# Text Line Detection in Historical Index Tables: Evaluations on a New French PArish REcord Survey Dataset (PARES)

Guillaume Bernard, Casey Wall, Mélodie Boillet, Mickaël Coustaty, Christopher Kermorvant, Antoine Doucet

# Text Line Detection in Historical Index Tables: Evaluations on a New French PArish REcord Survey Dataset (PARES)

Guillaume Bernard[1][0000−0001−5945−4865], Casey Wall[1], Mélodie Boillet[2][0000−0002−0618−7852], Mickaël Coustaty[1][0000−0002−0123−439X], Christopher Kermorvant[2][0000−0002−7508−4080], and Antoine Doucet[1][0000−0001−6160−3356]

[1] Laboratoire L3i – Université de La Rochelle, La Rochelle, France
{guillaume.bernard,casey.wall,mickael.coustaty,antoine.doucet}@univ-lr.fr
[2] TEKLIA, Paris, France
{boillet,kermorvant}@teklia.com

**Abstract.** In this paper, we address the challenge of document image analysis for historical index table documents with handwritten records. Demographic studies can gain insight from the use of automatic document analysis in such documents through the study of population movements. To evaluate the efficacy of automatic layout analysis tools, we release the PARES dataset [6], which contains 250 labeled index table images originating from French archives. Also, we run state-of-the-art algorithms (U-FCN, R-CNN and Transformers) in order to detect the lines within index tables, a common prerequisite for handwritten text recognition (HTR). Our results indicate that text line extraction works well with the U-FCN model, while also indicating that Transformer architectures show promise for accurate text line detection in such historical documents with great efficiency. This is a encouraging step towards a Transformer-based architecture for both layout and content detection. This process and dataset represent a first step to automatically analyze handwritten and historical index tables. In addition to this paper and the PARES [6] dataset of historical index tables of 250 images, we release segmentation masks, the code we used to train and test the models, and the models themselves.

**Keywords:** Dataset, Historical Documents, Document Image Analysis, Document Segmentation, Deep Learning

## 1 Introduction

National archives around the world hold historical documents of various kinds. These include handwritten archives, such as census tables, that may record births, deaths or marriages, to name but a few. As with other types of documents of interest in digital humanities, such as newspapers or photographs, the

indexing of such documents in digital libraries is intended to facilitate scholarly analysis in a controlled environment and to avoid potential damage to the original material.

The analysis of these documents can provide insights into demographic studies. For instance, understanding the diffusion of family names could provide valuable insights when put into the context of relevant historical events. This could also help us to understand patterns of population displacements across specific territories. However, the lack of publicly available historical census table datasets is a challenge. In a broad sense, our methodology of extracting valuable information from historical documents, such as named entities and dates, is not a novel concept within digital libraries. Nonetheless, it appears that only a limited number of researchers specifically focus on historical index or census tables [18,12], as a result of these challenges in analyzing such data. Therefore, tools developed for tables analysis and handwritten text recognition (HTR) shall be useful for the communities interested in historical documents.

The automatic analysis of historical documents, such as books or newspapers, has been widely explored, mainly due to the wide availability of high-quality digitized images. Given the advances in automated processing techniques applied to historical documents, it is now possible to handle more novel and intricate data types, such as tabular data. Therefore, state-of-the-art approaches should be evaluated on this kind of documents.

The contributions of this work are of twofold.

1. We present a novel dataset of historical index tables from the French National Archives: PARES, which stands for PArish REcord Survey [6]. This dataset contains annotated images specifically designed for document layout analysis (DLA). Both the original images and the segmentation annotations are made publicly available on Zenodo.
2. As the aim of processing these documents is to recognize and analyze the handwritten text, in this paper we focus on the preliminary step of text line detection. Hence, we compare multiple state-of-the-art document image analysis deep neural networks on the table text line detection task. We also share the implementations and source codes on Zenodo [5] and Software Heritage [4,2,3].

The paper is organized as follows. In Section 2, we introduce the Historical Index Table dataset. In Section 3, we present the state-of-the-art models for text line detection, including document image segmentation and instance segmentation models. Finally, in Section 4, we train and compare baseline models on the introduced dataset in order to extract the text records from our index tables.

## 2 Historical Index Table Dataset: PARES

Several recent projects have contributed to the release of handwriting table datasets. The READ ABP Tables is a collection of German handwritten records of Diocese Of Passau collected from 1847 to 1878 with more than 200 images [18].

The PoPP dataset [12] contains early 20$^{th}$ century historical census tables for the city of Paris, France. Similarly, the French Socface project also aims at contributing to the research for demographic studies. The collection includes census tables that were written over a span of a century, beginning in 1836. The research primarily emphasizes handwriting text recognition, and the datasets are soon to be made available in open access[3]. HisClima [29] is another partially handwritten set of historical tables of naval weather logbooks from the United States. Last, the Lectaurep project [21] provides French handwritten tables recording French notary deeds. The aforementioned datasets and their corresponding projects center around two primary challenges: analysing document layout and recognizing handwritten text. On another hand, while not being a table dataset, SIMARA [33] is a dataset of handwritten archive finding aids, comprised of metadata describing historical archives. Finding aids are handwritten and feature the same scientific challenges regarding handwritten text recognition. As an index, each finding aids contains expected data, such as the title of a document, its classification number, location, etc, just as in our index dataset.

