These steady features could be exploited to detect each record. As these patterns are stable, they would likely be easier to learn with few training examples.

In this article, we propose to compare two systems to find the records in these documents:

1. **The Object Detection system**. This strategy relies on object detection neural networks trained to directly detect the records. Three architectures are compared and several experiments are carried out. This approach is further described in section 3
2. **The Deep Syntax system**. This strategy combines u-shaped neural networks and a syntactic approach. It relies on the recurrent first text-lines and signatures to locate the beginning and the end of each record. This strategy is further described in section 4.

Finally, these approaches are compared in section 6.

3 Object detection networks

Deep neural networks have consistently outperformed most of the other methods for document layout analysis. With enough training data, learning-based methods achieve great performance in addressing complex layouts in both printed and handwritten documents.

We believe that this approach could be successfully applied to historical handwritten documents. Indeed, the records share similarities, both in structure and content: vertical and horizontal spacing, potential margin annotations, capital letters, recurrent keywords, signatures... These similarities could help the network to learn a representation of the records. The challenge of such a strategy is to learn to recognize complex, varying and overlapping objects, using few available training data.

3.1 Selected architectures

We perform several experiments using state-of-the-art neural networks: Mask R-CNN [35], RetinaNet [43] and YOLOv3 [56].

Mask R-CNN [35] is a two-stage approach. First, a Feature Pyramid Network is used as a backbone for feature extraction over the entire image, then the network head is used for bounding-box classification and regression. Mask R-CNN is based on Faster R-CNN [57], which has two outputs for each candidate object: a class label and a bounding-box. To this, Mask R-CNN adds a third branch that outputs the object mask, allowing instance segmentation. Mask R-CNN also introduces pixel-to-pixel alignment, which leads to consistent improvement over Faster R-CNN for object detection tasks. Thus, we choose Mask R-CNN over Faster R-CNN. Two-stage detectors such as Mask R-CNN are generally more accurate than one stage detectors, but are much slower.

RetinaNet [43] is a single shot detector. The network architecture is composed of a backbone network and two subnetworks. The backbone is a Feature Pyramid Network that computes convolutional feature maps over the entire image. The first subnetwork is used for object classification and the second for bounding box regression. The major improvement of RetinaNet comes from a novel focal loss function that handles the class imbalance. RetinaNet is able to match the speed of previous one-stage detectors while surpassing the accuracy of many state-of-the-art two-stage detectors, including Faster R-CNN [57].

YOLOv3 [56] is a single shot detector. The object detection task is tackled as a regression problem to spatially separate bounding boxes and class probabilities. In this way, bounding boxes and class probabilities are directly predicted from full images in one shot. The input image is divided into a grid where each cell predicts bounding boxes, confidence score, and class probabilities. In YOLOv3 the prediction is done across three different scales which improves the performance. This architecture is simple and fast, yet accurate.

3.2 Experimental protocol

We propose to train these architectures for record detection using similar experimental protocols. We now detail the training setup used to compare Mask R-CNN [35], Retina-Net [43] and YOLOv3 [56] for record detection.

Raw images are given as an input. They are scaled such as the larger side is equal to 1,000 pixels. Augmentation consists of random horizontal flips and Gaussian blur with sigma randomly chosen between 0.0 and 3.0. Each model is pre-trained on the COCO dataset [44]. Early stopping is used during training. The model yielding the best validation loss is saved and used. All implementations rely on the Keras framework. Training is done using NVIDIA RTX 2080 Ti GPUs. For post-processing, predictions with low confidence score (< 0.5) are discarded. Record widths are then normalized based on page borders if they are close enough to predicted borders. The training setup relative to each network is described as follows.

- Mask R-CNN[1]: Bounding boxes are transformed into masks such as there is one mask for each bounding box. ResNet-50 is used as a backbone. The backbone layers are frozen during the first stage of the training, then unfrozen.
- RetinaNet[2]: Bounding boxes are represented with their four coordinates and are shuffled before training. ResNet-50 is used as a backbone. The backbone layers are frozen during the first stage of the training, then unfrozen.
- YOLOv3[3]: Bounding boxes are represented with their four coordinates and are shuffled before training. Darknet-53 is used as a backbone. Early layers are frozen during the first stage of the training, then unfrozen.

Results are presented and discussed in section 6. One of the limitations of object detection networks is that they require a large training database. As previously mentioned, collecting and annotating such a database is beyond our means. To overcome this constraint, we introduce Deep Syntax, a hybrid approach that should be able to learn with less training data.

4 DeepSyntax: our proposed approach combining neural networks and logical rules

In this section, Deep Syntax, our original contribution, is presented. We propose to take advantage of recurrent

[1] https://github.com/matterport/Mask_RCNN
[2] https://github.com/fizyr/keras-retinanet
[3] https://github.com/qqwweee/keras-yolo3

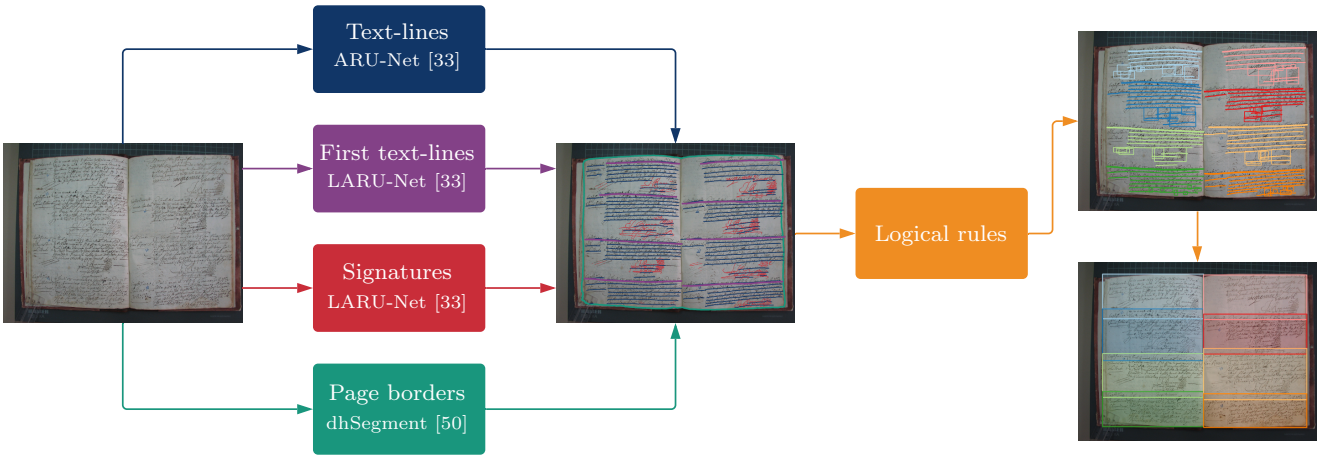


Fig. 4: Overview of Deep Syntax. First, neural networks are used to predict several patterns: text-lines, first text-lines, signatures and page borders. Then, logical rules are applied to group patterns belonging to the same record. Finally, record borders are computed by taking the bounding box of each group. Figure best viewed in color.

patterns to spot the records in parish registers. Neural networks are trained to find these patterns and logical rules are applied to group them based on prior knowledge regarding parish register layout. The workflow of this system is summarized in Fig. 4.

4.1 Taking advantage of useful patterns

We argue that the records share common features that can be used to spot the records [64]. In this section, some helpful patterns for record segmentation in parish registers are presented: signatures, first text-lines, page borders and text-lines. Signatures and first text-lines are especially helpful as they help to delimit the end and the beginning of each record.

4.1.1 Signatures

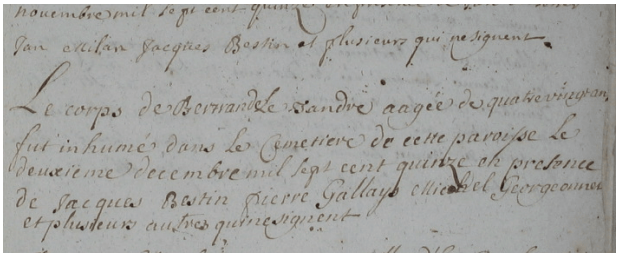
One of the most consistent patterns that can be found in the records is a signature. Indeed, each record was generally signed by the priest. In some cases, it was also signed by several witnesses. In some rare cases, the record is not signed by anyone, as in Fig. 5a. Consequently, extracting signatures can help to find the end of each record, but our system should rely on other patterns as well for record segmentation.

Extracting signatures can be done since they share some common features such as localization and style. Different methods were compared for signature segmentation, as described in [64]. We trained a LARU-Net using a database containing 200 images: 120 images are used for training, 40 for validation and 40 for testing. Five-fold cross-validation is used such as each image is in the testing set at some point. For each pixel, the network predicts its probability of belonging to a signature. A threshold is applied to probability maps and the connected components that are too small are removed.

