Digitizing History: Transitioning Historical Paper Documents to Digital Content for Information Retrieval and Mining-A Comprehensive Survey

Nancy Girdhar, Mickaël Coustaty, Antoine Doucet

Digitizing History: Transitioning Historical Paper Documents to Digital Content for Information Retrieval and Mining—A Comprehensive Survey

Nancy Girdhar, Mickaël Coustaty, *Member, IEEE*, and Antoine Doucet

***Abstract*—Historical document processing (HDP) corresponds to the task of converting the physical-bind form of historical archives into a web-based centrally digitized form for their *conservation*, *preservation*, and *ubiquitous access*. Besides the conservation of these invaluable historical collections, the key agenda is to make these *geographically distributed historical repositories* available for information mining and retrieval in a *web-centralized touchless mode*. Being a matter of interest for interdisciplinary scholars, the endeavor has garnered the attention of many researchers resulting in an immense body of the literature dedicated to digitization strategies. The present study first assembles the prevalent tasks essential for HDP into a pipeline and frames an outline for a generic workflow for historical document digitization. Then, it reports the latest task-specific state of the art which gives a brief discourse on the methods and open challenges in handling historical printed and handwritten script images. Next, grounded on various layout attributes, it further talks about the evaluation metrics and datasets available for observational and analytical purposes. The current study is an attempt to trail the contours of undergoing research and its bottlenecks thus, providing readers with a comprehensive view and understanding of existing studies and unfolding the open avenues for the future outlook.**

***Index Terms*—Archival images, handwritten, historical document, image processing, information retrieval, layout analysis, natural language processing, printed.**

I. Introduction

HISTORY speaks volumes. Historical documents narrate events exhaustively and are evidence of happenings of the past. These documents are good fetching grounds for the enormous knowledge about the past and the culturally poised heritage. There is a myriad of historical and culturally significant documents available in archives, data repositories, and collections in *handwritten* and *printed* forms (*articles*, *journals*, *books*, *newspapers*, etc.) that gives insights into the arts, languages, and politics, and even symbolize arbitrary power. However, *searching*, *indexing*, and *accessing* these vast repositories and libraries to gain information on a specific topic is a very arduous and challenging task. Due to the importance of these treasured historical documents *representing*, *registering*, and *symbolizing* culturally prominent records which are of great value to many disciplines including *digital heritage*, *digital humanities*, *digital discipline*, and so on, and with the forecasted potential impact of digitization over the way how the information will be accessed have gained the interests of many historians, researchers, and scholars from diverse fields to delve into the digitization efforts. Furnishing the artificial intelligence (AI) touch to these old collections not only preserves them but also can revolutionize the way historical knowledge is being served to the general public for information mining and retrieval purposes.

A plethora of work is dedicated to the literature that windows from fundamental model frameworks to tools development to provide *web-centralized touchless access* to the *wide geographically distributed historical collections* [1]. There have been prolific advances in machine [2], [3], [4], [5] and deep learning-based models [6], [7], [8], [9] for the digitization of both printed and handwritten historical documents. Numerous attempts based on historical text recognition (HiTR) [10], [11] and historical optical character recognition (HOCR) [12], [13], [14] are conducted to get satisfactory information mining and retrieval results. Many competitions and campaigns are organized for various tasks from *system development* to *model-evaluation*. Multiple large-scale projects are anticipated for developing document analysis systems (DASs) to let them be saved from *perishable pages* to *preserved digitized web pages*. Despite all the visible impetus in digitization efforts, there are still many gaps that need to be addressed given the *complex*, *diverse*, *heterogeneous*, and *variable layout structures* of historical documents specifically when it comes to the digitization of the historical handwritten documents (HHDs) [15], [16], [17] with the disparity in *font-types*, *text-sizes*, and *writing styles* [1], [9], [18], [19], [20].

The current article orbits around three major aspects of historical document processing (HDP). *Initially*, it presents a generic workflow pipeline by assembling the various essential tasks for HDP. *Next*, the presented work summarizes task-specific state of the art and their challenges that elucidate readers with information on the methodologies, algorithms, and

Manuscript received 6 September 2023; revised 5 February 2024; accepted 11 March 2024. This work was supported by the ANNA (2019-1R40226), TERMITRAD (AAPR2020-2019-8510010), Pypa (AAPR2021-2021-12263410), and Actuadata (AAPR2022-2021-17014610) projects through Nouvelle-Aquitaine Region, France. *(Corresponding author: Nancy Girdhar.)*

The authors are with L3i, University of La Rochelle, 17000 La Rochelle, France (e-mail: nancy.girdhar@univ-lr.fr).

