Pattern Analysis and Applications (2024) 27:22
https://doi.org/10.1007/s10044-024-01235-6

ORIGINAL PAPER

Segmenting large historical notarial manuscripts into multi-page deeds

Jose Ramón Prieto[1] · David Becerra[3] · Alejandro Hector Toselli[1] · Carlos Alonso[2] · Enrique Vidal[1,2]

Received: 19 October 2023 / Accepted: 19 January 2024 / Published online: 28 February 2024

Abstract

Archives around the world hold vast digitized series of historical manuscript books or "bundles" containing, among others, notarial records also known as "deeds" or "acts". One of the first steps to provide metadata which describe the contents of those bundles is to segment them into their individual deeds. Even if deeds are often page-aligned, as in the bundles considered in the present work, this is a time-consuming task, often prohibitive given the huge scale of the manuscript series involved. Unlike traditional Layout Analysis methods for page-level segmentation, our approach goes beyond the realm of a single-page image, providing consistent deed detection results on full bundles. This is achieved in two tightly integrated steps: first, we estimate the class-posterior at the page level for the "initial", "middle", and "final" classes; then we "decode" these posteriors applying a series of sequentiality consistency constraints to obtain a consistent book segmentation. Experiments are presented for four large historical manuscripts, varying the number of "deeds" used for training. Two metrics are introduced to assess the quality of book segmentation, one of them taking into account the loss of information entailed by segmentation errors. The problem formalization, the metrics and the empirical work significantly extend our previous works on this topic.

Keywords Handwritten document image processing · Multi-page layout analysis · Bundle segmentation · Historical manuscripts

✉ Jose Ramón Prieto
joprfon@prhlt.upv.es

David Becerra
davidbegut@gmail.com

Alejandro Hector Toselli
ahector@prhlt.upv.es

Carlos Alonso
carlos.alonso.ono@gmail.com

Enrique Vidal
evidal@prhlt.upv.es

[1] PRHLT Research Center, Universitat Politècnica de València, Camí de Vera s/n, 46021 València, València, Spain

[2] tranSkriptorium IA, València, Spain

[3] Universidad de Sevilla, Sevilla, Spain

1 Introduction

Extensive collections of historical manuscripts holding important notarial documents span across vast lengths of archive shelving globally. Many of these collections have been digitized, transforming them into high-resolution digital images.

Typically, these images are organized sequentially in various archival units like folders, books or boxes, here called "image bundles".[1] Each of these bundles can encompass thousands of individual page images which are sequentially organized into several, often many "image documents", also known as "files", "acts", or, specifically for notarial documents considered in this work, "deeds".

For massive series of documents of this kind, it is often unfeasible for archives to provide detailed metadata that accurately describes the content of each bundle. In particular, if available at all, metadata often goes without information about the specific location of individual deeds within a bundle. Automated solutions are needed to assist archival specialists in the arduous task of cataloging these expansive series. Segmenting bundles into distinct deeds stands as one of the preliminary phases in this operation. This is the primary focus of our current study.

Historically, efforts to address analogous challenges originated in the field of Layout Analysis (LA). This domain encompasses many document analysis tasks, from line detection and page layout segmentation to document understanding.

Numerous line detection and extraction techniques, a basic step for Handwritten Text Recognition (HTR) systems, lean on text baseline detection methods [6, 25] that employ diverse strategies. Some recent studies utilize convolutional networks [4] or employ encoder–decoder designs to concurrently achieve layout segmentation and line detection [3, 18, 27]. Additionally, some incorporate a spatial attention mechanism across various resolutions to existing architecture [9].

Region Proposal Networks (RPN) are also used for single-page LA. Tools like MaskRCNN [12] have proven effective for complex layout segmentations, as demonstrated in Refs. [2, 25]. Notably, in Ref. [1], the same methodology addresses challenges posed by having lines close together, facilitating data extraction from tables.

Furthermore, the LA domain has achieved considerable advancements by implementing transformer-based designs, which occasionally integrate HTR. For instance, the DONUT approach introduced in Ref. [14] offers document understanding by analyzing entire pages in an end-to-end fashion. Another solution, DAN [7], introduces a segmentation-free model that determines the logical layout and simultaneously recognizes textual content without necessitating geometrical data from LA.

The interest in these kinds of holistic approaches notwithstanding, none of these works consider text-image processing beyond the realm of a single-page image, which is the problem we face here.

Automatic classification techniques have been developed for image documents (deeds) to classify their specific typological categories, like "Letter of Payment" or "Will", with encouraging results documented in Refs. [8, 21, 23, 30]. However, these studies assume that successive page images for each deed are assumedly given. Contrarily, real-world scenarios often present deeds nestled within bundles without explicitly separating the distinct page images each deed comprises.

Therefore, to provide practical solutions for automated documentary management of these important series of manuscripts, a pending problem is how to segment a large bundle into its constituent deeds. This is the task considered in the present work. Previous approaches to this problem exist, but they should be considered only somewhat tentative. In studies like Refs. [5, 6] and [20], the authors introduce text-line-oriented, page-level segmentation techniques using Hidden Markov Models (HMMs) and recurrent CTC-based systems to identify the deeds' starting, middle, and ending sections within each page image. Further, Tarride et al. [29], outline a complete workflow for data extraction from registers. However, it is important to note that these studies do not provide results on segmenting entire bundles. Additionally, they lack methodologies to ensure consistent detection throughout a whole bundle.

In this paper, we propose a pipeline based on a simple neural network to obtain the posterior probabilities of each image over a closed set of classes: initial, middle, and final. Then, relying on the sequential order of the book's images, we apply a decoding approach based on the global context on these image posteriors to segment the entire book. The book segmentation is evaluated in terms of two different metrics which clearly show that far better results are obtained when considering the global book context rather than local image classification. One of the main advantages of this pipeline is that it does not require any kind of textual transcripts of the images for training. The cost of producing reference transcripts would be overly prohibitive for this task in practice. Unlike other approaches such as [7, 14], another advantage is that our method does not require large amounts of training data or the creation of any kind of synthetic data. It is also worth mentioning that this pipeline is hardware lightweight and can be trained with low memory GPU (6GB or less).

[1] The protocols used in this work were not really "bundles". However, once they are converted into digitized images, we no longer have anything like books, bundles, boxes, etc., but electronic files and folders. Since there is no commonly used term in archival science to refer to this kind of organized sets of page images, we have taken the liberty of using the term "image bundle".