Within this context, and alongside this paper, we release the PARES dataset consisting of 250 digitised index tables from the French National Archives, together with ground truth layout annotations. Based on visual observation, we employed heuristic methods to identify and label the table headers, table lines, page headers, and footers to create segmentation masks. This is a first step towards handwritten text recognition. Indeed, in this historical research context, we want to localize the text first, then apply HTR tools to make the most of the records. Finally, the extraction of named entities, which represent highly valuable content within these records, plays a crucial role in contextualizing and facilitating a comprehensive understanding of the dataset for demographic studies.

In the following sections, we describe the PARES dataset and the corresponding annotations in detail.

### 2.1 PARES Dataset Description

The dataset contains 250 images of handwritten index tables from about 1670 A.D. to 1862 A.D. They come from two French cities, Vic-sur-Seille (French department of Moselle) and Echevronne (French department of Côte d'Or). While they relate to the distant past, the documents are quite recent as they are handwritten transcriptions of older parish registers, the original index tables written from the period mentioned. Our index tables were copied during the 1960's and 1970' by only a few different writers for two different studies led by *INED* (*Institut National des Études Démographiques* – 'French National Institute for Demographic Studies'), one for each location [15,7,31]. As can be seen in Figure 1, two different and normalized paper templates were used, whose size ranges between A3 and A4 format (ISO 216). These historical tables were previously analyzed for a project managed by INED studying the French population movements before

[3] https://socface.site.ined.fr/ (in French)

1830 [15]. The 250 images we mention in this paper are part of a wider set of 537 images that are fully transcribed. Every token of each line is semantically described (name, surname, father of/mother of, profession...). We intend to release these transcriptions when we will study the handwritten text recognition on these tables.

In 2015, these documents were digitised using two industrial devices. For highly damaged documents, the *Mamiya 645 DF +* digital back captured documents with 60 Mpx and output documents in TIFF format. Digitised documents were rescaled down to 300 DPI and the acquisition tool was regularly recalibrated to maintain color consistency across documents. The other tables that were not damaged were digitised with a *Fujitsu FI 6800* scanner, which outputs 300 DPI images in TIFF format. These digitisations respect the Metamorfoze Preservation Imaging Guidelines [34].

In this publication, we study a subset of this corpus (250 randomly selected images among the 537 images), to identify both research questions and issues for the humanities and social sciences. The high annotation cost led us to annotate 250 images for a first, early release of documents that will be useful for the community. For convenience, we converted the TIFF images to PNG images, a format that compresses images without degrading them.

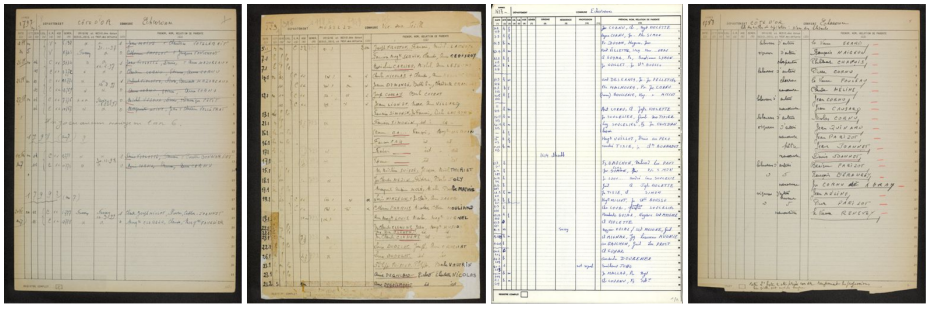


Fig. 1: The different types of documents of the *Historical Index Table Dataset*. From left to right, we will reference these types as 'Category n°1' or 'C1' up to 'C7'. Categories shown are C1, C4, C6 and C7. Details are given in Table 1.

We have categorized the pages into seven distinct groups based on their particularities, which encompass color shifts, layouts, and degradations. An example image for some of these categories is presented in Figure 1, and we report the category statistics in Table 1. C1 and C3 are, for the most part, high quality documents without serious damage, and account for 90% of the dataset. Other document categories include highly damaged pages or pages with unique characteristics.

A notable aspect of this dataset is that the records are written using only two different physical paper templates. Pages in categories 1, 2, 3, 6 and 7 have 25 recordings while those of categories 4 and 5 are larger and can record up to

Table 1: Distribution of the document categories over the 250 images of the dataset. It contains information on the image resolutions (in pixels) and the aspect ratio.

| Category | Description | Count | Image Height | Image Width | Aspect ratio |
|---|---|---|---|---|---|
| | | | Mean (Std) | Mean (Std) | |
| 1 | Clean image with no visible damage | 149 | 4007 (11.29) | 3191 (14.05) | ≈4:5 |
| 2 | Higly damaged documents | 2 | 3980 (1.41) | 3204 (19.09) | |
| 3 | Color-shifted variants of C1 | 74 | 3986 (9.80) | 3199 (11.25) | |
| 4 | Different layout with very light background | 9 | 4956 (8.41) | 3192 (3.00) | ≈2:3 |
| 5 | Yellow-hued variants of C4 | 9 | 4968 (7.62) | 3194 (1.86) | |
| 6 | Table headers differing from those in C1 to C5 | 1 | 3976 (NaN) | 3181 (NaN) | ≈4:5 |
| 7 | Incompliant of pre-defined layout | 6 | 4015 (0.00) | 3197 (0.00) | |

35 items. In Table 1, C4 and C5 images have higher resolutions and a different aspect ratio as compared to the other documents. They represent less than 8% of the dataset, which is hence homogeneous.

### 2.2 Document Layout and Annotations

Layout analysis is an important step prior to document understanding. With this dataset, we wish to retrieve the different components of the image – primarily the text lines which are the handwritten records. The tables have a very clear and organized structure because they are based on very simple templates. We identify four different regions in each document.