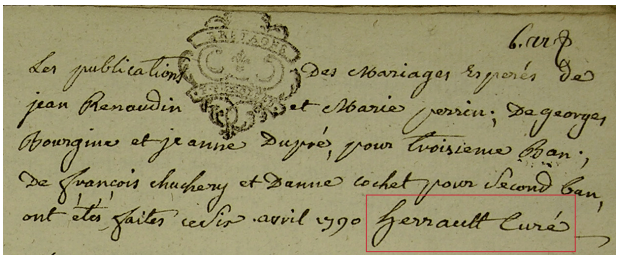
Signature extraction can be challenging, as they appear very similar to the main text. As a consequence, signature detection can lead to prediction errors. For instance, a signature can easily be missed, which triggers a merge error at the record level. This is illustrated in Fig 5b, where the signature looks a lot like the main text. In opposition, the network can also produce a false positive, which triggers a split error at the record level. This is especially true on images such as Fig. 5c, where names are more elaborate than signatures. Finally, there is a strong interaction between signatures and the main text located below making the frontier between records unclear and leading to small surface errors. Sometimes, signatures and text even overlap, such as in Fig. 5d.

4.1.2 First text-lines

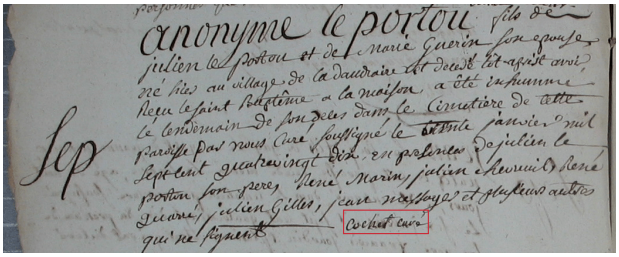
The first text-line of each record is another striking pattern as first text-lines present steady features over the records. Firstly, they share similar textual content as the same phrasing is used by different priests. As a consequence, they contain recurrent, letters, words and expressions. For example, *'fils/fille de'* ('son/daughter of'), *'ce jour'* ('this day'), and *'le corps de'* ('the body of') are frequently found in the first text-line. Also, grammatical articles such as *'le/la'* ('the') are commonly used at the beginning of a sentence in French, hence, first text-lines often begin with these words. Finally, names are usually located in the first text-line,



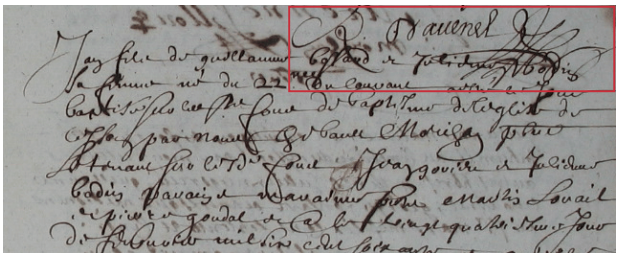
(a) There is no signature.



(b) Signatures are hard to distinguish from the main text.

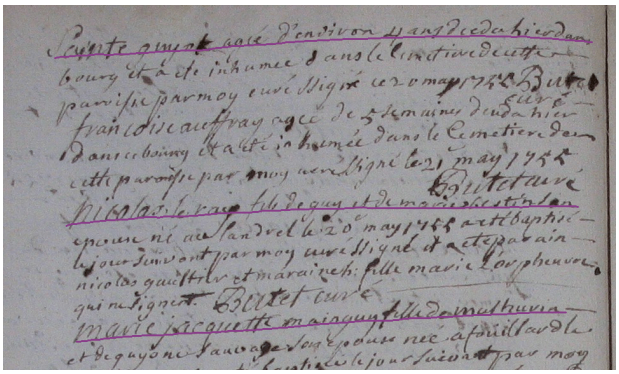


(c) Names are more stylized than signatures.

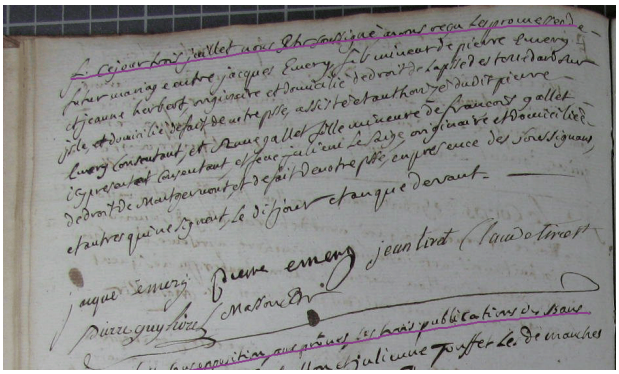


(d) Text and signatures overlap.

Fig. 5: Example of challenging records for signature detection. Signatures are enclosed by a red rectangle. On record (a), there is no signature to signal the end of the record. Signature from record (b) is not obvious and could easily be missed. Names and margin annotations from record (c) are elaborate and could be falsely detected as signatures. The text from record (d) overlaps the signature of the previous record. Figure best viewed in color.

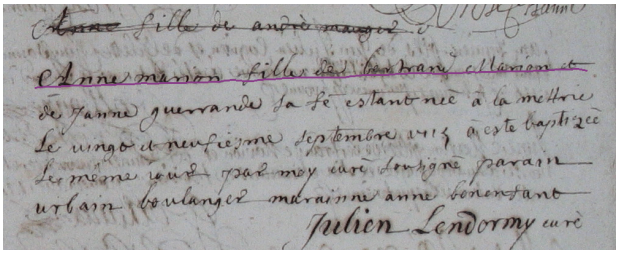


(a) Very compact handwriting.



(b) The group of signatures show features similar to first text-lines, mainly vertical spacing and capital letters.

(c) The first text-line is written around a seal.



(d) A first text-line is crossed out and written again below.

Fig. 6: Example of challenging records for first text-line detection. First text-lines are underlined in purple. In (a), first text-lines are hard to spot due to uniform line spacing. In record (b), the group of signatures could be falsely detected as first text-line. On record (c), the first line is broken and can be missed. In record (d), a first text-line is crossed-out but would likely be detected anyway. Figure best viewed in color.

and can be easily spotted since they begin with a capital letter.
Secondly, several context-based indications can help to localize first text-lines. On average, there is a white gap above the first text-lines since vertical spacing is larger between two records. Moreover, signatures are consistently located above the first text-lines. In the same way, margin annotations are often lined up with the first text-lines. For these reasons, we believe that the first text-line of each record can be found using neural networks. This would allow to delimit the records.

We compared several architectures for this task. LARU-Net clearly outperforms the other architectures. Experiments regarding the input format are also performed: we trained the network with two classes (first/other text-lines or first/all text-lines) or only one class (first text-lines). Experiments show that training using one class outputs better results. The training is done using a database containing 200 images: 120 images are used for training, 40 for validation and 40 for testing. Five-fold cross-validation is used such as each image is in the testing set at some point. For each pixel, the network determines the probability of belonging to a first text-line. Different post-processing have been compared. The best method consists of extracting blurred text-lines in predicted masks.

Fig. 6 gives an overview of some challenging records. The main issue is that first text-lines can look similar to other text-lines, especially in tight layouts, such as in Fig. 6a. In this page, there is almost no vertical spacing between two successive records. In this configuration, a first text-line would likely be missed, which would trigger a merge error at the record level. In Fig. 6c, the first text-line can be missed as well as it is written around two seals. In opposition, the network can produce false positive first text-lines, which triggers a split error at the record level. This is especially true on images such as Fig. 6b where groups of signatures appear similar to first text-lines: signatures are aligned with a significant vertical spacing above. In Fig. 6d, a first text-line is crossed out but would likely be detected anyway, as well as the text-line below. In this case, small surface errors would appear at the record level.

4.1.3 Text-lines

Text-lines are the main structural component of the records. Extracting them is a key step in record segmentation. Using logical rules based on signatures and first text-lines, text-lines that likely belong to a same record are grouped together. The records are found by extracting the bounding box of each group.

ARU-Net [33] trained on cBad is used to extract text baselines. Then, the post-processing introduced by Oliveira et al. [50] is applied. Probability maps are filtered using a Gaussian filter and hysteresis thresholding is applied. Connected components in the binary map are then converted to a polygonal line.

4.1.4 Page borders

Localizing page borders is useful to apply the analysis within the page. They are also used to normalize the width of each record.