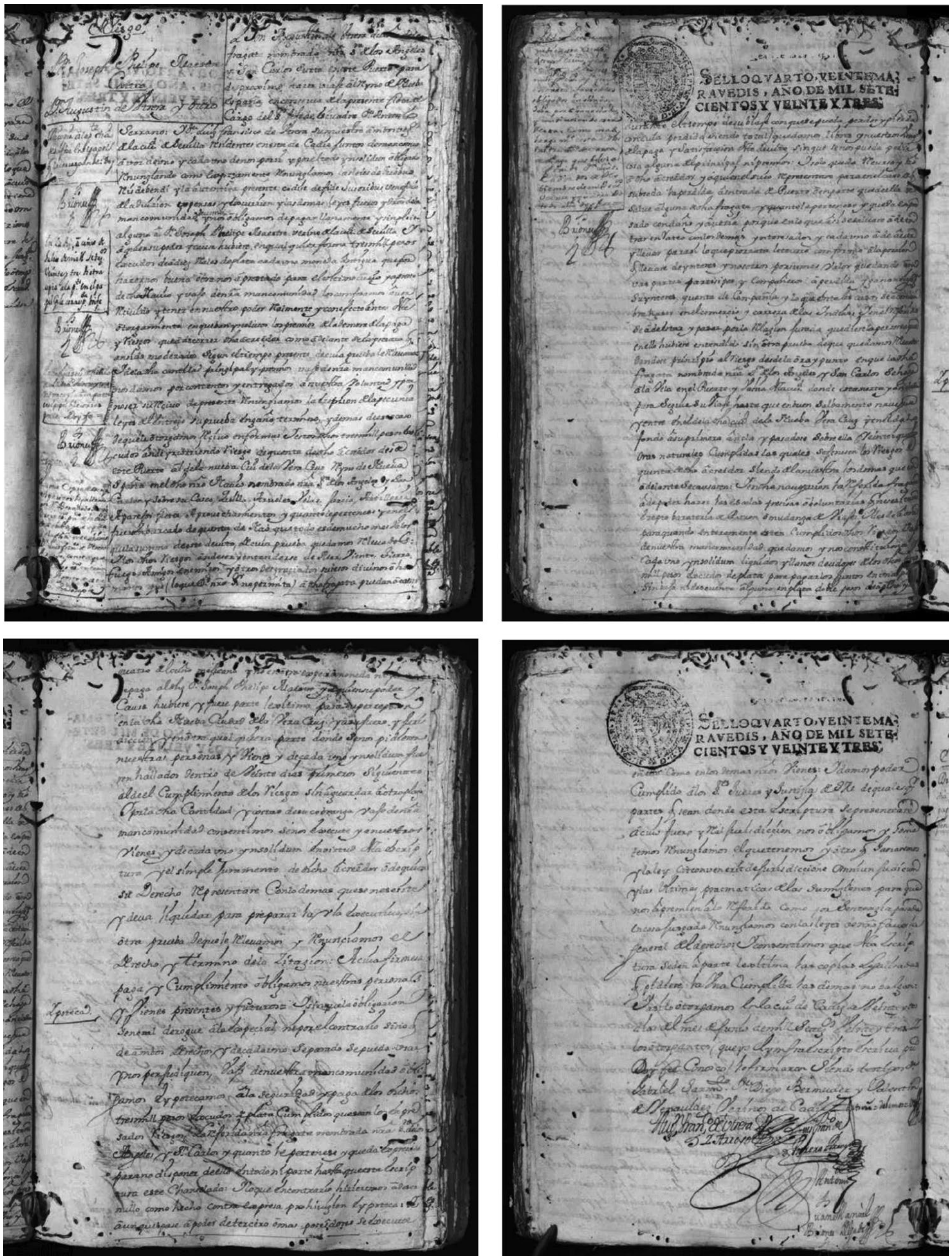
SELLO QVARTO, VEINTE MARAVEDIS, AÑO DE MIL SETECIENTOS Y VEINTE Y TRES
SELLO QVARTO, VEINTE MARAVEDIS, AÑO DE MIL SETECIENTOS Y VEINTE Y TRES

Fig. 1 A four-page deed from the JMBB-4949 bundle of the AHPC

This work extends the research started in Ref. [22]. Here, a detailed formalization of the problem and new ways of evaluating the results are proposed. Furthermore, we present more comprehensive experiments, with a larger and more complete dataset.

The presentation of our work is organized as follows. Section 2 describes the historical documents considered. Section 3 explains in detail the technical problem entailed by bundle segmentation and the solutions we propose. Section 4 introduces the metric used to assess bundle segmentation results. Section 5 provides dataset and empirical setting details. Section 6 presentes the experiments carried out and the results achieved. And Sect. 7 summarizes the work carried out and comes out with final remarks.

2 The JMDB series of notarial record manuscripts

For this study, we have worked on a portion of the 16,849 "notarial protocol books" which are more than three hundred years old and are preserved in the Provincial Historical Archive of Cadiz (AHPC, by its acronym in Spanish). Each protocol book (or "bundle") contains more than eight hundred pages, organized into two hundred and fifty deeds on average.

The portion considered in this work is a series of four bundles referred to as JMDB, which contain deeds written by the notary Jose Manuel Briones Delgado between 1712 and 1726. The deeds of each bundle are arranged chronologically, preceded by an onomastic and topographic index (about fifty pages) which has not been used in the experiments presented in this paper. Figure 1 shows a typical JMDB deed.[2]

For the JMBD series, the deeds are *page-aligned*. Each deed always begins on a new recto page and can contain from one to dozens of pages, some of which may be almost blank or without any textual content. The first and last pages of each deed are often visually identifiable because of slight layout differences compared to other pages. However, separating the deeds of each bundle is not straightforward and remains a challenging task. The main difficulty is the similarity many regular pages share with the beginning or ending pages. Various "data-centric" and rule-based techniques have been attempted, but none of them have worked sufficiently well. Additionally, another feature of this(and most other) collection(s) is the lack of textual transcripts. The deed in Fig. 1 showcases the starting, ending, and two intermediary page images.

While these details might appear very specific to these manuscripts, it is important to highlight this series's sheer magnitude, which deserves the development of ad-hoc methods. Furthermore, innumerable notarial record series exhibit a similar bundle structuring, especially where each deed begins on a fresh page. However, it is pertinent to point out that other notarial record series exist where the deeds are *not* aligned at the page-level. Among them, we can mention Chancery (HIMANIS) [20], Oficio de Hipotecas de Girona (OHG) [26, 28], and The Cartulary of the Seigneury of Nesle (Nesle) [13].[3] Though these particular manuscript types are not the focus of our current study, some of the concepts and methods introduced here might also be useful when addressing challenges associated with these kinds of bundles.

3 Problem statement and proposed approaches

The bundle segmentation problem is formalized in this section, along with the approaches we propose, based on individual page-image classification and whole-bundle consistence modeling.