- **Page header**: the name of the French department, the city, and the year. Some documents have extra annotations written by a pencil or a black pen. The headers occasionally add context, help to understand the documents, or are used for internal referencing. According to the paper templates, four instances are expected.
- **Page footer**: boolean information (full register or not) and sometimes pencil annotated page numbers. One instance is expected in the paper template document.
- **Table header**: labels on columns indicating the expected content. 19 instances are expected, except for categories n°4 to 6 where there are 23 instances expected.
- **Text line**: the recording itself. From the left boundary of the table to the right, the annotations span across all the columns. Occasionally, corrections were written above or below the main lines within the documents. In such cases, we consider this added information as an independent line unit. Depending on the templates, 25 or 35 text lines are expected.

For each image in the given set, we created segmentation masks in accordance with the aforementioned classes. An example of an annotation is shown in Figure

Table 2: Statistics of the components annotated in the whole dataset of 250 images.

| Component | Count | Mean | Std | Q1 | Median | Q3 | Min | Max |
|---|---|---|---|---|---|---|---|---|
| Page header | 1,451 | 5.80 | 1.57 | 5.00 | 5 | 7 | 3 | 10 |
| Page footer | 475 | 1.90 | 0.54 | 2.00 | 2 | 2 | 1 | 4 |
| Table header | 4,799 | 19.20 | 1.60 | 19.00 | 19 | 19 | 18 | 23 |
| Text line | 5,560 | 22.24 | 8.31 | 18.25 | 25 | 26 | 2 | 63 |

2. The four classes (page header, page footer, table header and text line) are described with four colors and extra page headers & line units are visible. The annotations are bounding polygons as the objects boundaries (such as text lines) are not as clear and precise as they could be in real world – street like – scene images.

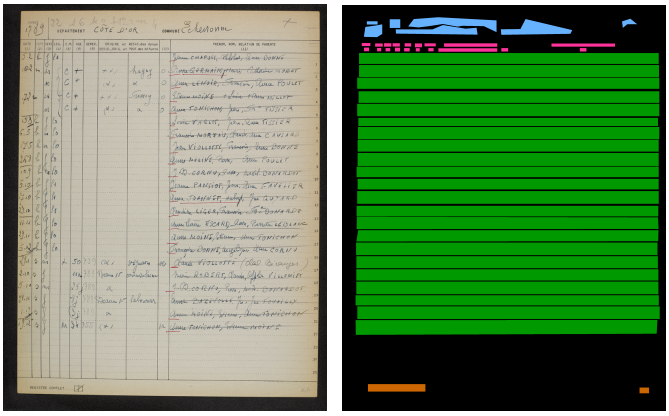
The scientific problem one aims to solve with such a dataset is the automatic detection of the element instances within the given documents, in this case, focusing on text lines. These documents are mainly handwritten, leading to inconsistent use of the template: some text is written out of the cells, overlaps with other text or lines, is crossed out, etc. Our aim is to identify the region of information to perform handwritten text recognition (HTR), not to extract the complex details of the layout of each table, such as columns or cells, with the intention of converting the table into something that looks like a CSV file, for example. We aim at localizing text regions to perform text extraction and named entity recognition before indexing them. This is where the added value resides for archives and digital libraries [33], as it can potentially serve as finding aids for census tables and directly contribute to demographic studies. As a result, the detection of text lines is the baseline task we perform on this dataset.

To annotate the images, we used the open-source Callico[4] platform. The annotations for all the 250 images were exclusively created by a small group, comprising solely the authors of this paper. This approach ensured consistency and adherence to the same annotation rules across the entire dataset. All selected images include each of the four classes just mentioned. In addition to the images, we provide the annotations as segmentation masks, as depicted in Figure 2b. Furthermore, we present various statistics about the annotations in Table 2.

## 3 Document Image Analysis

Document image analysis of digitized documents is generally a two-step process [27,19] although one-step approaches have also been explored [13]. In two-step approaches, the first step involves segmenting the image to extract and classify its components (street signs or cars for street scene images, text lines for document images). In this paper, we focus our attention on the text line detection, elements

[4] https://doc.callico.eu



(a) Original image (b) Segmentation mask

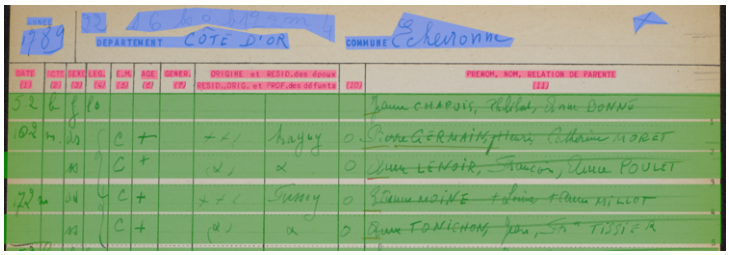


Fig. 2: Example of an image of a document with its annotations (Category n°1). Page header in blue, footer in orange, table header in pink and text lines in green.

that carry information relevant to our second task: the recognition and extraction of handwritten text.

Advances in image segmentation, not limited to document image analysis, mainly come from research based on challenges for natural scene image segmentation [20,10]. As such, the Pascal VOC [14] or Microsoft COCO datasets [23] are mainly used as baseline datasets to enhance image segmentation processes.

The detection of text lines has been widely explored in historical manuscript text books [26,9] and other historical documents of different natures, such as newspapers [25], meteorological tables [1] finding aids [33], as well as many other supports. With index tables, one can consider the issue as a two-class image segmentation task: we separate text lines from the background.