Digital Object Identifier 10.1109/TCSS.2024.3378419

techniques developed to achieve efficient historical document digitization. *Finally*, the study then compiles several datasets and the evaluation metrics employed in the literature by the researchers to measure the performance of their efforts and contributions toward enhancing the results. The significant contributions of the current survey and the questions it addresses can be summarized as follows.

1) *Q1:* What are the key aspects and challenges in HDP?
2) *Q2:* What are the specific tasks involved in the structural and semantic analysis of historical documents?
3) *Q3:* What strategies and techniques have been developed to address challenges in various historical document analysis tasks?
4) *Q4:* What tools and software are available for processing historical documents for layout analysis?
5) *Q5:* Which datasets and evaluation parameters are commonly used in this field for performance assessment?

The rest of the survey is structured as follows. In Section II, HDP challenges are discussed, and the current research rationale is presented. Section III introduces a general digitization workflow, covering historical data acquisition with preprocessing details, layout analysis, and automatic text recognition (ATR). Section IV includes information on historical datasets and evaluation metrics. Last, Section V provides a summary of the survey along with its conclusions.

II. Survey Framing: Historical Documents Processing Challenges and Research Rationale

This section discusses the challenges and rationale behind processing and digitizing historical documents. Although important, this task can be quite challenging and demanding due to various factors that must be considered when converting physical document images into digital form for mass access. While preserving the document is a straightforward task using facsimile, making it available for data searching, mining, and retrieving is an arduous process that requires significant effort.

A. Challenges in Handling Historical Documents

Processing and digitization of historical documents are undoubtedly important, however very tenacious and demanding tasks. There are many factors that need to be considered and taken care of while converting this *physical document images* into *web digital form*. Considering preservation as the only objective, transforming a physical document into digital form is a simple task by facsimile. However, besides preservation making available these documents for data *searching*, *mining*, and *retrieving* facilities is a very laborious and strenuous task. Handling and processing historical documents present numerous challenges as follows.

1) Document degradation factors [21] (discussed in Section III-A).
2) Vast array of document layouts featuring multiple languages, fonts, and writing styles [18], [22].
3) Different handwriting styles with varying font types, sizes, and line spacing [23].
4) Text-oriented issues such as skew, slant, and rotated text [24] (as illustrated in Fig. 1).
5) Lack of standardized metadata and formatting, leading to inconsistencies in labeling criteria and semantic annotations [25].

All these compelling factors add up to the bottlenecks in the processing of historical documents, particularly handwritten documents.

B. Previous Surveys and Target Audience

Several efforts and studies are contributing to the literature for handling the aforementioned challenges by the interdisciplinary research community resulting in a surplus number of methodologies, algorithms, and techniques [26]. Some works have collected, elucidated, and summarized these myriads of articles [1], [18], [27]. There are several other task-oriented surveys compiling the various studies and highlighting the trends, techniques, and gaps for a particular task (*page segmentation* [28], [29], *text line segmentation* [20], [29], *character segmentation* [30], [31], and so on). A few dimension-related studies focused on HiTR [22], historical document layout analysis (HDLA) [1], [9], and HOCR are also available in the literature that provides a summary of the various advances in the techniques and enhanced tools available specific to that dimension [18], [32]. However, it has been observed that a comprehensive synthesis of the entire HDP workflow is lacking in the existing literature. This study aims to address this gap by presenting a roadmap for general HDP, providing insights into methodologies and techniques developed over the past decades. The distinctive features of our survey that distinguish it from previous works are as follows.