3.1 Problem formalization

Let $B = D^1, \dots, D^K$ be a bundle which sequentially encompasses K deeds.[4] Each deed D^k, in turn, is a sequence of $M(k) \geq 2$ page images, denoted as $D^k = G^k_1, \dots, G^k_{M(k)}$. In the deed segmentation task, a bundle B is given just as a plain, unsegmented sequence of N page images $B = G_1, \dots, G_N$ and the problem is to find $K+1$ boundaries b_k, $0 \leq k \leq K$, such that $b_0 = 0$, $b_{k-1} < b_k$, $b_K = N$, and B becomes described as a sequence of deeds $D^1, \dots, D^K$, where $D^k = G_{b_{k-1}+1}, \dots, G_{b_k}$ and $M(k) = b_k - b_{k-1}, 1 \leq k \leq K$.

As discussed in Sect. 2, all the deeds within a bundle start on a new page and also end with a full page, even if part of the page is unused. Initial and final deed pages will be referred to as "Initial" (I) and "Final" (F), respectively. All the pages within I and F will be referred to as "Mid" (M). Let $\mathcal{C} \overset{\text{def}}{=} \{\mathsf{I}, \mathsf{M}, \mathsf{F}\}$ be de set of page classes[5] or labels. A sequence $c_1, \dots, c_N,\ c_j \in \mathcal{C},\ 1 \leq j \leq N$ of page labels that follows these rules is said to be *consistent*.

Given a label sequence $c_1, \dots, c_N$, the corresponding segmentation is readily obtained by successively setting the boundaries b_k to the positions of the pages labeled with F, as outlined by the following pseudocode:

$$\begin{aligned} &b_0 := k := 0; \\ &\text{for } (j := 1 \dots N) \ \text{ if } (c_j = \mathsf{F})\{k := k+1; b_k := j\};\ K := k \end{aligned} \tag{1}$$

3.2 Proposed approaches

Three increasingly complex solutions are proposed. The common idea is to model the likelihood of a (consistent) page label sequence, and then try to obtain as a result a sequence with maximum likelihood. In general, the probability of a label sequence $c_1, \dots, c_N$ for a given bundle $B = G_1, \dots, G_N$ can be decomposed as:

[2] The entire JMBD collection, among others from AHPC, can be seen in the first folder of the following demonstrator: https://www.prhlt.upv.es/htr/carabela/

[3] https://deeds.library.utoronto.ca/cartularies/0249

[4] To avoid cumbersome equations, we will abuse the notation and index the elements of a sequence of sequences with a plain, rather than parenthesized superindex. That is, we will write D^k, rather than $D^{(k)}$.

[5] A deed may also contain blank or otherwise useless ("nonessential") pages, which may appear either at the end of a deed, or within it. Nonessential pages are almost trivial to detect and, for the sake of simplicity, we assume they are eliminated from the bundle in a simple previous step.

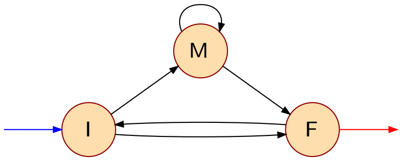

Fig. 2 Topology of the Consistency Constraints HMM. For convenience, states are labeled with I, M, and F, respectively corresponding to the initial, middle, and final pages of a deed

$$P(c_1,\dots,c_N \mid B) = P(c_1 \mid B)\prod_{j=2}^{N} P(c_j \mid B, c_1,\dots,c_{j-1}) \tag{2}$$

Assuming naive Bayes page class independence and also that $P(c_j \mid B)$ only depends on the corresponding image G_j:

$$P(c_1,\dots,c_N \mid B) \approx \prod_{j=1}^{N} P(c_j \mid G_j) \tag{3}$$

Now, rather than naive Bayes, a more context-aware decomposition of the right part of Eq. (2) can be adopted (a derivation is given in "Appendix A"):

$$P(c_1,\dots,c_N \mid B) \;\approx\; P(c_1 \mid G_1)\prod_{j=2}^{N} P(c_j \mid c_{j-1})\frac{P(c_j \mid G_j)}{P(c_j)} \tag{4}$$

To obtain this expression, two independence assumptions have been made: page class dependency is first-order Markov; and G_j is conditionaly dependent only on c_j and conversely c_j is conditionally dependent only on G_j. This decomposition essentially corresponds to that of a Hidden Markov Model (HMM), as discussed below.

Our proposals to segment B into deads will amount to obtaining a page class sequence with maximum probability; that is:

$$\hat{c}_1,\dots,\hat{c}_N \;=\; \arg\max_{c_1,\dots,c_N} P(c_1,\dots,c_N \mid B) \tag{5}$$

where $P(c_1,\dots,c_N \mid B)$ is approximated by Eqs. (3) or (4). This optimization process is called *decoding*.

3.2.1 Optical page class modeling

A classifier is required to estimate the class-posteriors $P(c \mid G)$, $c\in\mathcal{C}$ needed in Eqs. (3,4) for each page image G of a bundle. This classifier is fully local in that it completely ignores the context of G (i.e., the preceding and succeding pages) and relies only on "visual" or "optical" features of each individual page image. Possible optical classifiers to model these probabilities are proposed in Sect. 3.3, below.

Given a bundle $B = G_1,\dots,G_N$, the page classifier is used to estimate a sequence of class-posteriors $P(c \mid G_j)$, which will be referred to as the *posteriorgram*[6] of B:

$$\mathbf{g}_1,\dots,\mathbf{g}_N, \quad g_{jc} \overset{\text{def}}{=} P(c \mid G_j), \quad c\in\mathcal{C}, \quad 1\le j\le N \tag{6}$$

3.2.2 Consistency constraints model

The probability decomposition of Eq. (4) is that of a first-order HMM with a set of states $\mathcal{Q} = \mathcal{C} = \{\mathsf{I},\mathsf{M},\mathsf{F}\}$ and state transition probabilities $P(c \mid c')$, $c, c' \in \mathcal{Q}$. The state-emission probabilities would be the class-conditional likelihoods $P(G \mid c)$, where G is a page image and $c \in \mathcal{C}$. These likelihoods are proportional to $P(c \mid G)/P(c)$, as used in Eq. (4), where the posterior $P(c \mid G)$ is given by the image classifier (Eq. 6), and $P(c)$ is trivially estimated from Ground Truth (GT) segmented bundles.

Note that the ultimate goal of segmentation is to preserve the coherence of the textual information of each deed of a bundle. To this end, each deed segment $D^k, 1\le k\le K$, must fullfil the following *hard Consistency Constraints* (CC): $G_{b_{k-1}+1}$ is an I-page, G_{b_k} is an F-page and, if $M(K)>2$, $G_{b_{k-1}+2},\dots,G_{b_k-1}$ are all M-pages.

Correspondingly, only the state F can be final and the initial-state probability must be 1 for the state I and 0 for other states. In addition, $P(\mathsf{I}\,|\,\mathsf{M})=P(\mathsf{M}\,|\,\mathsf{F})=P(\mathtt{I}\,|\,\mathtt{I})=P(\mathsf{F}\,|\,\mathsf{F})=0$. The other five transition probabilities can be straightforwardly estimated from GT segmented bundles. This HMM topology is depicted in Fig. 2.