In this paper, we train and fine-tune state-of-the-art algorithms for the purpose of text-line detection. The used neural network models include: Fully Convolutional Networks (FCN), Region-based CNNs (R-CNN) and Transformers.

Fully Convolutional Networks (FCN) output pixel-level probability maps, from which the predicted elements can be extracted. `dhSegment` [28] or more recently `Doc-UFCN` [9] rely on this strategy to detect elements in document images.

For both models, a post-processing task is needed to export masks or polygon coordinates of the detected components.

Other kinds of image segmentation algorithms come from studies that address real-time image segmentation of natural scenes. The most widely used state-of-the-art networks are two-step algorithms ('detect then segment' [16]). These are the `Mask-RCNN` [17] and PANet [24] models, with backbones based on ResNet and Feature Pyramid Networks (FPN) [22] and their derivatives. These models have been evaluated and compared on COCO challenges [16], and are fully integrated in popular toolkits such as Detectron2 or LayoutParser. `Mask-RCNN` has been prior used for document understanding such as on historical newspapers [25]. In contrast, YOLACT and YOLACT++ [10] are single-step approaches focusing on efficiency and increasing the number of frames per second (FPS), a metric that indicates the number of images processed in one second. This enables them to be used in real time applications. When processing historical documents, the FPS metric can aid in estimating the scaling potential of an automated document analysis method when confronted with the task of processing thousands, if not millions, of images.

Recent advances with the transformer architecture [35] have led to improvements within document image analysis fields, including table understanding. The `TableTransformer` [32] is a model for detecting tables and extracting table structure from images and PDF documents. SegFormer [36] and later DocSegTr [8] are other attempts to use the transformer architecture for general document and image segmentation. The latter approach focuses on document segmentation utilizing attention masks to predict instances with segmentation masks, while the former more generally predicts object instances with segmentation. Other transformer-type models like Pix2Seq [11] encode images and output a sequence of bounding box coordinates and the corresponding object classes.

In this paper, we evaluate and compare state-of-the-art image segmentation approaches on the index table dataset to build a baseline for historical handwritten table analysis. We specifically address the challenges posed by tabular data, which is not effectively handled by existing models and presents difficulties for both image segmentation and handwritten text recognition. Moreover, the quality of the latter depends on the efficiency of the document image analysis.

We evaluate the efficiency of the models using the metrics defined in Pascal VOC [14] and COCO [23]. The first is pixel-based and called IoU or Intersection over Union. It is a metric used in segmentation tasks to evaluate the detection of objects in images. Based on IoU thresholds, the mean AP (average precision) and its derivatives ($AP_{50}$, $AP_{75}$) are object based and evaluate the efficiency of object detection [23]. This type of ratio metric helps when comparing the predicted and ground truth segmentations. For example, an IoU of 1 indicates a perfect detection. We also compute the error value for the difference in the number of objects (DiCr – Difference in Count, ratio). This is the mean error based on the number of predicted objects when comparing predictions with corresponding ground truth masks. Zero is the best value, it means there is no difference in count between the prediction and the ground truth. Finally,

we report the inference time in frames per second (FPS). The last two metrics become relevant as the number of detected lines increases, since missing or added predicted lines can lead to huge losses of information and more lines will increase inference time during later HTR processes. Inference time is also a key indicator, as digitization of historical collections can be carried out on massive amounts of data, leading to potentially lengthy and expensive processing.

## 4 Baselines for Text Line Extraction

In this section, we train different models on our tabular data. The aim of this research is to understand how the different models behave in the presence of historical documents, as most relevant models are designed for scene understanding or image analysis of modern documents. As such, text line extraction is often a prerequisite for document understanding and handwritten text recognition. Among the technologies mentioned in Section 3, we run baseline experiments using neural networks from three categories. First, U-shaped Fully Convolutional Networks [30] (U-FCN) that produce a single mask of pixel-level probabilities at the resolution of the input image. Second, Region-based CNNs (R-CNN), such as `Mask-RCNN` [17], detect Regions of Interest (RoI) in images to predict segmentations. Here, each detected instance comes with its own segmentation mask. Finally, recent advances with transformers for computer vision have proved their effectiveness on image segmentation and document understanding [8]. With this, we have chosen to focus our research on the model known as `TableTransformer` [32] which outputs bounding boxes for detected table components.

In the following section, we introduce the PARES dataset splits, pre-processing steps and baseline algorithms, and we comment on the results.

### 4.1 Data Splits and Pre-Processing

Within our dataset, the category distribution is very unbalanced. The dataset is mainly composed of two different categories, C1 and C3, which represent about 90% (respectively 149 and 74 images) of the entire dataset. The models should learn general features to perform well on unseen data. Therefore, we have created a split with 80% of images in `train`, 10% in `validation` and 10% in `test`. We defined the split and the ratios of documents such that we will be able to train models on very unified data, and evaluate their abilities to generalize on unseen documents. The decision to assign an image to a split is still random, we only force proportion distribution. Since C1 and C3 are very similar and without many errors, we use them to train the models. The `validation` set contains C1 and C3 images, to check the learning on these categories, but also images from other categories. In the `test` set, we keep images with the unique peculiarities and damages. Table 3 gives details about the distribution of categories in each set.