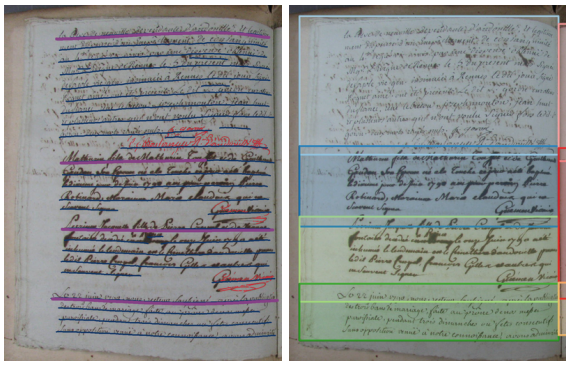
Available pre-trained networks are usually designed to handle single-page documents. As a result, we had to train a model for single and double-page documents. Experiments led us to select dhSegment [50] to perform page segmentation. Training is done using a database containing 200 images: 120 images are used for training, 40 for validation and 40 for testing. Five-fold cross-validation is used such as each image is in the testing set at some point. For each pixel, the network predicts its probability of belonging inside the page. Probability maps are then threshold and post-processed by finding the smallest enclosing rectangle.

4.2 Building logical rules

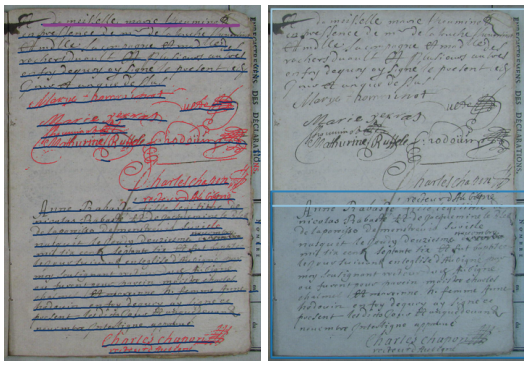
Once these patterns are extracted, logical rules are applied so that the patterns that belong to a same record are grouped together. Logical rules are implemented using an open framework based on DMOS and Enhanced Position Formalism (EPF) [19]. Although most images depict double-page documents, some of them feature single-page documents. As a consequence, the first step is to delimit each page in double-page documents. To this end, we try to find a page separator by applying a filter or by using text alignment. If no separator is found, the assumption is made that the image contains only one page.

Logical rules are then applied to each page. A page is defined as a group of records, but a record can be defined by several rules. The main rule states that a record is composed of a first text-line followed by a group of text-lines and a signature. The pseudo-code of this main rule is provided below:

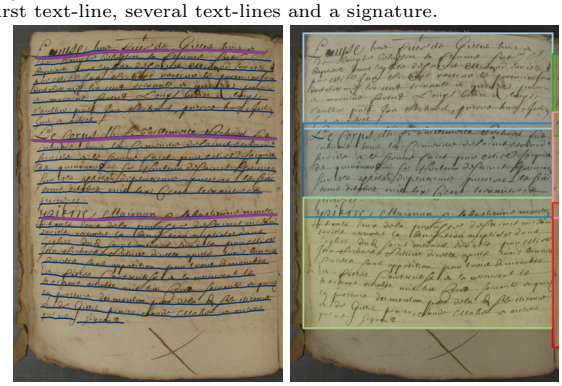
```
record := AT(topPage) &&
          firstTextLine FTL &&
          AT(under FTL) &&
          signature S &&
          AT(between FTL S) &&
          textLines TLS.
```



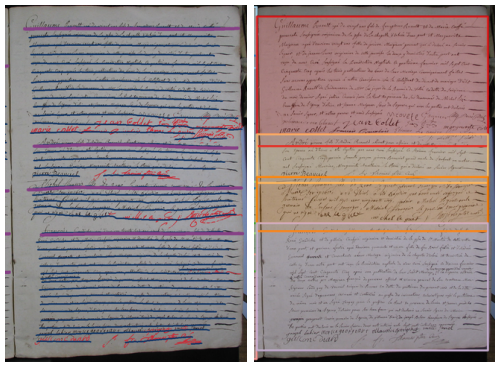
(a) Expected configuration: each record is composed of a first text-line, several text-lines and a signature.



(b) Missed first text-line: the signature is used to spot the end of the record if the first text-line is missed.



(c) Missing signatures: the records are not signed but are found thanks to the first text-lines.



(d) False positive signature: the small signature located in the last record is ignored.

Fig. 7: Illustration of frequent record configurations. For each sub-figure, the image located at the left depicts the patterns predicted by neural networks: first text-lines are shown in purple, text-lines in blue and signature in red. The image located at the right shows the output of the rules. Figure best viewed in color.

But several other rules are designed to overcome frequent prediction errors of signatures and first text-lines. The rules have been designed to obtain a trade-off between split and merge errors. These rules mostly rely on first text-lines to delimit the records, as they are more accurately predicted than signatures. However, in some configurations, signatures are also used to find the end of the record. After these rules are applied, the bounding box of each group is extracted to obtain the outline of each record, as shown in Fig. 4.

Fig 7 illustrates some frequent cases. Fig. 7a shows the main configuration: each record is composed of a first text-line followed by text-lines and a signature. The last record is cut due to the end of the page. In Fig 7b, the first text-line of the second record is not found. In this case, the record is detected thanks to the signature of the first record that is big enough to be considered reliable. In Fig. 7c, the records are not signed by the priest. In such cases where signatures are missing or missed by the network, the first text-lines are used to delimit the records. Finally, in Fig. 7d, a small false positive signature is found in the middle-right of the last record. In this case, the signature is not considered reliable and is ignored.

In this section, we have introduced our original contribution: Deep Syntax. In section 5, we introduce the database and metrics used for evaluation. Finally, results are presented and discussed in section 6.

5 Databases and evaluation protocols

In this section, the two databases used for this work are presented, as well as the protocol used to evaluate record detection.

5.1 Databases

We introduce the BMS database that contains images of French parish registers. Due to proprietary reasons, this database cannot be published. As a result, this work is evaluated on the Esposalles public database [59] as well. Table. 1 summarizes their main characteristics, and a sample of each database is presented in Fig. 8.

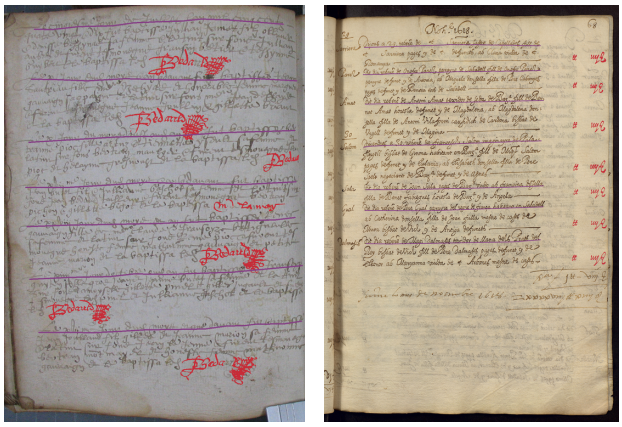
5.1.1 The BMS database

These documents are provided by Les Archives Départementales d'Ille-et-Vilaine (35, France). The corpus contains over 300,000 images of parish registers from 50 parish churches, dating from 16th to 19th century. Documents from this database are heterogeneous. Most images feature double-page documents, but some of them feature single-page documents. Since the documents come from different churches and time periods, each document was written by a different writer. As a result, there is a wide variety of writing styles. This variability has already been partially presented in Fig. 3. From this corpus, we have annotated two subsets:

1. **The experimental subset: BMS-1-expe**. This subset contains 200 images (1,565 records) from four different years: 1675, 1715, 1750, and 1775. Four registers were extracted from each one of the 50 churches, and the seventh image of each register was selected to avoid blank pages. This subset is used for training, validation and evaluation using five-fold cross-validation.
2. **The testing subset: BMS-2-test**. This subset contains 209 images (2,143 records) from 1500 to 1775. The registers were selected at random from the 50 churches so that they span the whole period. From each register, the image number is also selected at random. This subset is used exclusively for evaluation.

The language and writing style of priests strongly evolved over time. Thus, the BMS-1-expe subset is biased, as only four years are represented. Moreover, the image number is fixed. As opposed, the BMS-2-test subset uniformly covers three centuries, and the image number is randomized. One of the aims of our work is to evaluate the ability of each method to generalize well on the BMS-2-test subset, while learning from the small, non-representative BMS-1-expe subset.

Each record is manually annotated such as its bounding box contains the corresponding text and signatures. The width of each record is then normalized to be consistent with page borders. Since the text can be skewed and signatures often overlap on text, successive bounding boxes often overlap as well. The recurrent patterns used for training Deep Syntax are annotated for the BMS-1-expe subset only: page borders, text-lines, first text-lines and signatures.



(a) BMS **(b)** Esposalles

Fig. 8: Samples from each evaluation database. First text-lines are overlaid with a purple line. Ending patterns are overlaid in red: signatures for the BMS database, tax for the Esposalles database. Figure best viewed in color.

5.1.2 The Esposalles database

The Esposalles database was introduced in [59]. It was notably used for the ICDAR2017 Competition on Information Extraction in Historical Handwritten Records [29].