3.2.3 Unconstrained decoding

Using Eq. (3), the optimization (5) becomes trivial (but consistency is not guaranteed):

$$\hat{c}_j \;=\; \arg\max_{c\in\mathcal{C}} g_{jc}, \qquad 1\le j\le N \tag{7}$$

3.2.4 Greedy decoding

Before attempting to obtain a globally optimal solution to Eq.(5), a much simpler *greedy decoder* can be devised by locally applying the CCs as follows:

$$\begin{aligned} &\hat{c}_1 = \mathtt{I};\quad \hat{c}_j \;=\; \arg\max_{c\in\rho(\hat{c}_{j-1})} g_{jc},\ 2\le j\le N-2;\\ &\quad \hat{c}_{N-1} = \pi(\hat{c}_{N-2});\quad \hat{c}_N = \mathsf{F} \end{aligned} \tag{8}$$

[6] Following time-honored tradition in signal processing and automatic speech recognition, the term *posteriorgram* is used for this type of (variable-length) sequences of posterior probability vectors.

where the function $\rho : \mathcal{C} \rightarrow 2^{\mathcal{C}}$ is defined as: $\rho(\mathsf{I}) = \rho(\mathsf{M}) = \{\mathsf{M}, \mathsf{F}\}$; $\rho(\mathsf{F}) = \{\mathtt{I}\}$, and $\pi : \mathcal{C} \rightarrow \mathcal{C}$ is a "previous label function" defined as: $\pi(\mathtt{I}) = \pi(\mathsf{M}) = \mathsf{M}$; $\pi(F) = \mathtt{I}$.

By construction, this greedy decoder yields a consistent label sequence. However, it does not guarantee that the probability of this sequence is maximum.

3.2.5 Viterbi decoding

Following Eqs. (4) and (6) the optimization (5) becomes:

$$\hat{c}_1, \dots, \hat{c}_N = \arg\max_{c_1,\dots,c_N} g_{1c_1} \prod_{j=2}^{N} \tilde{g}_{jc_j} P(c_j \mid c_{j-1}) \tag{9}$$

where $\tilde{g}_{jc_j} \overset{\text{def}}{=} g_{jc_j}/P(c_j),\ g_{jc_j}, 1 \leq j \leq N, c_j \in \mathcal{C}$, are the components of the bundle posteriorgram, and $P(c_j)$ are the page class priors.

To achieve a globally optimal solution to this equation, a *Dynamic Programming* decoder is needed. The Viterbi algorithm, which is one of the most popular of this kind of decoders, can be outlined as follows.

Let $V(j, q)$ denote the probability of a max-probability state sequence which ends in state q and generates the first j labels. Set $V(1, \mathtt{I}) = g_{1,\mathtt{I}},\ V(1, \mathsf{M}) = V(1, F) = 0$. Then the following recurrence relation holds for $2 \leq j \leq N$:

$$V(j, q) = \max_{q' \in \mathcal{Q}} \tilde{g}_{jq}\, P(q \mid q')\, V(j-1, q'), \quad q \in \mathcal{Q} \tag{10}$$

Once $V(N, \mathsf{F})$ is computed, backtracing yields a globally optimal consistent sequence of states and the corresponding sequence of I,M,F labels.

3.3 Individual page image classification

To compute the posteriorgram introduced in Eq. (6), a page-image classifier is needed to obtain the class posterior $P(c \mid G)$ for each page image G and each $c \in \{\mathtt{I}, \mathsf{M}, \mathsf{F}\}$.

In the present work, several classifiers have been tried; namely, ResNet-{18,50,101} [11], ConvNeXt [16], and other transformer-based image classifiers, such as Swin [15]. All of them have been pre-trained on ImageNet from [32].

ResNet and ConvNeXt are both convolutional neural networks with a last, linear classification layer. ResNet is a commonly used architecture for image classification tasks. It consists of multiple convolutional layers with residual connections between them. Residual connections enable the network to train much deeper architectures by mitigating the vanishing gradient problem. The specific architecture is determined by the number of ResNet blocks. The more blocks we set, the larger and deeper the network, and the greater number of parameters to be trained.

ConvNeXt, as the authors explain in Ref. [16], is a "modernized" ResNet with the latest advances from training ConvNets and Vision Transformers. In the present work, only ConvNeXt base has been used. ConvNeXt tiny was tried too, but the results were worse.

We also tried to train Vision Transformers, such as Swin [15], but the model had convergence difficulties and the results were rather poor. A possible cause is the difference in the resolution of images from the pre-training stage, where the original images were of 224 × 224 pixels, while we need to use an image size of 1024 × 1024. A larger size is required because, in our classification task, the clues are usually words-sized visual features or relatively small boxes in specific parts of the image and a higher resolution is needed to avoid losing this information.

4 Evaluation measures

In the approaches described in Sect. 3, several alternatives are proposed for the different components needed to build a complete bundle segmentation system.

To select the best system components, we first need to assess the performance of the different individual page image classifiers (Sect. 3.3), and then evaluate, end-to-end, the segmentation performance of the alternative decoders proposed in Sect. 3.2.

4.1 Assessing individual page classifiers

The performance of the page image classifiers discussed in Sect. 3.3 could be assessed just using the conventional classification error rate. However, the impact of these classifiers on the end-to-end performance of the different methods discussed in Sect. 3.2 is more dependent on the faithfulness of the class probability distribution than on their ability to make the best class choice without taking the context into account.

Therefore, we are interested in well-calibrated class probabilities, not just low classification errors. Even though incontextual classification by max class posterior fails to yield correct class hypotheses, decoding can achieve flawless segmentation if the probabilities of the correct classes are not too low.

The faithfulness of a posterior distribution can be measured by the *cross-entropy*:

$$\begin{aligned} H(P_t, P) &= -\frac{1}{N} \sum_{j=1}^{N} \sum_{c \in C} P_t(c \mid G_j) \log_2 \\ P(c \mid G_j) &= -\frac{1}{N} \sum_{j=1}^{N} \log_2 g_{j,c_j} \end{aligned} \tag{11}$$

where P_t is the "target distribution",[7] N is the total number of samples (pages) and g_{j,c_j} is the class-posterior of the image G_j for the reference class c_j, as defined in Sect. 3.2. It will be used to evaluate and compare the page-image classifiers between the probability distributions of the hypotheses and the reference distributions. This way, we will assess the uncertainty between both probability distributions, selecting the model with the lowest cross-entropy for later decoding and segmenting the books. It should be noted that cross-entropy is rather preferable to classification error, since our goal is to measure the goodness of the image posterior probabilities provided by the model to segment properly the image sequence.

4.2 Assessing bundle segmentation performance end-to-end

Most importantly, to assess the ultimate goal of the proposed approaches, we need to measure how well a whole bundle is segmented into its deeds. To this end, we propose two metrics called Bundle Segmentation Error Rate (BSER) and Content Alignment Error Rate (CAER).[8] BSER aims to measure structural deed errors due to wrong segmentation boundaries, without considering the (textual) content of the deeds. On the other hand, CAER explicitly aims to measure the amount of textual information lost because of deed segmentation errors.