In their works, Yang *et al.* [37] and Boillet *et al.* [9] preprocessed the input images by downsizing the images so that the longest size does not exceed respectively 384 and 768 pixels. The quality of the prediction is hence affected

Table 3: Category distribution in the split

| | C1 | C2 | C3 | C4 | C5 | C6 | C7 | Total |
|---|---|---|---|---|---|---|---|---|
| train | 135 | 0 | 65 | 0 | 0 | 0 | 0 | 200 |
| validation | 8 | 0 | 5 | 4 | 5 | 0 | 3 | 25 |
| test | 6 | 2 | 4 | 5 | 4 | 1 | 3 | 25 |

since the network has fewer features to map. We evaluate the impact of different input sizes on the quality of image segmentation. To do so, we downsize the images (Figure 2a) and labels (Figure 2b) to 512 and 768 pixels. The height of all the images in full resolution is around 4000 pixels, except for the documents belonging to the C4 and C5 categories, where the height is around 5000 pixels. This information is reported in Table 1 (page 5). The downsized image height is ≈13% of the original height (for 512 pixels) and ≈19% (for 768 pixels) for all categories except C4 and C5, where the ratios are ≈10% and ≈16%. While the latter resolution (768 pixels in height) was used to match the number of pixels used in `Doc-UFCN`, we chose the former after visual analysis of images smaller than 500 pixels showed large amounts of missing information in the images and labels due to resizing interpolation. Consequently, instead of using the 384 pixel value given by Yang *et al.*, we then used 512 pixels in height for the smaller resolution. We resize the labels and images using `ImageMagick`[5] with a `nearest` pixel interpolation and a `point` filter. To prevent elements from overlapping, when downsizing the segmentation masks, we used an square erosion kernel is of size 2. In the dataset [6], segmentation masks are eroded to prevent overlap, but the JSON files that describe the bounding boxes are not. The implementation of `Doc-UFCN` requires a square image as input. To address this, we introduced zero padding on the sides of the images to transform them into squares after resizing.

### 4.2 Implementations and Experiments

In this paper, we compare `Doc-UFCN`, a U-FCN network, with `Mask-RCNN` and `TableTransformer`, to test and understand how they behave for the task of table document image analysis. `Doc-UFCN` [9] and `TableTransformer` [32] have public implementations and are made for document analysis. We use them directly, as they provide command line tools to train and test the models on custom data. For `Mask-RCNN`, we implemented a model using PyTorch and released it on PyPi[6] as well as on Software Heritage [4].

`Mask-RCNN` and `TableTransformer` rely on a ResNet backbone and pretrained model weights are publicly available. We use these pretrained backbones and models as a starting point for training on our data. For `Doc-UFCN`, no such pretrained backbone models exist, so we trained the model from scratch.

[5] ImageMagick is a command-line image manipulation tool. `https://imagemagick.org/`

[6] `https://pypi.org/project/mask-rcnn-documents/`

Table 4: Baseline experiments on the split of the historical index table dataset. `Doc-UFCN` is trained from scratch. For other architectures, we start from the backbone weights.

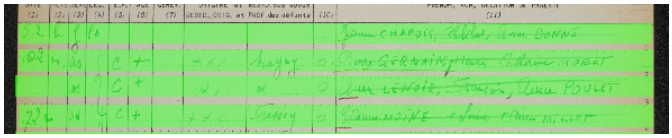
| Model | Parameters | Input size | Pixel metrics | | | | Object metrics | | | DiCr | FPS |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | IoU | P | R | F1 | $AP_{50}$ | $AP_{75}$ | $mAP_{.5,.95}$ | | |
| `Doc-UFCN` | 4.1 M | 512 | 0.54 | 0.75 | 0.62 | 0.67 | 0.36 | 0.14 | 0.18 | 0.85 | 3.28 |
| | | 768 | 0.55 | 0.73 | 0.64 | 0.67 | 0.39 | 0.21 | 0.21 | 1.15 | 1.86 |
| `Mask-RCNN` | 45.9 M | 512 | 0.55 | 0.86 | 0.61 | 0.71 | 0.57 | 0.08 | 0.20 | 0.19 | 1.93 |
| | | 768 | 0.59 | **0.92** | 0.62 | 0.74 | **0.68** | 0.11 | 0.24 | **0.22** | 1.59 |
| `TableTransformer` | 28.8 M | 512 | 0.47 | 0.63 | 0.65 | 0.62 | 0.28 | 0.08 | 0.11 | 1.08 | **8.01** |
| | | 768 | **0.65** | 0.81 | **0.76** | **0.78** | 0.56 | **0.32** | **0.32** | 0.31 | 3.43 |

Table 5: Baseline experiments on the split of the historical index table dataset. Models were pre-trained with the text lines dataset introduced and shared in [9].

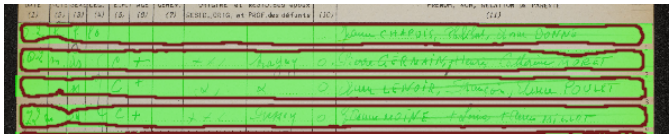
| Model | Parameters | Input size | Pixel metrics | | | | Object metrics | | | DiCr | FPS |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | IoU | P | R | F1 | $AP_{50}$ | $AP_{75}$ | $mAP_{.5,.95}$ | | |
| `Doc-UFCN` | 4.1 M | 512 | **0.66** | **0.91** | 0.71 | **0.79** | **0.65** | 0.28 | **0.34** | 0.21 | 3.35 |
| | | 768 | 0.64 | **0.91** | 0.69 | 0.77 | 0.60 | **0.35** | **0.34** | 0.44 | 1.90 |
| `Mask-RCNN` | 45.9 M | 512 | 0.43 | 0.89 | 0.46 | 0.60 | 0.32 | 0.01 | 0.07 | **0.18** | 2.09 |
| | | 768 | 0.44 | 0.90 | 0.46 | 0.61 | 0.32 | 0.00 | 0.07 | 0.20 | 1.77 |
| `TableTransformer` | 28.8 M | 512 | 0.56 | 0.77 | 0.67 | 0.70 | 0.37 | 0.11 | 0.16 | 0.32 | **7.87** |
| | | 768 | 0.64 | 0.75 | **0.80** | 0.77 | 0.56 | 0.30 | 0.31 | 0.30 | 3.52 |