1) *Comprehensive Workflow Synthesis*: Unlike existing surveys that often focus on specific tasks or dimensions of HDP, our study provides a holistic and integrated view of the entire HDP workflow. We systematically connect the essential steps, offering a comprehensive understanding of the complete pipeline.
2) *Bridge Across Subfields*: While there are task-oriented surveys in subfields such as page segmentation [28], [29], text line segmentation [20], [29], and character segmentation [30], [31], our survey acts as a bridge across these subfields. We integrate insights from various studies, compiling methodologies and techniques that are scattered throughout different dimensions of HDP.
3) *Addressing the Research Gap*: We recognize and address the observed dearth of studies that comprehensively review and discuss the entire HDP workflow. By synthesizing existing works, approaches, algorithms, tools, software, and datasets, we provide a consolidated roadmap, filling the gap and facilitating research efforts, especially for novice researchers.
4) *Practical Utility for Archivists and Historians*: Beyond catering to the research community, our survey aims to serve as a practical aid for archivists and historians involved in digitization projects. We provide insights into

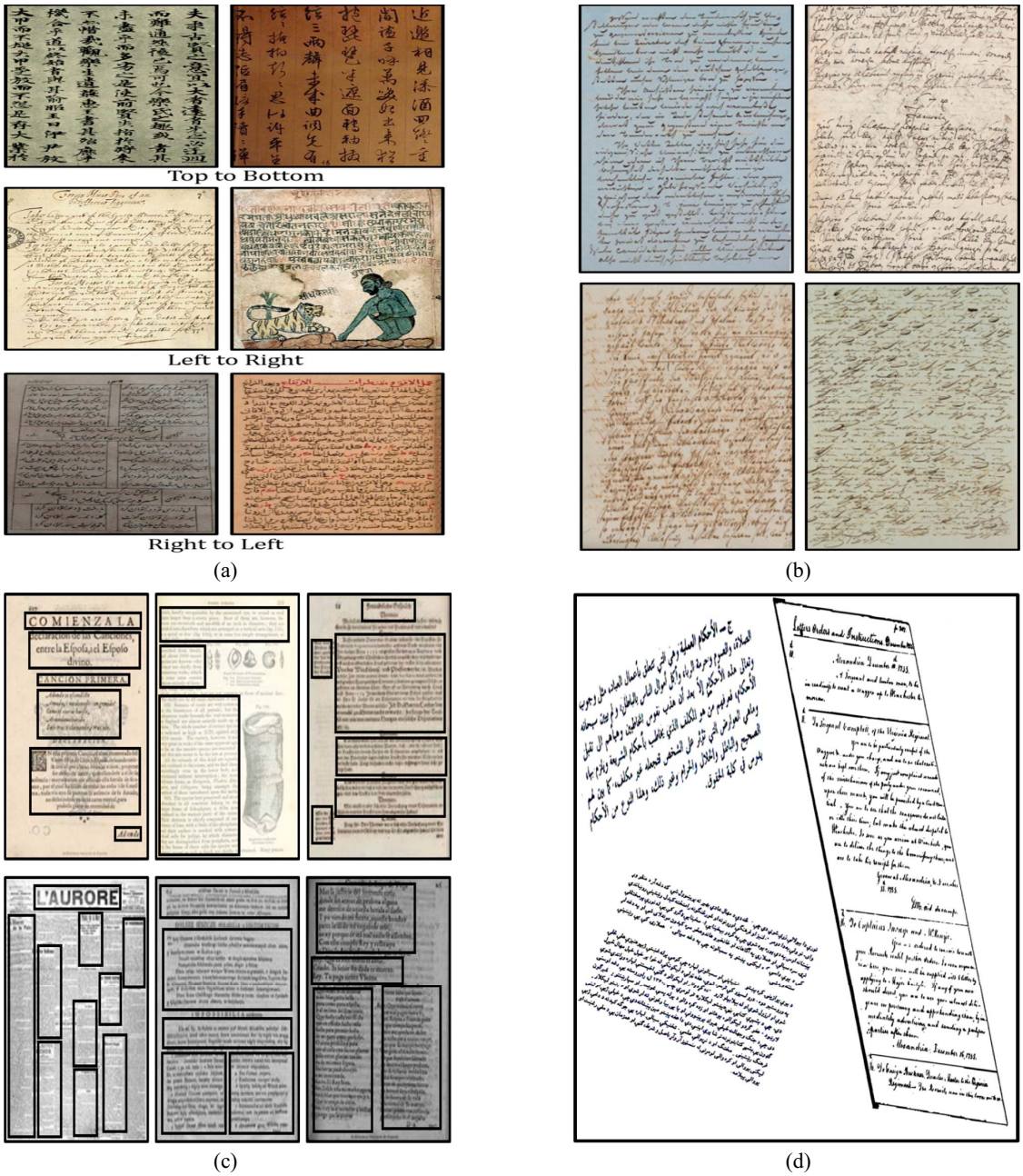

Fig. 1. (a) Various writing directions (top-left) [22], (b) font styles (top-right) [23], (c) layouts (bottom-left) [18], and (d) text orientations (bottom-right) [24] exhibited by historical documents.

efficient tools, enhancing the applicability and impact of our work in real-world archival and historical preservation scenarios.