In both cases, the same Dynamic Programming (DP) algorithm is used to align reference and hypothesis deeds. However, the individual deed alignment cost is different for BSER and CAER. For the moment, let us just assume this cost is given by a function $L : \mathcal{D} \times \mathcal{D} \to \mathbb{R}^{\geq 0}$, where $\mathcal{D}$ is the set of deeds which includes the *"empty deed"* (a deed with no pages or words), denoted by ϵ.

Let $\hat{B} = \hat{D}^1, \dots, \hat{D}^{\hat{K}}$ be a sequence of $\hat{K}$ deeds, obtained as a hypothesis for a whole bundle as explained in Sec. 3, and let $B = D^1, \dots, D^K$ be the corresponding reference GT. Using these sequences, the minimum *deed editing cost* to transform B into $\hat{B}$ is $\mathrm{E}(K, \hat{K})$, which can be computed by DP according to the following recurrence relation:

$$\begin{aligned} \mathrm{E}(i,j) = \min \big(& \mathrm{E}(i,j-1) + \mathrm{L}(\epsilon, \hat{D}^j), \\ & \mathrm{E}(i-1,j-1) + \mathrm{L}(D^i, \hat{D}^j), \\ & \mathrm{E}(i-1,j) + \mathrm{L}(D^i, \epsilon) \end{aligned} \tag{12}$$

The decisions associated to the minimal cost $\mathrm{E}(K, \hat{K})$ are interpreted as *deed edit operations*; namely, deed insertions, substitutions and deletions, respectively corresponding to the three terms of the min function. An *insertion* indicates that a deed that was not in the reference does appear in the hypothesis. Conversely, a *deletion* means that a deed of the reference has disappeared in the hypothesis. A *substitution*, finally, corresponds to matching a pair of reference and hypothesis deeds.

4.2.1 Bundle segmentation error rate

A deed here is considered just as a *set* of page images. Therefore, a direct way to measure how much two deeds differ is just to compute the difference between the sets of pages of the deeds.

Let $D, \hat{D} \in \mathcal{D}$ denote the sets of page images of a pair of reference and hypothesis deeds of some bundle. The individual alignment cost for these two deeds is defined as the *symetric set difference* $L(D, \hat{D}) \overset{\text{def}}{=} |D \ominus \hat{D}|$, which can be also computed as[9]

$$L(D, \hat{D}) = |D \cup \hat{D}| - |D \cap \hat{D}| \tag{13}$$

According to this definition, the cost of a deed insertion, $L(\epsilon, \hat{D}) = |\hat{D}|$, is the number of page images in $\hat{D}$. Similarly for deletion, where the cost $L(D, \epsilon)$ is the number page images in D. And the cost of a substitution $L(D, \hat{D})$ is just the number of page images which are in D but not in $\hat{D}$ and those page images that are in $\hat{D}$ but not in D.

Finally, for the reference deed sequence of a bundle $B = D^1, \dots, D^K$ and the corresponding segmentation hypothesis $\hat{B} = \hat{D}^1, \dots, \hat{D}^{\hat{K}}$, the BSER is defined as:

$$\mathrm{BSER}(B, \hat{B}) = \frac{1}{T}\, \mathrm{E}(K, \hat{K}) \tag{14}$$

where $\mathrm{E}(K, \hat{K})$ is computed using Eq. (12) with the cost function given by Eq. (13) and $T \overset{\text{def}}{=} \sum_{i=0}^{K} |D^i|$ is the total number of page images in B.

4.2.2 Content alignment error rate

To define a metric that measures the information loss caused by segmentation errors, some issues must be taken into account. First, not solely the quality of structural segmentation should impact the evaluation metrics; also the deed's textual content should be taken into account. However, recognition errors in nonessential pages or in pages containing minimal or low informational text should exert only a minor penalty on the overall bundle segmentation performance.

[7] defined as $P_t(c \mid G_j) = 1$ iff G_j is of class c, according to GT.

[8] In previous publications this metric was called "Bundle Alignment Error Rate" (BAER).

[9] Page image sets can be represented just as sets of page ID integers. Moreover, if these integers are sequentially assigned to the page images of a bundle, as is typically our case, $L(D^i, \hat{D}^j)$ can be trivially computed using the segmentation *boundaries*: of D^i and $\hat{D}^j$, as: $b_i - b_{i-1} + \hat{b}_j - \hat{b}_{j-1} - 2\max(0, \min(b_i, \hat{b}_j) - \max(b_{i-1}, \hat{b}_{j-1}))$ (cf. Sect. 3.1).

Table 1 Number of page images and deeds for the bundles JMBD4946, JMBD4949, JMBD4950 and JMBD4952

	JMBD4946	JMBD4949	JMBD4950	JMBD4952
Number of pages	1399	1615	1481	980
Number of deeds	248	295	260	236
Average pages per deed	5.9	5.5	5.7	4.2
Min–max pages per deed	2–200	2–122	2–62	2–38
St-dev of pages per deed	14.3	9.9	8.2	4.1

Taking these issues into account and following [8, 20], we rely on Probabilistic Indexing (PrIx) [24] and Information Gain (IG) to compute a feature vector $\mathbf{D} \in \mathbb{R}^n$ for a reference deed and similarly $\hat{\mathbf{D}} \in \mathbb{R}^n$ for a deed hypothesis. The n components of these vectors correspond to the n words with higher IG, as determined in Refs. [20, 23], and the value of each component is the expected number of occurrences of the corresponding word, estimated from the image PrIx as also discussed in Refs. [20, 23]. In addition, an empty deed ϵ is represented by the null vector $\mathbf{0}$. Such a document characterization provides a compact representation of the information provided by the most relevant textual contents of the page images that constitute a deed.

This way, the sequence of deeds obtained by automatic bundle segmentation becomes represented as a sequence of vectors $\hat{\mathbf{D}}^1, \ldots, \hat{\mathbf{D}}^{\hat{K}}$ and, similarly, the corresponding GT reference as $\mathbf{D}^1, \ldots, \mathbf{D}^K$.

As explained in Ref. [23], the IG calculation depends on the documents' classes. Since our interest lies in ranking words to select the top n, we have used the same order as in Ref. [23], where 12 deed typologies were adopted as classes. And also in accordance with the results of Prieto et al. [23], we have set the number of high IG words at $n = 16\,384$.