For training, we propose two configurations. For the first, we fine-tune (`Mask-RCNN`, `TableTransformer`) or train (`Doc-UFCN`) directly to detect table lines from only the historical index table dataset with the split presented in Table 3. Results for this experiment are reported in Table 4. For the second configuration, we first pre-train the models on a large database of document images annotated for the standard text line detection task (around 4,000 images in `train`, 1,300 in `validation` and 2,000 in `test`). They are the datasets used by Boillet *et al.* [9] and listed in the original paper in Table 2. Please refer to this paper to have more information on the datasets used to train the models and the model parameters. Then, we fine-tune these new model weights on the historical index table dataset. Given the limited number of 200 table images in the `train` set, we aim to determine whether fine-tuning with similar data types can enhance the final results. This approach allows the models to encounter a larger number of training examples of text lines, potentially leading to improved performance. The results of this configuration are shown in Table 5.

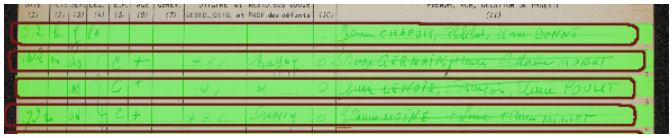
We train several models in different scenarios: with two images sizes (512 and 768 pixels in height) and three types of neural networks (U-FCN, R-CNN and Transformers). Training is done using two `NVIDIA A40` GPUs and inference is done on a computer with an `Intel(R) Core(TM) i7-10850H CPU @ 2.70GHz` and a `NVIDIA Quadro T2000 Mobile` GPU. It is the configuration of a workstation that could likely be used to perform a real document analysis without requiring extra GPUs. The number of epochs during training never exceeded 40



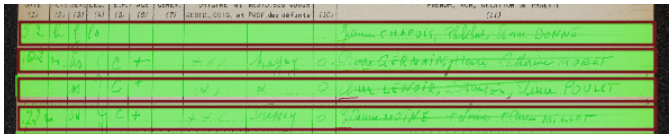
(a) Ground Truth



(b) `Doc-UFCN` with ground truth above it



(c) `Mask-RCNN` with ground truth above it



(d) `TableTransformer` with ground truth above it

Fig. 3: Superposition of manual annotations (in green) and preditions (in red), obtained by the models pre-trained on the large text lines dataset. For `Doc-UFCN` and `Mask-RCNN`, we show the contours of the lines extracted from the segmentation masks, and for `TableTransformer`, the predicted bounding box.

epochs when fine-tuning the models with a batch size of 4, and the training time never exceeded 2 hours.

### 4.3 Results

Depending on the model, the outputs are either segmentation masks or bounding boxes. `Doc-UFCN` outputs a single segmentation mask, as shown in Figure 3b, and `Mask-RCNN` produces multiple masks, one for each detected instance. We merged them into a single mask during post-processing, shown in Figure 3c, to compare the two predictions. The `TableTransformer` model outputs bounding boxes (Figure 3d). We freely share the results of our experiments and the models for further analysis [5].

During inference, we feed the neural networks with images of the same height (512 or 768 pixels) as the one with which the model was trained. This is the resolution of the output. Since we are comparing segmentations at multiple scales, we upscale the predicted masks to the original image resolution before comput-

ing the evaluation metrics. Unlike previous work, we do not evaluate on small sub-resolutions.

We provide two types of results. First, we report in Table 4 the results of models trained from scratch on the index table images. In Table 5, we report the evaluation metrics obtained by the models pre-trained on a standard text line detection task, as described previously. We expect to see improvement in document segmentation as the model would have seen more examples of text lines. For each experiment, we provide both pixel and object metrics. To compute them, we use a free, publicly available software [9]. We keep only $AP_{50}$ and $AP_{75}$ as higher threshold values are not pertinent in our case. The primary information within a text line is typically the text itself, which often gets obscured by background pixels.

Observing Table 4, it becomes apparent that `TableTransformer` outperforms `Doc-UFCN` and `Mask-RCNN` in terms of pixel metrics for 768 pixel images and consistently surpasses `Doc-UFCN` in terms of object metrics. With regard to pixel metrics, `Mask-RCNN` and `Doc-UFCN` behave similarly, with comparable results. Enhancing the image resolution consistently leads to improved results; however, the effect is minimal with `Doc-UFCN`. The same pattern is observed in Table 5. The results on both resolutions are very comparable, indicating that `Doc-UFCN` is able to work well on smaller images. Also, pre-training `Doc-UFCN` using generic text lines proved to be highly beneficial, whereas the impact on other models appears to be minimal. When pre-trained with generic text lines, `Mask-RCNN` performs less well and the improvement on `TableTransformer` is not significant on 768 pixel images; however, it improves a lot on lower resolution images. Alternatively, when trained from scratch, `Mask-RCNN` appears to be more interesting as it performs better than `Doc-UFCN` and `TableTransformer` on small images. Furthermore, it is worth noting that the results obtained from `Mask-RCNN` are consistent across both resolutions. We postulate that this occurrence is attributed to the ResNet-50 backbone based on its deep feature masks which are 224×224 pixels. In contrast, the implementation of `Doc-UFCN` does not exhibit such a limitation, despite yielding comparable results. Nevertheless, it is evident with `TableTransformer` that higher image resolutions yield superior results. We hypothesize that these results are attributed to the attention mechanism in the transformer architecture that is more efficient than the Region of Interest (RoI) detection methods used in `Mask-RCNN` or pixel classification of `Doc-UFCN`. Furthermore, by generating a sequence of bounding box coordinates, the model guarantees the coherence of the coordinates in relation to the page layout. Since `TableTransformer` does not predict at a pixel level, small components (noise) within its predicted bounding box sequences are typically thresholded based on confidence values.