This database is composed of 125 pages: 75 for training, 75 for validation, and 25 for testing. It consists of historical handwritten marriage records from the Archives of the Cathedral of Barcelona. Each image features a single-page document extracted from a volume written in old Catalan, from the 17th century.

Each marriage record contains information about the husband, his wife, as well as their parents. In this regard, these documents are similar to French parish registers. However, there are some major differences. Besides the change of language, the structure also appears more clearly. The records were all written by the same writer over a short period, as a result the corpus is very homogeneous. Also, the records are not signed, but rather are marked with a tax symbol at the end. We propose to take advantage of this symbol as the ending pattern for Deep Syntax.

Table 1: Comparison of the two databases used for this work.

Database	BMS	Esposalles
Writers	One writer for each document	One writer for the database
Period	1500-1790	1617-1619
Origin	50 churches, France	1 church, Spain
Layout	Mostly double page	Simple page
Records	Baptism, marriage, burial	Marriage
Patterns	First text-lines and signatures	First text-lines and tax symbol
Variability	+	-

Previous work on this database focus on handwritten text recognition, yet we use this database for structure detection. We produced a ground truth annotation for each record. The bounding boxes of each record have been partially built from the ground truth proposed in [29], by taking the enclosing rectangle of the words belonging to the same text region. Then, the width of each record has been normalized to page borders to include marginal annotations and tax symbols. We have also annotated page first text-lines and tax symbols. These annotations are freely available[4]. One particularity of this corpus is that some marriage records were voided. These voided records are not annotated in the ground truth, but look very similar to regular records.

5.2 Evaluation protocols

Many evaluation metrics have been proposed to assert the quality of document layout analysis systems. However, few of them are able to process overlapping ground truth zones belonging to the same class. We select several metrics designed to assert the quality of object detection methods, as well as quantify each type of errors.

5.2.1 Surface evaluation

This ZoneMap metric [31] has been specifically designed for the detection and classification of areas in scanned documents, as part of the Maurdor International evaluation campaign [10]. The ZoneMap score summarizes a surface error that is computed on foreground pixels. Reference and hypothesis zones are incrementally associated based on their overlap such as they are in one of the following configurations: Match (one-to-one), Miss (one-to-zero), False Alarm (zero-to-one), Split (one-to-many), Merge (many-to-one). For each configuration, a specific surface error is computed on foreground pixels. A perfect detection corresponds to a ZoneMap score of 0. However, ZoneMap scores can exceed 100 if large zones are inaccurately detected. As a consequence, the metric score is hard to interpret. Nevertheless, it is useful for comparing different methods on a given dataset.

We also report the average precision (AP) as it is very common in object detection tasks. This metric is more clear and straightforward than ZoneMap, but was not designed for this task. We report AP for two Intersection over Union (IoU) thresholds. We denote AP@.50 the average precision computed with IoU = 0.5 and AP@.75 when computed with IoU = 0.75.

[4] https://gitlab.inria.fr/starride/structure-esposalles

5.2.2 Matching evaluation

Counts of *Miss*, *False Alarm*, *Split*, *Merge* and *Match* are computed by the ZoneMap evaluation tool [31]. These numbers are presented to compare qualitative errors between methods. As the number of matching records must be analyzed with respect to the number of ground truth and predicted records, we also compute the precision, recall and F1-score.

6 Evaluation of the Deep Syntax and Object Detection systems for record detection

In this section, we present the experiments performed for each strategy on the BMS-1-expe subset. Finally, these two strategies are compared on three databases: the BMS-1-expe subset, the BMS-2-test subset, and the Esposalles database.

6.1 Experiments on the Object Detection system

First, we present the experiments performed on object detection networks, and select the optimal architecture and training setup.

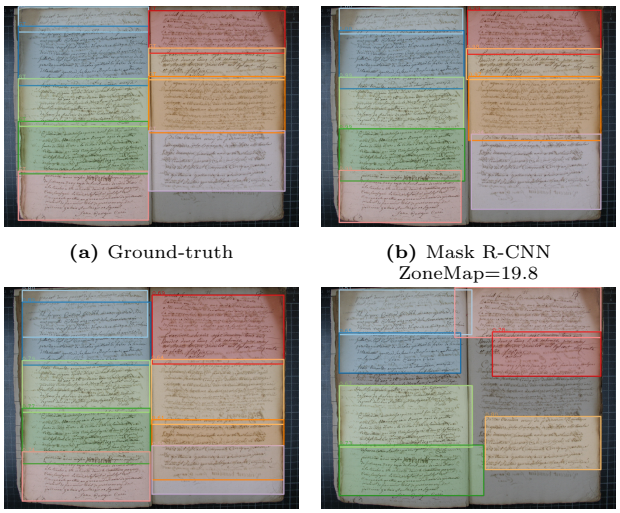
6.1.1 Comparison of three architectures

We compare the performance of Mask R-CNN [35], Retina-Net [43] and YOLOv3 [56] on the BMS-1-expe subset. All networks were pre-trained on COCO. Five-fold cross-validation is used, as described in section 3, using 120 images for training, 40 images for validation and 40 images for testing. As a result, 200 images (1,565 records) are evaluated.

Table 2 summarizes the performance of each network. The BMS database presents several main difficulties: no clearly delimited frontiers, high intra-class variability, overlapping objects and class imbalance. We observe that YOLOv3 struggles to detect the records. One possible explanation could be that the feature extractor might not be able to learn such difficult objects using few training data. As opposed, Mask R-CNN and RetinaNet output acceptable results. They are both based on Feature Pyramid Networks that generate multi-scale feature maps. This feature extraction strategy can explain their performance. RetinaNet also takes advantage of the focal loss to overcome the class imbalance. Despite this, Mask R-CNN outperforms RetinaNet. The superiority of Mask R-CNN likely comes from its proposal mechanism. In two-stage detectors, such as Mask R-CNN, the model proposes a set of regions of interest, then a classifier processes the region candidates.

Table 2: Performance of each network on the BMS experimental subset (200 images, 1,565 records)

Model	ZoneMap	AP@.50	AP@.75
Mask R-CNN	**31.9**	**86.8**	**66.1**
RetinaNet	47.4	68.9	36.5
YOLOv3	76.3	24.3	0.8



(a) Ground-truth

(b) Mask R-CNN ZoneMap=19.8

(c) Retina-Net ZoneMap=32.2

(d) YOLOv3 ZoneMap=76.7

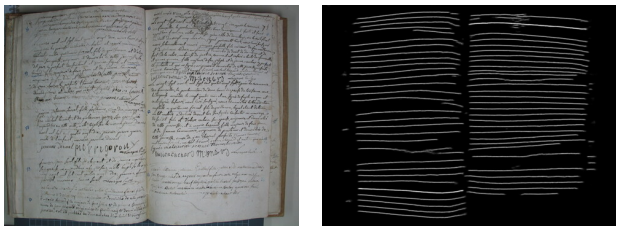
Fig. 9: Performance of each method on a single image from the BMS experimental subset - Each record is depicted with a distinct color. A ZoneMap score is computed for each prediction. Figure best viewed in color.

Two-stage approaches are generally more accurate than one-stage networks.

Fig. 9 shows an illustration of the results on a single image for each network. The output from Mask R-CNN looks close to perfect since all the records are correctly found. However, the score is penalized by small surface errors, mostly due to bleed-through in the last record. RetinaNet outputs zones with correct width on both pages, however, they are highly imprecise, with large overlaps between two successive records. It also triggers many merge errors. YOLOv3 struggles to find relevant zones, especially on the right page. Zone widths are not consistent with page borders, and many records are missed or merged. Moreover, records frontiers are highly imprecise.

6.1.2 Experiments on Mask R-CNN

We have shown that Mask R-CNN outperforms other architectures. In this section, further experiments are performed, with a focus on the input, backbone architecture and data augmentation. Once the prediction is done, the width of the records are post-processed based on predicted page borders. All the results are presented in Table 3.



(a) Original image (I). **(b)** Predicted text-lines (T).

Fig. 10: Input: the original image (I) is concatenated with the text-lines (I+T) predicted using ARU-Net [33].

First, we compare the performance of Mask R-CNN when trained on raw images (I) or on images associated with their predicted text-lines (I+T), as depicted in Fig. 10. This idea is motivated by the structure of the records that strongly depends on text-lines. Overall, we find that using the image concatenated with predicted text-lines as an input consistently removes recurrent errors. It especially helps on images featuring bleed-through, as the text-line detector is able to distinguish bleeding text-lines from actual text-lines. Using text-line prediction as an input reduces the number of false positives records.

We also compare the results when using a ResNet-50 or ResNet-101 as a backbone. Both architecture were pre-trained on COCO. Overall, the backbone architecture does not have a significant impact on the performance.