If $D, \hat{D} \in \mathcal{D}$ is a pair of reference and hypothesis deeds, $L(D, \hat{D})$ is defined in similar way as the "Bag of Words Error Rate" (bWER) in Ref. [31], but without normalization. It estimates the number of *word* insertion, substitution and deletion edit operations to transform the text in D into the text in $\hat{D}$, but *ignoring* the *order of words* in D or in $\hat{D}$. Specifically:

$$L(D, \hat{D}) \overset{\text{def}}{=} \frac{1}{2}\Big(\|\mathbf{D} - \hat{\mathbf{D}}\|_1 + \big| \|\mathbf{D}\|_1 - \|\hat{\mathbf{D}}\|_1 \big| \Big) \tag{15}$$

According to this definition, the cost of a deed insertion, $L(\epsilon, \hat{D}) = \|\hat{\mathbf{D}}\|_1$, is the number of (high IG) running words in $\hat{D}$. Similarly, the cost of a deletion $L(D, \epsilon)$ is the number of (high IG) running words in D. And, in the case of a substitution, $L(D, \hat{D})$ is the unnormalized bWER calculated with the two aligned deeds.

Finally, for the reference deed sequence of a bundle $B = D^1, \ldots, D^K$ and the corresponding segmentation hypothesis $\hat{B} = \hat{D}^1, \ldots, \hat{D}^{\hat{K}}$, the CAER is defined as:

$$\mathrm{CAER}(B, \hat{B}) = \frac{1}{T}\,\mathrm{E}(K, \hat{K}) \tag{16}$$

where $\mathrm{E}(K, \hat{K})$ is computed using Eq. (12) with the cost function given by Eq. (15) and $T \overset{\text{def}}{=} \sum_{j=0}^{K} \|\mathbf{D}^j\|_1$ is the total number of (high IG) running words in B.

5 Data set and experimental setup

The data set statistics are shown in this section, along with empirical details needed to make the experiments reproducible.

5.1 Data set

Among the AHPC Series of Notarial Record Manuscripts described in Sect. 2, 50 bundles were included in the collection compiled in the Carabela project [30].[10, 11] From these bundles, in the present work we selected four: JMBD4946, JMBD4949, JMBD4950 and JMBD4952 to be manually tagged with segmentation GT annotations.

Table 1 shows statistics for this dataset. Each bundle has more than, or close to one thousand pages. The number of deeds is also shown, with more than two hundred deeds per bundle, and almost three hundred in JMBD4949. An interesting fact is the large variation in the number of pages per deed, with a standard deviation of more than 14 pages per deed in JMBD4946. In fact, one of these deeds spans up to 200 pages. This large deed size variability poses a challenge for automated segmentation, since methods which may rely solely on page count or structure are not sufficiently effective. Additional strategies, such as the methods presented in the present work, need to be employed to accurately identify and separate individual deeds within a bundle.

[10] The GT data is available in https://zenodo.org/records/10418179. The images are available under request from PARES (Portal de Archivos Españoles)

[11] Inhttps://www.prhlt.upv.es/htr/carabela/ the images of this collection and a search interface based on Probabilistic Indexing are available.

Table 2 Cross-entropy (bits/page) between the hypothesis page-image class posterior and the reference (0/1) probability distributions. Training with one bundle and testing with the other three bundles

Classifier	JMBD4946	JMBD4949	JMBD4950	JMBD4952	Average
ResNet18	0.028	0.027	0.031	0.116	0.050
ResNet50	0.013	0.009	0.014	0.022	0.015
ResNet101	0.043	0.016	0.028	0.065	0.038
ConvNeXt	0.017	0.009	0.022	0.027	0.019

Table 3 BSER and CAER achieved by different decoders, training with one bundle and testing with the other three bundles. The page image classifier was ResNet50. Results are in percentage. The highest standard deviation of fluctuations of the Viterbi decoding results (BSER and CAER) obtained for the 10 trials is 4.55%, while 9.31% for the other decoding results

Metric	Decoder	JMBD4946	JMBD4949	JMBD4950	JMBD4952	Average
BSER	Unconstrained	23.3	24.5	23.3	37.1	27.0
	Greedy	22.7	24.5	23.1	36.2	26.2
	Viterbi	5.6	9.0	5.7	13.4	8.4
CAER	Unconstrained	19.1	20.1	21.0	29.4	22.4
	Greedy	18.4	20.1	20.6	28.6	22.0
	Viterbi	4.5	7.4	4.8	10.8	6.9

5.2 Empirical settings

In a real use of the proposed methods, the most expensive step is the production of the required training data. Therefore, we try to limit the amount of training deeds in our experiments as much as possible. To this end, we will obtain results using only one bundle as a training set and the other three as a test set. That is, we perform four experiments—each using a different bundle for training. This protocol will allow us to study training data requirements in more detail by further limiting the amount of deeds used from each training bundle. Note that, for training it is not feasible to use isolated samples from bundles, since it would then be impossible to employ the entire contexts of that bundles for decoding them. Thus training with full bundles in this approach is more aligned with a real-world use case.

Note that the metrics discussed in Sect. 4 are defined at bundle level; for instance in Eq. (16) CAER is defined as the cost of all deed edit operations for a bundle, normalized by the total number of page images of the bundle. Since in the proposed protocol each experiment entails testing with three bundles, we do micro-averaging to compute the metrics; that is, the cost is accumulated over the three bundles and finally normalized by the total number of pages of these bundles.

The image-classifiers explained in Sect. 3.3 are trained, at least, for 15 epochs and a maximum of 30, with an early stopping of 5 epochs concerning to the evaluation loss. For all the models, a 15% of the training set was randomly selected for validation. The AdamW optimiser [17] was used with a learning rate of 0.001 with a batch size of 4, except for ConvNeXt, where it was reduced to 2 due to memory limitations. A learning rate decay of 0.5 was applied after every 10 epochs. All of the images were resized to 1024×1024.

To mitigate fluctuations due to the initialization of parameters of the learning algorithms, all the values reported in tables and curves in Sect. 6 are averages of results obtained with 10 random parameter initializations.

6 Results

First, we evaluated the image classifiers discussed in Sect. 3.2, to measure the quality of the image class-posterior distribution they provide. Table 2 reports the cross-entropy for the different models trained with each bundle, along with the overall average.

ResNet50 clearly outperforms the other classifiers, followed by the ConvNeXt model. Recall that no decoder was used in these experiments. According to these results, the model with the lowest cross-entropy, ResNet50, was selected to conduct the following tests where the test bundles are actually segmented.

These end-to-end bundle segmentation results were obtained using the posteriorgrams produced by ResNet50 and each of the decoders proposed in Eqs. (7–9) of Sect. 3.2. Results in terms of both BSER and CAER metrics are reported in Table 3.