The number of layers and parameters of the models have an impact on the frame per second value. Additionally, reducing the resolution of images boosts the FPS of a given the model due to the reduced number of pixels to process in the input images. Contrary to the general expectation, we observe that `Doc-UFCN`

outperforms `Mask-RCNN`, despite having significantly fewer parameters (10 times fewer) and fewer layers (three times fewer). Since `Doc-UFCN` is able to predict even better segmentation, it is a good candidate for the table line segmentation task for our dataset. `TableTransformer`, even with a high number of parameters compared to `Doc-UFCN`, performs relatively fast and the increase in speed follows the same ratio as `Doc-UFCN` whereas the increase in speed is very low with `Mask-RCNN`, probably due to the hypothesis we already mentioned before.

## 5 Conclusion

Alongside this paper, we share the PARES dataset, a historical dataset of 250 index tables images coming from the French National Archive [6]. Each image in PARES was meticulously annotated for document layout analysis with table headers, text lines, page headers, and footers using bounding polygons. The dataset annotations, presented as segmentation masks, are made available without any overlap between the elements, as eases the extraction of connected components. All identified components are labeled with different colors in segmentation masks to ease the conversion into the formats of your choice: ALTO, PageContent or COCO for instance. The documents are homogeneous (90% are based on a single physical paper template), but we identified variations in each, differences that might fail automatic methods to automatically detect and segment text lines. To that extent, we propose and share a dataset split that focuses on unique document characteristics.

We experimented with three types of neural networks: a U-shaped FCN (`Doc-UFCN`), a region-based CNN (`Mask-RCNN`) and a Transformer (`TableTransformer`) to test their ability to extract table text lines from our historical documents. Their implementations were already public, except for `Mask-RCNN`, which was easily implemented with PyTorch[7]. We release our own version of Software Heritage [4]. The two others are very promising and recent advances with Transformers for document analysis [11,13] lead us to the conclusion that, given sufficient data, the results of these segmentation networks will increase in the domain of historical document analysis. We also foresee the design of a single Transformer-based neural network that could detect and segment documents (by predicting bounding box coordinates and their classes), as well as the handwritten text, all in one.

## Acknowledgements

The authors would like to thank the people of Teklia for providing the infrastructure to annotate document (https://callico.teklia.com). This work was supported by the French government in the framework of the France Relance program and by the TEKLIA company. Full gratitude is expressed to Isabelle Séguy, researcher at INED (Institut national d'études démographiques) and the institution who allowed us to use and share the original images of the PARES dataset.

[7] `https://pytorch.org/vision/main/models/mask_rcnn.html`

## References


1. Andrés, J., Prieto, J.R., Granell, E., Romero, V., Sánchez, J.A., Vidal, E.: Information Extraction from Handwritten Tables in Historical Documents. In: Document Analysis Systems. Lecture Notes in Computer Science, vol. 13237, pp. 184–198. Springer International Publishing (2022). https://doi.org/10.1007/978-3-031-06555-2_13
2. Bernard, G.: doc-ufcn-test (2023), https://archive.softwareheritage.org/swh:1:dir:7ca17e4a36ff25cf4d68513a2af99074a3af4f3f
3. Bernard, G.: doc-ufcn-utilities (2023), https://archive.softwareheritage.org/swh:1:dir:ca5daf53c31def70e46c9aa8f887abe60cdd1d27
4. Bernard, G.: mask-rcnn-documents (2023), https://archive.softwareheritage.org/swh:1:dir:981ec0052f93e37505eba3d47e085a255483441f
5. Bernard, G., Wall, C.: Experiments of 'Line Detection in Historical Index Tables: Evaluations on a New French PArish REcord Survey Dataset (PARES) (09 2023). https://doi.org/10.5281/zenodo.8334664
6. Bernard, G., Wall, C., Boillet, M., Coustaty, M., Kermorvant, C., Doucet, A.: Pares: Parish registry survey – historical census table dataset (19th, 20th centuries) – france (May 2023). https://doi.org/10.5281/zenodo.8337504
7. Biraben, J.N., Brouard, N., Blanchet, D.: Pour reconstituer le mouvement de la population aux xvie et xviie siècles. Annales de Démographie Historique **1980**(1), 39–52 (1980). https://doi.org/10.3406/adh.1980.1452, https://www.persee.fr/doc/adh_0066-2062_1980_num_1980_1_1452, included in a thematic issue : La démographie avant les démographes (1500-1670)
8. Biswas, S., Banerjee, A., Lladós, J., Pal, U.: Docsegtr: An instance-level end-to-end document image segmentation transformer. CoRR **abs/2201.11438** (2022), https://arxiv.org/abs/2201.11438
9. Boillet, M., Kermorvant, C., Paquet, T.: Robust Text Line Detection in Historical Documents: Learning and Evaluation Methods. In: International Journal on Document Analysis and Recognition (IJDAR). vol. 25, pp. 95–114 (2022). https://doi.org/10.1007/s10032-022-00395-7
10. Bolya, D., Zhou, C., Xiao, F., Lee, Y.J.: YOLACT++: Better Real-time Instance Segmentation. In: IEEE Transactions on Pattern Analysis and Machine Intelligence. vol. 44, pp. 1108–1121 (2022). https://doi.org/10.1109/TPAMI.2020.3014297
11. Chen, T., Saxena, S., Li, L., Fleet, D.J., Hinton, G.E.: Pix2seq: A language modeling framework for object detection. CoRR **abs/2109.10852** (2021), https://arxiv.org/abs/2109.10852
12. Constum, T., Kempf, N., Paquet, T., Transuez, P., Chatelain, C., Bree, S., Merveille, F.: Popp datasets : Datasets for handwriting recognition from french population census. https://doi.org/10.5281/zenodo.6581158
13. Coquenet, D., Chatelain, C., Paquet, T.: DAN: A Segmentation-free Document Attention Network for Handwritten Document Recognition. In: IEEE Transactions on Pattern Analysis and Machine Intelligence. pp. 1–17 (2023). https://doi.org/10.1109/TPAMI.2023.3235826
14. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The Pascal Visual Object Classes (VOC) Challenge. In: International Journal of Computer Vision. vol. 88, pp. 303–338 (2010). https://doi.org/10.1007/s11263-009-0275-4