This current study attempts to provide a road map for general HDP which can then acquaint novice researchers with insight into the existing methodologies and techniques proposed over the past decades. We present a comprehensive survey to put together the various advances achieved as well as the challenges and bottlenecks that still need to be addressed to achieve effective document processing and information retrieval. This study offers a reasonable understanding of the field and will serve as an aid for archivists and historians to elect efficient tools for their digitization projects. Overall, this study canvasses a complete pipeline imperative for HDP, connecting all the essential steps.

III. Generic Workflow Pipeline for Historical Documents Processing and Digitization

This section presents the generic workflow pipeline for historical document image processing and digitization (see Fig. 2). The section elaborates on various essential stages in the pipeline each representing a necessary task, including historical document image acquisition, preprocessing, structural and semantic

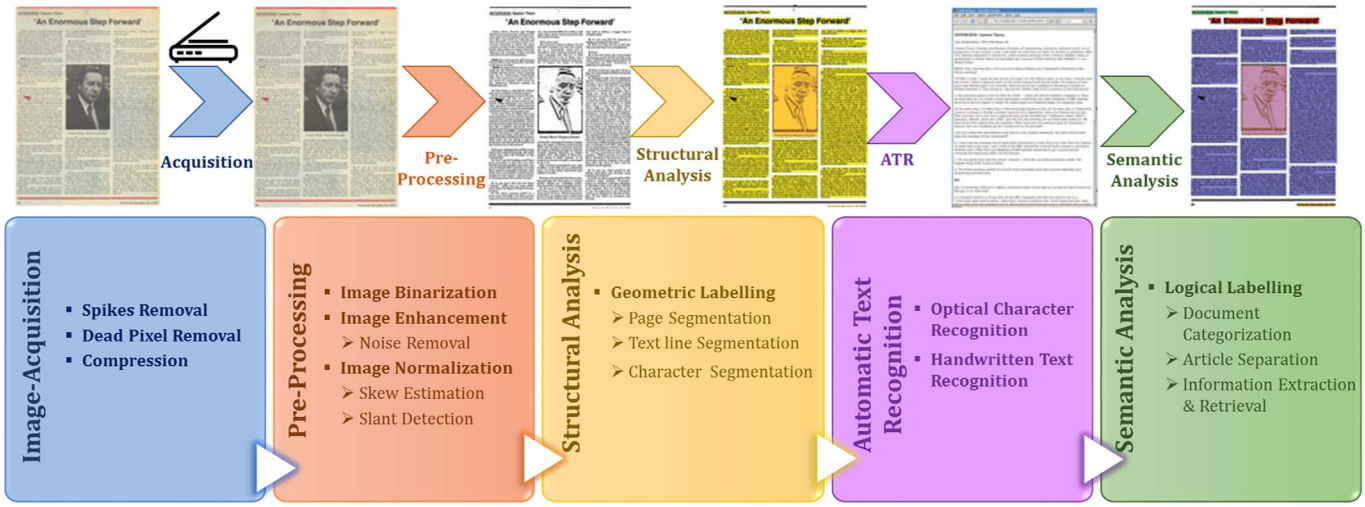

Fig. 2. Workflow for historical documents processing.