We first notice that the trends shown by BSER and CAER are very similar, indicating a high correlation between the two metrics, with BSER's average slightly higher. This is

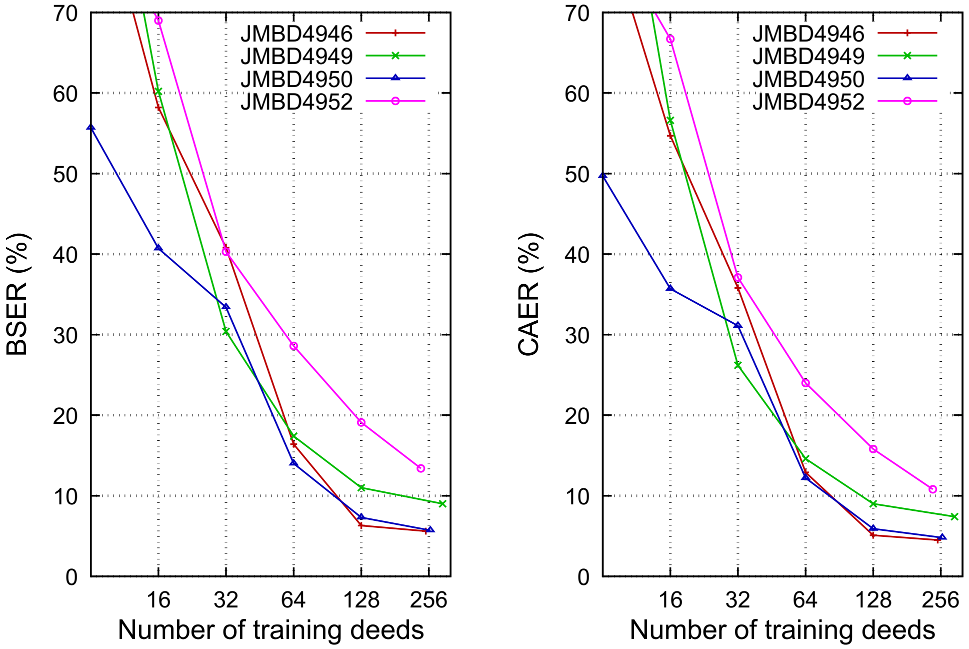

Fig. 3 Learning curve when training with each bundle. Page class posteriors (posteriorgrams) were obtained with a ResNet50 classifier and decoding was carried out with the Viterbi algorithm

because, while BSER considers all the book pages equally important, giving them the same weight, CAER does not. As explained in Sect. 4.2, CAER will penalize errors more heavily on a page with more textual content than on an empty one since it penalizes the wrong segmented content, not the wrong segmentation itself.

As expected, Viterbi decoding performs significantly better than the other (unconstrained and greedy) approaches in all the four experiments. The last column shows the overall average, noting that the Viterbi performance is more than three times better than that of the other decoders. This makes it clear that considering the global context of the entire book, we have chances of achieving a far better segmentation. On the other hand, approaches that consider local context of the posterior probabilities yield results significantly worse. This demonstrates that the problem is not trivial enough to rely only on image classification without considering the global context.

The best result is achieved when training on JMBD4946, with 5.6% BSER and 4.5% CAER. The worst result is for training on JMBD4952, reaching a 13.4% BSER and 10.8% CAER. These results align with the average number of pages per deed, showing that training with larger deeds lead to a lower segmentation error on the test bundles. Each of the first three bundles have an average of close to 6 pages per deed, leading the low segmentation errors. In comparison, JMBD4952, with 4.2 pages per deed (the lowest count), exhibits the highest segmentation error.

Finally, learning curves were obtained to study the training needs of the final proposed approach; namely, individual page image class-posteriors computed by a ResNet50 classifier, followed by Viterbi decoding. For each bundle, an increasing number of deeds are considered for training and the other three complete bundles are used for testing, as in the experiments of Table 3. The number of training deeds was chosen in powers of two, preserving the order in which they appear in the bundle, without repetition. As in the previous experiments, each test was repeated 10 times, and average results were reported. Figure 3 shows these results.

Segmentation error decreases monotonically with the number of training deeds. Notably, from 64 deeds onwards, we begin to obtain useful segmentation performance, except for JMBD4952 which, as already noted, seems to be less informative for training because of the smaller size of its deeds. When training with any of the other three bundles, 128 deeds appear to be enough since errors do not decrease significantly when twice as many training samples are used. On the other hand, the trend at the end of the JMBD4952 curve suggests that better results could be achieved if more deeds of this bundle where available for training.

7 Conclusion

We have proposed a system that, for the first time, can segment a large historical manuscript image bundle into the (many) multi-page deeds it encompasses. The proposed approach entails training an individual page image classifier to compute the class posterior of each page image of a full test bundle. This sequence of posteriors is then “decoded” to obtain a maximum probability class-label sequence which fullfills the full-bundle consistency constraints required for a proper deed segmentation.

Different classifiers and decoders have been studied and compared to finally select the best combination. An important conclusion of this study is that adequately using the bundle's global context not only ensures segmentation consistency but it leads to much better segmentation decisions.

On the other hand, we have introduced two new end-to-end metrics to evaluate the raw and content-aware segmentation accuracy at the bundle level. While the first metric, BSER, measures the segmentation errors, the second, called CAER, measures the loss of textual information caused by segmentation errors.

In future works, we will study whether an appropriate calibration of the posteriors may improve the segmentation using the Viterbi decoder [10]. We also plan to integrate deed segmentation and deed typology classification, as studied in Ref. [8, 21, 23], taking into account both textual and visual information in a multimodal way [20]. In this way, we also plan to work with deeds without the page-level simplifying restrictions exhibited by the corpora of this paper. In this regard, progress has been made with unrestricted deed corpora, utilizing both visual and textual information, where we hope results will be published soon [19].

Appendix A

Derivation of Eq. (4)

Departing from Eq. (2):

$$\begin{aligned} P(c_1,\dots,c_N \mid B) &\approx P(c_1 \mid G_1)\prod_{j=2}^{N} P(c_j \mid G_j, c_{j-1}) \\ &= P(c_1 \mid G_1)\prod_{j=2}^{N} \frac{P(c_j, c_{j-1}, G_j)}{P(c_{j-1}, G_j)} \\ &= P(c_1 \mid G_1)\prod_{j=2}^{N} \frac{P(c_{j-1})P(c_j \mid c_{j-1})P(G_j \mid c_j, c_{j-1})}{P(c_{j-1})P(G_j \mid c_{j-1})} \qquad (17) \\ &\approx P(c_1 \mid G_1)\prod_{j=2}^{N} P(c_j \mid c_{j-1})\frac{P(G_j \mid c_j)}{P(G_j)} \\ &= P(c_1 \mid G_1)\prod_{j=2}^{N} P(c_j \mid c_{j-1})\frac{P(c_j \mid G_j)}{P(c_j)} \end{aligned}$$

In the first step, two independent assumptions have been made: $P(c_j \mid G_1, \dots, G_N)$ only depends on G_j, and page class dependency is first-order Markov. In the fourth step, it is assumed that G_j is conditionally dependent only on c_j and inconditionally independent of c_{j-1}—so $P(G_j \mid c_{j-1}) = P(G_j)$. In the last step, the Bayes' rule has been used (again) to express the conditional likelihood $P(G_j \mid c_j)$ as a function of the posteriors $P(c_j \mid G_j)$.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature. Work partially supported by the research grants: the SimancasSearch project as Grant PID2020-116813RB-I00a funded by MCIN/AEI/ 10.13039/501100011033 and ValgrAI - Valencian Graduate School and Research Network of Artificial Intelligence and the Generalitat Valenciana, co-funded by the European Union. The second author's work was partially supported by the Universitat Politècnica de València under grant FPI-I/SP20190010. The third author's work is supported by a María Zambrano grant from the Spanish Ministerio de Universidades and the European Union NextGenerationEU/PRTR.