Experiments on data augmentation have also been carried out. By default, a simple augmentation composed of random flips and Gaussian blur is applied. The best model was also trained with more advanced augmentation consisting of a combination of horizontal flips, crops, Gaussian blur, contrast normalization, Gaussian noise, color variations and affine transformations. However, this did not improve the performance.

This study shows that Mask R-CNN outperforms other architectures. Using simple augmentation and training on both images and text-lines (I+T) also improves performance, as it helps to learn the underlying structure faster. In the following, we refer to the best model, trained with ResNet-101, using I+T and simple data augmentation, as the Object Detection system.

6.2 Experiments on Deep Syntax

In this section we evaluate the segmentation of recurrent patterns. We also compare several hybrid approaches and show the interest of relying on multiple patterns.

Table 3: Performance of each experiment on the testing set composed of 1,565 records. 5-fold cross-validation is used, with 120 images for training, 40 for validation and 40 for testing. As an input, the image (I) can be associated with predicted text-lines (T). For a description of the experiments, see section 6.1.2.

	Training parameters			Scores		
Comments	Backbone	Input	Augmentation	ZoneMap	AP@.50	AP@.75
Model from Table 2	ResNet-50	I	Simple	31.9	86.8	66.1
	ResNet-50	I+T	Simple	30.1	**91.9**	73.5
	ResNet-101	I	Simple	29.6	88.5	70.6
Selected model	ResNet-101	I+T	Simple	**29.1**	89.6	**73.9**
	ResNet-101	I	Advanced	39.2	81.8	32.2
	ResNet-101	I+T	Advanced	35.9	87.1	38.8

6.2.1 Evaluation of segmented patterns

All patterns were annotated and evaluated on the experimental database. For each training, five-fold cross-validation was used so that each image is in the test set once. The results presented in this section are obtained on the experimental subset of the BMS database.

Predicted page borders are evaluated pixel-wise. The precision, recall, F1 and IoU scores are presented in Table 4. Results are close to perfect for most images, but errors appear on documents featuring paper in the background. However, these errors have little impact on record segmentation as long as the left and right border are correctly found.

First text-lines are evaluated using the method described in [34]. The evaluation tool computes several scores for each image in order to assert the quality of the detection. The scores are then averaged over the images to get the final metrics. The results are presented in Table 4. Recurrent errors have been described in section 4. Many errors come from overlaps between first text-lines and seals or signatures. Despite these recurrent confusions, first text-line prediction is overall acceptable.

Post-processed signatures are evaluated pixel-wise. Evaluation on foreground pixels is presented in Table 4. Although not all relevant pixels are selected, those that are selected are relevant: hence the low recall but high precision. Ground truth often contains noise induced by binarization, especially in documents featuring ink stains, bleed-through or low contrast. As a consequence, false positives are often due to imprecise segmentation of signature outlines: 52% of false positive pixels are located on the frontiers of ground truth signatures. As a result, these false positives do not lead to false positive records.

Table 4: Evaluation of patterns on the BMS experimental subset. Pixel-wise evaluation of predicted page borders and signatures are presented. For signatures, scores are computed on foreground pixels. For first text-lines, the metric described in [34] is used.

Task	Method	Precision	Recall	F1
Page borders	Pixel-wise	0.97	0.99	0.98
Signatures	Pixel-wise	0.85	0.48	0.61
First text-lines	Custom [34]	0.92	0.91	0.92

Table 5: Surface evaluation of Deep Syntax on the BMS experimental subset when using different patterns.

Patterns used	ZoneMap	AP@0.50	AP@0.75
Signatures	32.1	78.8	49.1
First text-lines	28.4	87.2	54.4
Both	**27.1**	**89.5**	**69.8**

6.2.2 Advantage of using multiple patterns

In this section, we show the advantage of using multiple patterns for record segmentation in the BMS database.

In a previous section 4.2, we have presented several examples where using multiple patterns is relevant. Combining multiple patterns allows to overcome two main difficulties. First, the ending pattern does not appear systematically in the BMS database because some records are not signed. Secondly, predicted signatures and first text-lines can be missed or mistaken due to high intra-class variability (multiple writing styles, phrasing and layout). Using both patterns helps to increase performance on the BMS database, as shown in Table 5. Applying the logical rules with both patterns yields to an increase of the AP@0.75 score of 28% when compared with only first text-lines, and 42% when compared with only signatures.

For easier databases, such as the Esposalles database, combining multiple patterns might not be essential. As there is only one writer that uses consistent phrasing, layout and writing style is used for each record. Con-

sequently, first text-lines and tax symbols have a lower intra-class variability and are more easily learned. Besides, the tax symbol appears systematically at the end of each record. As a consequence, using either one of these patterns would likely yield good performance.

6.3 Comparison on the BMS-1-expe heterogeneous subset

In this section, the Deep Syntax and Object Detection systems are evaluated on the BMS-1-expe subset. The training is carried out using five-fold cross-validation. As a result, the dataset is split into 5 sets of 40 images, and 5 models are trained using 120 images for training, 40 images for validation, 40 images for testing. At a result, each image has been tested once, and the mean test error can be computed.

This subset allows the evaluation of both systems on documents from the same period as the training set, but written by different priests. Scores are presented in Table. 6. Both systems obtain acceptable performance on this subset. Results suggest that more match configurations are found using Deep Syntax, but that the object detection network outputs better bounding boxes. It is also worth noting that the object detection network tends to merge records, especially the smallest ones. As opposed, Deep Syntax produces more split errors, that might be due to false positive patterns. There are several false alarms that mostly appear on pages featuring paragraphs written by priests to describe the registers. In some cases, false alarms also appear on titles or on pages featuring bleed-through. If both strategies yield good precision, the Deep Syntax system yields a higher recall. In that regard, Deep Syntax outperforms slightly the Object Detection system.

However, our main concern is to select a system that could be applicable to a wide variety of parish registers. In that regard, we are interested in finding which system is able to adapt to documents from different time periods, with different phrasing and writing styles.

6.4 Assessment of the generalization ability when trained on few examples

In this section, both systems, Deep Syntax and Object Detection are trained and evaluated on the Esposalles public database, showing that they can both learn homogeneous layouts from very few training data. The two approaches are also compared on the BMS-2-test subset to evaluate the ability of both systems to process heterogeneous documents from different time periods. A study regarding the number of training examples needed in each situation is also carried out, as this assessment is critical within our industrial context.

Table 6: Evaluation on the BMS-1-expe subset (1,565 records).

(a) Surface evaluation

	Object Detection	Deep Syntax
ZoneMap	29.1	**27.1**
AP@0.5	**89.6**	89.5
AP@0.75	**73.9**	69.8

(b) Matching evaluation

	Object Detection	Deep Syntax
Match	1293	**1401**
Split	**40**	71
Merge	107	**42**
False Alarm	22	**16**
Miss	2	**1**
Precision	0.86	**0.87**
Recall	0.83	**0.90**
F1-score	0.84	**0.88**

6.4.1 Processing homogeneous documents with few training examples

First, both systems are evaluated on the Esposalles database. These registers feature a different layout than French parish registers: records are smaller, there are no signatures, and the language is different. As a result, both systems must be re-trained on this database. We investigate the influence of the training set size on the performance of each system. Then the results of both strategies are compared and discussed.

Both systems are trained on sub-sampled subsets of the Esposalles training database, using 10, 25, 50 or 75 training examples. Detailed scores are presented in Table. 7. They show that good performance can be achieved using few training data on this database. Indeed, the results suggest that both systems become efficient from 25 training documents, which corresponds to approximately 250 records. This fast learning can be explained by the homogeneity of the records, as there is only one writer over a short period of time. For Deep Syntax, experiments show that the tax symbol is well learned from 10 images, while first text-lines are well learned from 25 images. Fig. 11 depicts the prediction produced by each strategy on a single image in this condition.

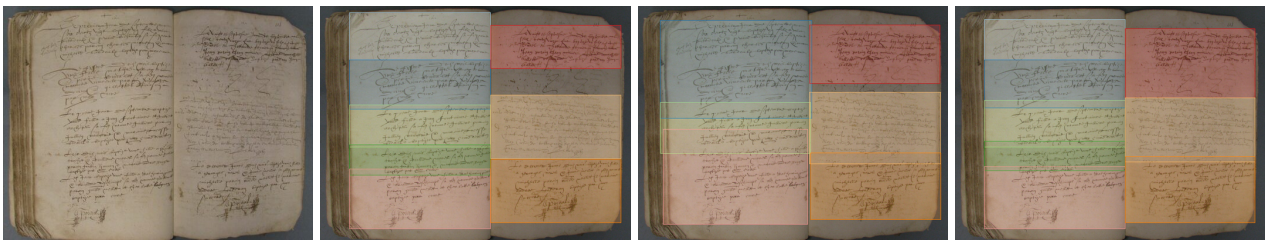
If both systems manage to accurately recognize the records, the Object Detection system manages to output boxes that fit very well to the ground truth. The ZoneMap surface error consistently decreases as the training set size increases for the Object Detection,

(a) Original image **(b)** Ground truth **(c)** Object Detection **(d)** Deep Syntax

Fig. 11: Comparison of both systems on the Esposalles database, when trained on 25 images. Figure best viewed in color.