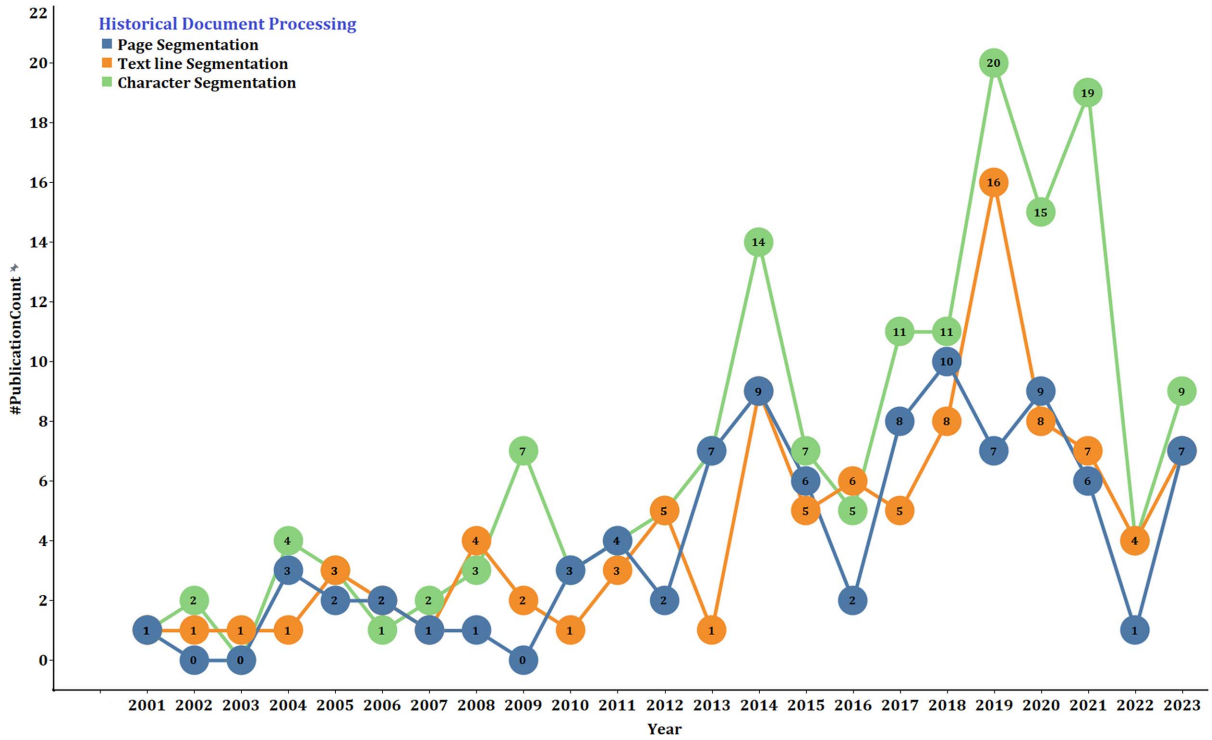

Fig. 3. Publication counts illustrating trends across various tasks in HDP (2001–2023).

analyses of both handwritten and printed layout images, with a brief on panoptic-based recent advancements, and lastly ATR with details on both handwritten text recognition (HTR) and HOCR.

For the current study, we considered publications from the Scopus database, focusing on the English language from 2001 to 2023. The search string, TITLE-ABS-KEY ((historical AND document AND (processing OR digitization))) AND PUBYEAR $>$ 2000 AND PUBYEAR $<$ 2024 AND (LIMIT-TO (LANGUAGE, "English")), yielded 1753 articles. We then refined these articles through structural, semantic, and panoptic analyses. In the structural analysis, we filtered and categorized articles based on tasks such as *page segmentation*, *text line segmentation*, and *character segmentation*. Fig. 3 illustrates the publication counts of trends from various tasks during the survey period.

Document digitization necessitates good document quality to *detect*, *recognize*, and *analyze* the text correctly for a human to computer-readable form. In contrast to humans for which the image quality can be defined as the ability to read the content comfortably, for computers, the image quality is determined in terms of *decipherability* (able to recognize and read a specified

character form accurately), *legibility* (able to recognize and read a specified word form correctly), and *readability* (able to recognize and read a specified chain of words correctly) [21], [33]. Overall, the error rate of recognizing a word or a character by a computer can be used as an indication to establish the quality of an image; a high rate of recognition error suggests poor image quality and vice versa. The image quality depends on several factors classified broadly as *on-creation* and *on-acquisition* parameters.

The parameters related to the quality of *material*, *type of content*, and *production techniques* used for creating a document are referred to as *on-creation* parameters, while the parameters related to the *methodology* used for acquiring images are known as *acquisition parameters*. In the case of historical documents, it is often difficult to obtain information about on-creation parameters, so researchers tend to focus more on acquisition parameters in the process of image processing and digitization of historical documents.