15. Fleury, M., Henry, L.: Pour connaître la population de la france depuis louis xiv. — plan de travaux par sondage. Population **13**(4), 663–686 (1958). https://doi.org/10.2307/1525088, https://www.persee.fr/doc/pop_0032-4663_1958_num_13_4_5737
16. Gu, W., Bai, S., Kong, L.: A Review on 2D Instance Segmentation Based on Deep Neural Networks. In: Image and Vision Computing. vol. 120, p. 104401 (2022). https://doi.org/10.1016/j.imavis.2022.104401
17. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2017)
18. Hervé, D., Eva, L., Florian, K.: READ ABP Table datasets. https://doi.org/10.5281/zenodo.1226879
19. Kiessling, B.: CurT: End-to-End Text Line Detection in Historical Documents with Transformers. In: Frontiers in Handwriting Recognition. pp. 34–48. Lecture Notes in Computer Science, Springer International Publishing (2022). https://doi.org/10.1007/978-3-031-21648-0_3
20. Kirillov, A., He, K., Girshick, R., Rother, C., Dollár, P.: Panoptic Segmentation. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2019)
21. LECTAUREP, Rostaing, A., Durand, M., Chagué, A.: Notaires de Paris - Répertoires, ground truth for various Parisian registries of notary deeds (French 19th and 20th centuries). https://doi.org/10.5072/zenodo.977691
22. Lin, T.Y., Dollar, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature Pyramid Networks for Object Detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 936–944. IEEE (2017). https://doi.org/10.1109/CVPR.2017.106
23. Lin, T.Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C.L., Dollár, P.: Microsoft COCO: Common Objects in Context. In: Computer Vision - ECCV 2014. pp. 740–755. Springer International Publishing (2014)
24. Liu, S., Qi, L., Qin, H., Shi, J., Jia, J.: Path Aggregation Network for Instance Segmentation. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8759–8768 (2018). https://doi.org/10.1109/CVPR.2018.00913
25. Michael, J., Weidemann, M., Laasch, B., Labahn, R.: ICPR 2020 Competition on Text Block Segmentation on a NewsEye Dataset. In: Pattern Recognition. ICPR International Workshops and Challenges. pp. 405–418. Lecture Notes in Computer Science, Springer International Publishing (2021). https://doi.org/10.1007/978-3-030-68793-9_30
26. Neche, C., Belaid, A., Kacem-Echi, A.: Arabic Handwritten Documents Segmentation into Text-Lines and Words Using Deep Learning. In: 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW). pp. 19–24. IEEE (2019). https://doi.org/10.1109/ICDARW.2019.50110
27. Nion, T., Menasri, F., Louradour, J., Sibade, C., Retornaz, T., Metaireau, P.Y., Kermorvant, C.: Handwritten Information Extraction from Historical Census Documents. In: 2013 12th International Conference on Document Analysis and Recognition. pp. 822–826. IEEE (2013). https://doi.org/10.1109/ICDAR.2013.168
28. Oliveira, S.A., Seguin, B., Kaplan, F.: dhSegment: A Generic Deep-Learning Approach for Document Segmentation. In: 2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR). pp. 7–12 (2018). https://doi.org/10.1109/ICFHR-2018.2018.00011
29. PRHLT: HisClima Dataset. https://doi.org/10.5281/zenodo.7442971

30. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015. pp. 234–241. Springer International Publishing (2015)
31. Séguy, I.: La population de la France de 1670 à 1829: l'enquête Louis Henry et ses données. Ined (2001)
32. Smock, B., Pesala, R., Abraham, R.: PubTables-1M: Towards Comprehensive Table Extraction From Unstructured Documents. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4634–4642 (2021)
33. Tarride, S., Boillet, M., Moufflet, J.F., Kermorvant, C.: Simara: a database for key-value information extraction from full pages. arXiv preprint arXiv:2304.13606 (2023)
34. Van Dormolen, H.: Metamorfoze Preservation Imaging Guidelines. National programme for the preservation of paper heritage **5**(1), 162–165 (2012). https://doi.org/10.2352/issn.2168-3204.2008.5.1.art00032
35. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017), https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
36. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. In: Neural Information Processing Systems (NeurIPS) (2021)
37. Yang, X., Yumer, E., Asente, P., Kraley, M., Kifer, D., Giles, C.L.: Learning to Extract Semantic Structure from Documents Using Multimodal Fully Convolutional Neural Network. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4342–4351 (2017). https://doi.org/10.1109/CVPR.2017.462