Table 7: Evaluation on 253 records from the Esposalles database.

		Object Detection				Deep Syntax			
Training set size (Esposalles)		10	25	50	75	10	25	50	75
Surface evaluation	ZoneMap	18.5	14.6	12.8	**11.9**	17.1	13.9	14.3	14.1
	AP@0.5	96.1	98.3	98.7	**99.1**	95.1	97.6	97.2	98.2
	AP@0.75	70.0	85.6	89.5	**89.6**	74.2	82.3	82.6	81.4
Matching evaluation	Match	241	249	**252**	251	240	248	248	249
	Split	4	2	**1**	2	5	6	5	4
	Merge	4	1	**0**	**0**	1	**0**	**0**	**0**
	False Alarm	2	5	3	**1**	3	3	3	3
	Miss	**0**	**0**	**0**	**0**	6	**0**	**0**	**0**
	Precision	0.95	0.98	0.98	0.98	0.95	0.95	0.95	0.96
	Recall	0.95	0.98	1.00	0.99	0.95	0.98	0.98	0.98
	F1-score	0.95	0.98	0.99	0.98	0.95	0.96	0.96	0.97



(a) Original image **(b)** Ground truth **(c)** Object Detection **(d)** Deep Syntax

Fig. 12: Comparison of both systems on the BMS-2-test subset when trained on 120 images of the BMS-1-expe subset. Figure best viewed in color.

Table 8: Evaluation on the 2,143 records of the BMS-2-test subset.

		Object Detection			Deep Syntax		
Training set size (BMS-1-expe)		60	120	180	60	120	180
Surface evaluation	ZoneMap	23.0	20.2	17.9	14.7	14.1	**13.7**
	AP@0.5	80.1	77.4	82.0	83.3	**84.0**	83.1
	AP@0.75	52.3	59.0	**63.8**	59.6	62.8	61.1
Matching evaluation	Match	1576	1547	1653	1762	1787	**1794**
	Split	54	**14**	21	80	56	67
	Merge	205	224	183	122	121	**105**
	False Alarm	**2**	**2**	4	4	3	5
	Miss	5	5	4	**1**	**1**	4
	Precision	0.83	0.86	**0.88**	0.85	**0.88**	**0.88**
	Recall	0.74	0.72	0.77	0.82	0.83	**0.84**
	F1-score	0.78	0.78	0.82	0.84	**0.86**	**0.86**

while it remains almost constant for Deep Syntax. A possible explanation is that Deep Syntax produces bounding boxes constrained by rules. As a consequence, small surface errors remain, even with perfect pattern predictions. As opposed, object detection networks learn to adapt to each record.

The matching evaluation shows that errors occurs on few records, while the large majority of records are correctly found. Deep Syntax produces 3 false positives. They correspond to the three voided records that appear in the testing set. We observe almost no merge errors or miss errors, however, several split errors are created. For the Object Detection system, there are also a few split and merge errors, however, they tend to disappear when the training set size increases. The three records that have been voided in the testing set present difficulties for both system, even when trained with the maximum number of images.

6.4.2 Processing heterogeneous documents when trained on a small, non reprentative subset

Both systems are evaluated on the BMS-2-test, when trained on 60, 120, or 180 documents from the BMS-1-expe subset. This BMS-2-test subset is difficult to process, as it contains images from periods that are not represented in the training set. As the writing style and the language evolve with time, older records appear different. This subset is representative of the difficulties faced in an industrial context, as collecting and annotating a large, representative database is not achievable. The results are presented in Table. 8, and an example of prediction can be observed in Fig. 12.

As compared to the evaluation on the BMS-1-expe subset, a significant drop of performance is observed on the Object Detection system: the F1-score decreases by 7% when using 120 training images. A possible explanation is that the documents from the BMS-2-test subset are older and feature tighter layouts. As a result, successive records are often merged. If increasing the number of training examples does increase the performance a bit, it is not sufficient to compete with those obtained by Deep Syntax. Fig. 12 shows that the bounding boxes produced by the Object Detection system are less accurate, and are therefore not usable in practice. That being said, the object detection system would certainly improve if more documents from more time periods were also annotated.

As opposed, Deep Syntax's performance tends to remain stable over both subsets. The main difference is that the system produces more merge errors than split errors on the BMS-2-test subset, while it produces more split errors than merge errors on the BMS-1-expe subset. This difference can be linked to the tighter layouts that compose the BMS-2-test subset. As the size of the training set increases, the performance slowly increases.

When trained on the same training set size, Deep Syntax produces in average 12% more match configurations while reducing the ZoneMap surface error by 30%. When trained on three times less data, Deep Syntax still manages to output 7% more match configurations and to lower the ZoneMap score by 18%. Thus, it would be more easily applicable for massive processing of French parish registers.

6.5 Discussion

In this section, we discuss the strengths and weaknesses of both approaches.

The Object Detection system yields good performance on the Esposalles database because the records are structured and appear similar. On this database, it easily outperforms Deep Syntax, as bounding boxes fit the ground truth very well. However, performance decreases on more complex databases. On the BMS database, the Object Detection system tends to retrieve paragraphs. As a consequence it struggles to find small records and has trouble performing well on tight layouts. Predicted bounding boxes are imprecise, with large overlaps between successive records. The Object Detection system also produces recurrent errors on this database: small records are merged or missed, and large records are split. We believe that more training images should be used to capture the variability of the records. Indeed, the performance drops even more on the BMS-2-test, that contains images that look different from training data. However, producing more ground truth annotations would require a tremendous amount of time and effort.

In contrast, the Deep Syntax system relies on simpler objects, e.g. signatures and text-lines, that can be learned from few examples. As a result, this strategy outputs strong results for record detection, even if few available training data are available. The main limitation of Deep Syntax is that its workflow is complex. Moreover, bounding boxes are constrained by rules, so they do not adapt to the specificity of each record. Despite these, we argue that Deep Syntax is perfectly applicable in the context of record detection. Taking advantage of structural patterns helps to simplify the detection task. Rather than learning to recognize a complex object, it can be easier to learn to recognize the separation between these objects. Moreover, using multiple patterns helps to strengthen the output, but is not required to obtain acceptable results.

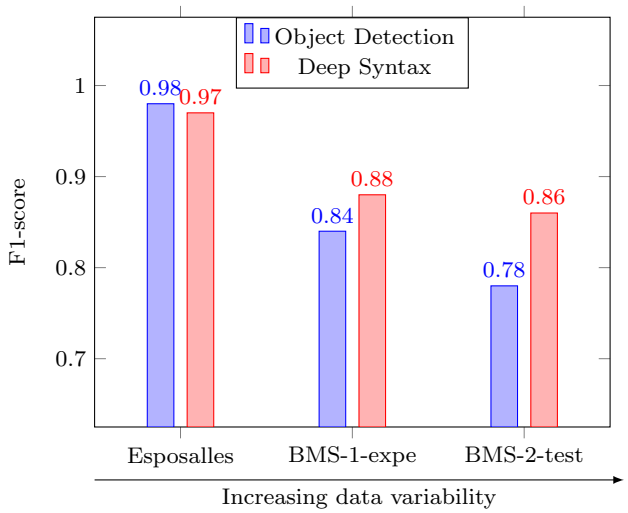


Fig. 13: F1-score on each database for both systems, when both systems are trained on 25 documents for the Esposalles database, and 120 documents for the BMS database. The database variability increases from left to right, showing that Deep Syntax generalizes better than the Object Detection system.

7 Conclusion

In this paper, we have presented two strategies for record detection in historical parish registers. The first one is based on object detection networks. We have compared three architectures in similar training conditions and performed several experiments on the architecture that seems best adapted for this task: Mask R-CNN. The second one is our original contribution, Deep Syntax, that relies on a combination of u-shaped networks and logical rules. Recurrent patterns are predicted using neural networks: page borders, text-lines, first text-lines and signatures.

We have studied their applicability within the context of massive data processing, where only a few data are available for training. To this end, we have compared both systems when trained with different training set sizes. We have also applied both systems to a complex subset of parish registers, featuring documents from various time periods that were not represented in the training set. Finally, we have applied them to the homogeneous records of the Esposalles public database, to ease future comparison with this work.