A. Image Acquisition

Image acquisition refers to the process of converting a document into a digitized form. There are several methods to convert a physical spatially documented image into its digital representation. While traditional scanners have been widely used, advancements in technology have introduced alternative devices such as digital cameras with various sensors. The choice of the capturing device significantly influences the quality of the digital representation. The quality of the digitized image is susceptible to numerous factors, including defects in the paper (*stains*, *speckles*, *wrinkles*, etc.), faults introduced during printing (*ink bleeding*, *toner dropout*, *baseline variations*, *show-through*, etc.), and flaws induced during scanning (*blur*, *skew resolution reduction*, *noise*, *compression artifacts*, *binarization*, *threshold*, *sensitivity*, *jitter*, etc.).

Various physical factors can affect the degradation of document quality over time, such as the *printing material quality*, *paper quality*, *gloss*, *yellowing*, *watermarks*, and *environmental factors* such as mold and infestation. The quality of documents can also be influenced by *photocopying* or *facsimile* methods. In addition to these physical and environmental factors, typography factors such as *typeface*, *font size*, *line length*, *margins*, and *boldness* can also introduce distortions that impact image quality. These factors, when combined, can significantly reduce the contrast between the text and the background, making it difficult to differentiate them visually. The accuracy of content recognition in a document is highly dependent on the document's creation and acquisition process. During the data acquisition process, unintended *artifacts* may also be introduced, which can further alter the results of analysis and recognition.

To improve the quality of HDP, different techniques are used, such as *spikes removal*, *dead-pixel removal*, and *compression*, to mitigate various degradation factors that hinder image quality [21], [33]. The goal of these techniques is to restore or reproduce the original image without altering the pixel values, and this is referred to as lossless compression schemes. Fig. 4 illustrates various degradation factors and challenges that can occur during the acquisition phase of historical document images.

B. Document Image Preprocessing

A colossal amount of work has been dedicated to the literature on image processing [33], [34], [35]. Different algorithms are evolved that are adopted at various phases for document image analysis for instance *skeletonization*, *projection profiles*, *morphological operations*, and *Hough transformation* approaches. Other methods based on *distance calculation*, *labeling connected components,* and *run-length smoothing* [36] have also been developed. Different imaging methods exist for document image binarization, which operates at different thresholding levels. Several efforts have been made for document image enhancement using techniques such as *histogram equalization* and *low-pass Wiener Filter*. There are contributions made to develop algorithms for document image normalization to address issues related to document alignment, such as *skew*, *slant*, and *text rotation* [22], [23], [24].

1) Document Image Binarization: There are various names used in the literature to refer to image binarization, including *thresholding*, *text segmentation*, and *background segregation and elimination*. These terms describe the process of separating text from background areas by converting a gray-scale (or colored) image to a binary image using a thresholding technique. This process not only saves storage space but also enhances content readability and enables quick and efficient page segmentation and recognition. Threshold selection methods are typically categorized into three main groups: *global*, *local*, and *hybrid*. Table I presents various image binarization techniques existing in the literature.

a) *Local and Global Thresholding:* Binarization is an essential however challenging task given the image degradation, nonuniform background intensity, low contrast, smear, and shadows on historic documents [26], [37]. The *global thresholding* approach works better for images having a distinctive separation between *foreground area* and *background area*; however, it falls short when there is overlap between two areas exists [52], [53], [54]. To counter the limitation, *local (adaptive) thresholding* approaches are taken into account that can perform with *spatially uneven grey levels* resulting from degradation [38], [39], [40], [41], [42], [43].

b) *Hybrid Thresholding:* To take complementary advantage of both, there exist some *hybrid thresholding* approaches that can classify between a *text pixel* and a *background pixel* [26], [34], [44], [45], [46].
In addition to this, there are some machine learning techniques that use training set [2], [47], [48] to distinguish between text and background pixels. There are some studies for the binarization of colored documents, where information on colored pixels is used for document processing [49], [50], [51].