Availability of data and materials The datasets generated and analyzed during the current study are available from the corresponding author on reasonable request https://zenodo.org/records/10418179.

Declarations

Ethical approval Not applicable.

References

1. Andrés J, Prieto JR, Granell E et al (2022) Information extraction from handwritten tables in historical documents. Document analysis systems (DAS), vol 13237. LNCS Springer, Cham, pp 184–198
2. Biswas S, Riba P, Lladós J et al (2021) Beyond document object detection: instance-level segmentation of complex layouts. Int J Doc Anal Recognit 24:269–281
3. Boillet M, Kermorvant C, Paquet T (2021) Multiple document datasets pre-training improves text line detection with deep neural networks. 2020 25th International conference on pattern recognition (ICPR). IEEE Computer Society, Los Alamitos, pp 2134–2141
4. Boillet M, Kermorvant C, Paquet T (2022) Robust text line detection in historical documents: learning and evaluation methods. Int J Doc Anal Recognit 25:95–114
5. Bosch V, Toselli AH, Vidal E (2012) Statistical text line analysis in handwritten documents. In: 2012 International conference on frontiers in handwriting recognition (ICFHR), IEEE, pp 201–206
6. Campos VB (2020) Advances in document layout analysis. Ph.D. thesis, Universitat Politècnica de València

7. Coquenet D, Chatelain C, Paquet T (2023) DAN: a segmentation-free document attention network for handwritten document recognition. In: IEEE transactions on pattern analysis and machine intelligence pp 1–17
8. Flores JJ, Prieto JR, Garrido D, et al (2022) Classification of untranscribed handwritten notarial documents by textual contents. In: Proceeding of the 2022 Iberian conference on pattern recognition and image analysis (IbPRIA), Springer, Cham pp 14–26
9. Grüning T, Leifert G, Strauß T et al (2019) A two-stage method for text line detection in historical documents. Int J Doc Anal Recognit 22:285–302
10. Guo C, Pleiss G, Sun Y, et al (2017) On calibration of modern neural networks. In: Precup D, Teh YW (eds) Proceedings of the 34th ICML, Proceedings of machine learning research, vol 70. PMLR, pp 1321–1330
11. He K, Zhang X, Ren S, et al (2015) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), Las Vegas, pp 770–778
12. He K, Gkioxari G, Dollar P, et al (2017) Mask R-CNN. vol 2017-October. Institute of Electrical and Electronics Engineers Inc., Venice, Italy, pp 2980–2988
13. Hélary X (2006) Le cartulaire de la seigneurie de nesle [chantilly, 14 f 22]
14. Kim G, Hong T, Yim M et al (2022) Ocr-free document understanding transformer. In: Avidan S, Brostow G, Cissé M et al (eds) Eur Conf Compu Vis (ECCV). Springer, Cham, pp 498–517
15. Liu Z, Lin Y, Cao Y, et al (2021) Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF ICCV, pp 10012–10022
16. Liu Z, Mao H, Wu CY, et al (2022) A convnet for the 2020s. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR) pp 11976–11986
17. Loshchilov I, Hutter F (2019) Decoupled weight decay regularization. In: 7th International conference on learning representations, ICLR 2019, New Orleans, LA, USA, May 6–9, 2019. OpenReview.net, New Orleans
18. Oliveira SA, Seguin B, Kaplan F (2018) DhSegment: a generic deep-learning approach for document segmentation. In: 16th International conference on frontiers in handwriting recognition (ICFHR) 2018-August:7–12
19. Prieto JR (To be defended) Deep learning methodologies for textual and graphical content-based analysis of handwritten text images. Ph.D. thesis, University of Politècnica de Valéncia
20. Prieto JR, Bosch V, Vidal E, et al (2020) Text content based layout analysis. In: 17th International conference on frontiers in handwriting recognition (ICFHR), IEEE, pp 258–263
21. Prieto JR, Bosch V, Vidal E, et al (2021) Textual-content-based classification of bundles of untranscribed manuscript images. In: 2020 25th International conference on pattern recognition (ICPR), IEEE, pp 3162–3169
22. Prieto JR, Becerra D, Toselli AH et al (2023) Segmentation of large historical manuscript bundles into multi-page deeds. In: Pertusa A, Gallego AJ, Sánchez JA et al (eds) Pattern recognition and image analysis. Springer, Cham, pp 121–133
23. Prieto JR, Flores JJ, Vidal E et al (2023) Open set classification of untranscribed handwritten text image documents. Pattern Recognit Lett 172:113–120
24. Puigcerver J (2018) A probabilistic formulation of keyword spotting. Ph.D. Thesis, University of Politècnica de València
25. Quirós L (2022) Layout analysis for handwritten documents. a probabilistic machine learning approach. Ph.D. thesis, Universitat Politècnica de València
26. Quirós L, Bosch V, Serrano L, et al (2018) From HMMs to RNNs: computer-assisted transcription of a handwritten notarial records collection. In: 16th International conference on frontiers in handwriting recognition (ICFHR), IEEE, pp 116–121
27. Quirós L, Toselli AH, Vidal E (2019) Multi-task layout analysis of handwritten musical scores. In: Proceedings of the 2019 Iberian conference on pattern recognition and image analysis (IbPRIA), Springer, Berlin, pp 123–134
28. Quirós L, Serrano L, Bosch V, et al (2018) Oficio de Hipotecas de Girona. A dataset of Spanish notarial deeds (18th Century) for Handwritten Text Recognition and Layout Analysis of historical documents. Zenodo
29. Tarride S, Maarand M, Boillet Mea (2023) Large-scale genealogical information extraction from handwritten quebec parish records. Int J Doc Anal Recognit
30. Vidal E et al. (2020) The Carabela project and manuscript collection: large-scale probabilistic indexing and content-based classification. In: 17th International conference on frontiers in handwriting recognition (ICFHR), pp 85 – 90
31. Vidal E, Toselli AH, Ríos-Vila A et al (2023) End-to-end page-level assessment of handwritten text recognition. Pattern Recognit 142:109695
32. Wolf T, Debut L, Sanh V, et al (2020) Transformers: State-of-the-art natural language processing. In: Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations. Association for computational linguistics, Online, pp 38–45