We observe that object detection networks achieve very good performance when they are applied to a homogeneous database: only 25-50 training documents are required to obtain a F1-score of 0.99 on the Esposalles database. However, they struggle on heterogeneous documents. Fig. 13 shows that their performance drops when the corpus variability increases. For instance, when trained on 120 pages of parish registers, Mask R-CNN is not able to generalize well to parish registers from other time periods. The results suggest that object detection networks require a lot more training data to handle heterogeneous documents, as the training database must be representative of the corpus to process. Typically, thousands of annotated documents are used to detect tables and figures in PDF documents [61, 68]. But in the context of massive processing of archival documents, the task is even more complex: documents are poorly-structured, unevenly photographed, and feature various writing styles and degradation. As a result, object detection networks would require a substantially large training database, including documents from various time periods and locations. Collecting and annotating such a database is not always achievable in practice. This limitation highlights the interest of using hybrid methods that learn from few examples, such as Deep Syntax. As it relies on simpler, recurrent patters, it learns from few training examples while being able to generalize well on documents from different time periods. When trained on the same training database, Deep Syntax is able to produce 12% more matching configurations than Mask R-CNN, while reducing the ZoneMap surface error metric by 30%, in average. Deep Syntax also outperforms Mask R-CNN when trained on a database three times smaller. We plan to apply Deep Syntax to parish registers from all over France, from 1550 to 1790. This project aims to ease the reading of these registers by genealogists.

Future works will focus on the textual content of each record. The first step would be to spot recurrent keywords in parish registers, as it would help to classify each record into its corresponding type, e.g. marriage, baptism, burial. But recurrent keywords could also be integrated to Deep Syntax to make record detection more reliable. The second step would be to achieve text recognition, in order to extract information that would be helpful to genealogists.

Acknowledgements The BMS database is provided by Les Archives Départementales d'Ille-et-Vilaine, 35, France.

References

1. Alaasam, R., Kurar, B., El-Sana, J.: Layout analysis on challenging historical arabic manuscripts using siamese network. In: 2019 International Conference on Document Analysis and Recognition (ICDAR), pp. 738–742 (2019)
2. Alberti, M., Pondenkandath, V., Würsch, M., Ingold, R., Liwicki, M.: Deepdiva: A highly-functional python framework for reproducible experiments. CoRR **abs/1805.00329** (2018)
3. Alberti, M., Vögtlin, L., Pondenkandath, V., Seuret, M., Ingold, R., Liwicki, M.: Labeling, cutting, grouping:

an efficient text line segmentation method for medieval manuscripts. CoRR **abs/1906.11894** (2019)
4. Antonacopoulos, A., Gatos, B., Bridson, D.: Page segmentation competition. In: Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), vol. 2, pp. 1279–1283 (2007)
5. Asi, A., Cohen, R., Kedem, K., El-Sana, J.: Simplifying the reading of historical manuscripts. In: 2015 13th International Conference on Document Analysis and Recognition (ICDAR), pp. 826–830 (2015)
6. Baechler, M., Liwicki, M., Ingold, R.: Text line extraction using dmlp classifiers for historical manuscripts. In: 2013 12th International Conference on Document Analysis and Recognition, pp. 1029–1033 (2013)
7. Barlas, P., Adam, S., Chatelain, C., Paquet, T.: A typed and handwritten text block segmentation system for heterogeneous and complex documents. In: 2014 11th IAPR International Workshop on Document Analysis Systems, pp. 46–50 (2014)
8. Benjlaiel, M., Mullot, R., Alimi, A.M.: Multi-oriented handwritten annotations extraction from scanned documents. In: 2014 11th IAPR International Workshop on Document Analysis Systems, pp. 126–130 (2014)
9. Bolshakov, I.A., Gelbukh, A.: Text segmentation into paragraphs based on local text cohesion. In: V. Matoušek, P. Mautner, R. Mouček, K. Taušer (eds.) Text, Speech and Dialogue, pp. 158–166. Springer Berlin Heidelberg, Berlin, Heidelberg (2001)
10. Brunessaux, S., Giroux, P., Grilhères, B., Manta, M., Bodin, M., Choukri, K., Galibert, O., Kahn, J.: The maurdor project: Improving automatic processing of digital documents. In: 2014 11th IAPR International Workshop on Document Analysis Systems, pp. 349–354 (2014)
11. Bukhari, S., Shafait, F., Breuel, T.: Coupled snakelets for curled text-line segmentation from warped document images. International Journal on Document Analysis and Recognition (IJDAR) **16**, 1–21 (2011). DOI 10.1007/s10032-011-0176-2
12. Bukhari, S.S., Shafait, F., Breuel, T.M.: High performance layout analysis of arabic and urdu document images. In: 2011 International Conference on Document Analysis and Recognition, pp. 1275–1279 (2011)
13. Bulacu, M., Koert, R., Schomaker, L.: Layout analysis of handwritten historical documents for searching the archive of the cabinet of the dutch queen (2007)
14. Carel, E., Burie, J.C., Courboulay, V., Ogier, J.M., Poulain d'Andecy, V.: Multiresolution approach based on adaptive superpixels for administrative documents segmentation into color layers. pp. 566–570 (2015). DOI 10.1109/ICDAR.2015.7333825
15. Chen, K., Seuret, M., Liwicki, M., Hennebert, J., Ingold, R.: Page segmentation of historical document images with convolutional autoencoders. In: 2015 13th International Conference on Document Analysis and Recognition (ICDAR), pp. 1011–1015 (2015)
16. Chen, K., Wei, H., Liwicki, M., Hennebert, J., Ingold, R.: Robust text line segmentation for historical manuscript images using color and texture. In: 2014 22nd International Conference on Pattern Recognition, pp. 2978–2983 (2014)
17. Chen, K., Yin, F., Liu, C.: Hybrid page segmentation with efficient whitespace rectangles extraction and grouping. In: 2013 12th International Conference on Document Analysis and Recognition, pp. 958–962 (2013)
18. Clausner, C., Antonacopoulos, A., Pletschacher, S.: A robust hybrid approach for text line segmentation in historical documents. In: Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012), pp. 335–338 (2012)
19. Coüasnon, B.B., Lemaitre, A.: DMOS, It's your turn ! In: 1st International Workshop on Open Services and Tools for Document Analysis (ICDAR-OST). Kyoto, Japan (2017). URL https://hal.inria.fr/hal-01659131
20. Coüasnon, B.: Dmos, a generic document recognition method: Application to table structure analysis in a general and in a specific way. IJDAR **8**, 111–122 (2006). DOI 10.1007/s10032-005-0148-5
21. Cruz, F., Terrades, O.R.: Em-based layout analysis method for structured documents. In: 2014 22nd International Conference on Pattern Recognition, pp. 315–320 (2014)
22. Diem, M., Kleber, F., Sablatnig, R.: Text classification and document layout analysis of paper fragments. In: 2011 International Conference on Document Analysis and Recognition, pp. 854–858 (2011)
23. Diem, M., Kleber, F., Sablatnig, R.: Text line detection for heterogeneous documents. In: 2013 12th International Conference on Document Analysis and Recognition, pp. 743–747 (2013)
24. Diem, M., Kleber, F., Sablatnig, R., Gatos, B.: cbad: Icdar2019 competition on baseline detection. In: 2019 International Conference on Document Analysis and Recognition (ICDAR), pp. 1494–1498 (2019)
25. Ferilli, S., Biba, M., Esposito, F., Basile, T.M.A.: A distance-based technique for non-manhattan layout analysis. In: 2009 10th International Conference on Document Analysis and Recognition, pp. 231–235 (2009)
26. Fernández, F.C., Terrades, O.R.: Document segmentation using relative location features. In: Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012), pp. 1562–1565 (2012)
27. Filippova, K., Strube, M.: Using linguistically motivated features for paragraph boundary identification. pp. 267–274 (2006). DOI 10.3115/1610075.1610114
28. Fischer, A., Baechler, M., Garz, A., Liwicki, M., Ingold, R.: A combined system for text line extraction and handwriting recognition in historical documents. In: 2014 11th IAPR International Workshop on Document Analysis Systems, pp. 71–75 (2014)
29. Fornés, A., Romero, V., Baró, A., Toledo, J.I., Sánchez, J.A., Vidal, E., Lladós, J.: Icdar2017 competition on information extraction in historical handwritten records. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), vol. 01, pp. 1389–1394 (2017)
30. Gaceb, D., Eglin, V., Lebourgeois, F., Emptoz, H.: Application of graph coloring in physical layout segmentation. In: 2008 19th International Conference on Pattern Recognition, pp. 1–4 (2008)
31. Galibert, O., Kahn, J., Oparin, I.: The zonemap metric for page segmentation and area classification in scanned documents. In: 2014 IEEE International Conference on Image Processing (ICIP), pp. 2594–2598 (2014). DOI 10.1109/ICIP.2014.7025525
32. Garz, A., Sablatnig, R., Diem, M.: Layout analysis for historical manuscripts using sift features. In: 2011 International Conference on Document Analysis and Recognition, pp. 508–512 (2011)
33. Grüning, T., Leifert, G., Strauß, T., Labahn, R.: A two-stage method for text line detection in historical documents. CoRR **abs/1802.03345** (2018)
34. Grüning, T., Labahn, R., Diem, M., Kleber, F., Fiel, S.: Read-bad: A new dataset and evaluation scheme for baseline detection in archival documents. In: 2018 13th IAPR

International Workshop on Document Analysis Systems (DAS), pp. 351–356 (2018)
35. He, K., Gkioxari, G., Dollár, P., Girshick, R.B.: Mask R-CNN. CoRR **abs/1703.06870** (2017)
36. Hebert, D., Paquet, T., Nicolas, S.: Continuous crf with multi-scale quantization feature functions application to structure extraction in old newspaper. In: 2011 International Conference on Document Analysis and Recognition, pp. 493–497 (2011)
37. Jaekyu Ha, Haralick, R.M., Phillips, I.T.: Document page decomposition by the bounding-box project. In: Proceedings of 3rd International Conference on Document Analysis and Recognition, vol. 2, pp. 1119–1122 vol.2 (1995). DOI 10.1109/ICDAR.1995.602115
38. Journet, N., Ramel, J.Y., Eglin, V., Mullot, R.: Document Image Characterization Using a Multiresolution Analysis of the Texture: Application to Old Documents. International Journal on Document Analysis and Recognition **Volume 11**(Number 1), 9–18 (2008). DOI 10.1007/s10032-008-0064-6
39. Kamola, G., Spytkowski, M., Paradowski, M., Markowska-Kaczmar, U.: Image-based logical document structure recognition. Pattern Anal. Appl. **18**(3), 651–665 (2015). DOI 10.1007/s10044-014-0412-8
40. Kumar, J., Abd-Almageed, W., Kang, L., Doermann, D.: Handwritten arabic text line segmentation using affinity propagation. pp. 135–142 (2010). DOI 10.1145/1815330.1815348
41. Lemaitre, A., Camillerapp, J., Coüasnon, B.: Multiresolution cooperation makes easier document structure recognition. IJDAR **11**, 97–109 (2008). DOI 10.1007/s10032-008-0072-6
42. Lemaitre, A., Camillerapp, J., Coüasnon, B.: A perceptive method for handwritten text segmentation. Document recognition and retrieval XVIII **7874** (2011). DOI 10.1117/12.873037
43. Lin, T., Goyal, P., Girshick, R.B., He, K., Dollár, P.: Focal loss for dense object detection. CoRR **abs/1708.02002** (2017)
44. Lin, T., Maire, M., Belongie, S.J., Bourdev, L.D., Girshick, R.B., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: common objects in context. CoRR **abs/1405.0312** (2014)
45. Mehri, M., Gomez-Krämer, P., Héroux, P., Boucher, A., Mullot, R.: Texture feature evaluation for segmentation of historical document images. In: Proceedings of the 2nd International Workshop on Historical Document Imaging and Processing, HIP '13, p. 102–109 (2013). DOI 10.1145/2501115.2501121
46. Mehri, M., Heroux, P., Gomez-Krämer, P., Boucher, A., Mullot, R.: A pixel labeling approach for historical digitized books. pp. 817–821 (2013). DOI 10.1109/ICDAR.2013.167
47. Mehri, M., Héroux, P., Mullot, R., Moreux, J., Coüasnon, B., Barrett, B.: Icdar2019 competition on historical book analysis - hba2019. In: 2019 International Conference on Document Analysis and Recognition (ICDAR), pp. 1488–1493 (2019)
48. Moysset, B., Kermorvant, C., Wolf, C., Louradour, J.: Paragraph text segmentation into lines with recurrent neural networks. In: 2015 13th International Conference on Document Analysis and Recognition (ICDAR), pp. 456–460 (2015)
49. Oliveira, D., Viana, M.: Fast cnn-based document layout analysis. pp. 1173–1180 (2017). DOI 10.1109/ICCVW.2017.142
50. Oliveira, S.A., Seguin, B., Kaplan, F.: dhsegment: A generic deep-learning approach for document segmentation. CoRR **abs/1804.10371** (2018)
51. Ouwayed, N., Belaïd, A.: A general approach for multi-oriented text line extraction of handwritten document. International Journal on Document Analysis and Recognition **14**(4) (2011). DOI 10.1007/s10032-011-0172-6
52. Papavassiliou, V., Stafylakis, T., Katsouros, V., Carayannis, G.: Handwritten document image segmentation into text lines and words. Pattern Recognition **43**(1), 369 – 377 (2010). DOI https://doi.org/10.1016/j.patcog.2009.05.007
53. Peng, X., Setlur, S., Govindaraju, V., Sitaram, R.: Handwritten text separation from annotated machine printed documents using markov random fields. International Journal on Document Analysis and Recognition (IJDAR) **16**, 1–16 (2011)
54. Pinson, S.J., Barrett, W.A.: Connected component level discrimination of handwritten and machine-printed text using eigenfaces. In: 2011 International Conference on Document Analysis and Recognition, pp. 1394–1398 (2011)
55. Prusty, A., Aitha, S., Trivedi, A., Sarvadevabhatla, R.K.: Indiscapes: Instance segmentation networks for layout parsing of historical indic manuscripts (2019)
56. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. CoRR **abs/1804.02767** (2018)
57. Ren, S., He, K., Girshick, R.B., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. CoRR **abs/1506.01497** (2015)
58. Renton, G., Soullard, Y., Chatelain, C., Adam, S., Kermorvant, C., Paquet, T.: Fully convolutional network with dilated convolutions for handwritten text line segmentation. International Journal on Document Analysis and Recognition (IJDAR) (2018). DOI 10.1007/s10032-018-0304-3
59. Romero, V., Fornés, A., Serrano, N., Sánchez, J.A., Toselli, A.H., Frinken, V., Vidal, E., Lladós, J.: The esposalles database: An ancient marriage license corpus for off-line handwriting recognition. Pattern Recognition **46**(6), 1658 – 1669 (2013). DOI https://doi.org/10.1016/j.patcog.2012.11.024
60. Ryu, J., Koo, H.I., Cho, N.I.: Language-independent text-line extraction algorithm for handwritten documents. IEEE Signal Processing Letters **21**(9), 1115–1119 (2014)
61. Saha, R., Mondal, A., Jawahar, C.V.: Graphical object detection in document images. 2019 International Conference on Document Analysis and Recognition (ICDAR) pp. 51–58 (2019)
62. Shafait, F., v. Beusekom, J., Keysers, D., Breuel, T.M.: Structural mixtures for statistical layout analysis. In: 2008 The Eighth IAPR International Workshop on Document Analysis Systems, pp. 415–422 (2008)
63. Tang, Y., Wu, X., Bu, W.: Text line segmentation based on matched filtering and top-down grouping for handwritten documents. pp. 365–369 (2014). DOI 10.1109/DAS.2014.14
64. Tarride, S., Lemaitre, A., Couasnon, B.B., Tardivel, S.: Signature detection as a way to recognise historical parish register structure. In: HIP 2019, pp. 54–59. ACM Press, Sydney, Australia (2019). DOI 10.1145/3352631.3352636
65. Wei, H., Baechler, M., Slimane, F., Ingold, R.: Evaluation of svm, mlp and gmm classifiers for layout analysis of historical documents. In: 2013 12th International Conference on Document Analysis and Recognition, pp. 1220–1224 (2013)

66. Wei, H., Chen, K., Ingold, R., Liwicki, M.: Hybrid feature selection for historical document layout analysis. In: 2014 14th International Conference on Frontiers in Handwriting Recognition, pp. 87–92 (2014)
67. Weliwitage, C., Harvey, A.L., Jennings, A.B.: Handwritten document offline text line segmentation. In: Digital Image Computing: Techniques and Applications (DICTA'05), pp. 27–27 (2005)
68. Yi, X., Gao, L., Liao, Y., Zhang, X., Liu, R., Jiang, Z.: Cnn based page object detection in document images. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), vol. 01, pp. 230–235 (2017). DOI 10.1109/ICDAR.2017.46
69. Yin, F., Liu, C.: A variational bayes method for handwritten text line segmentation. In: 2009 10th International Conference on Document Analysis and Recognition, pp. 436–440 (2009)
70. Yin, F., Liu, C.L.: Handwritten chinese text line segmentation by clustering with distance metric learning. Pattern Recognition **42**(12), 3146 – 3157 (2009). DOI https://doi.org/10.1016/j.patcog.2008.12.013. New Frontiers in Handwriting Recognition
71. Ziaratban, M., Faez, K.: An adaptive script-independent block-based text line extraction. In: 2010 20th International Conference on Pattern Recognition, pp. 249–252 (2010)