2) Document Image Enhancement: To enhance image quality by mitigating various degradation factors such as bleed/shadow through, noise, uneven/low contrast, and damaged

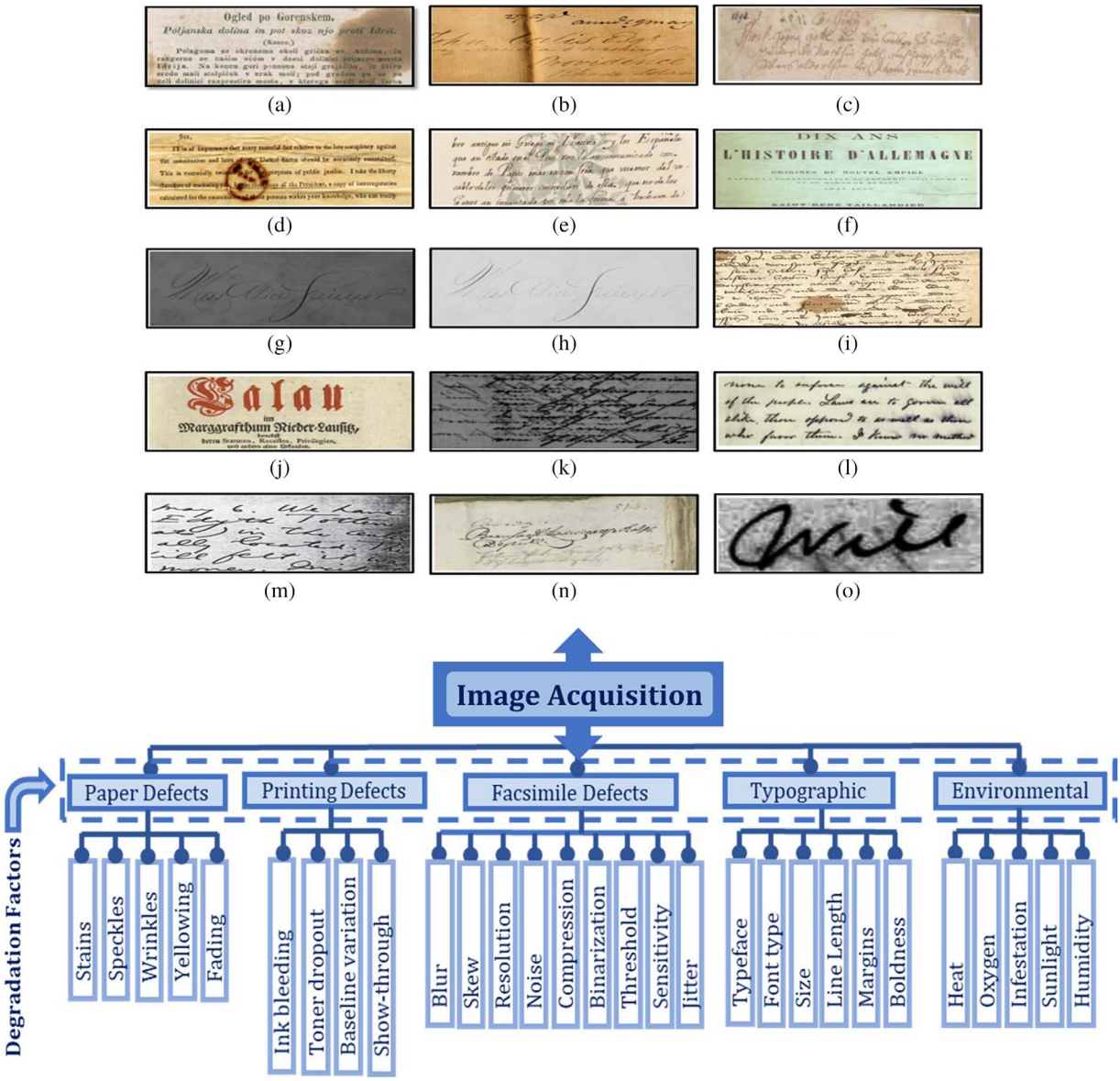

Fig. 4. Historical documents image degradation factors and acquisition challenges [21]. (a) Stained. (b) Creased. (c) Faded. (d) Stamped. (e) Background texture. (f) Varying font text. (g) Varying font size. (h) Thin strokes. (i) Text alternation. (j) Varying font color. (k) Bleed through ink. (l) Smeared ink. (m) Complex texture. (n) Border noise. (o) Compression artifacts.

characters image enhancement is utilized. Table II provides an overview of various image enhancement techniques existing in the literature. One of the methods is the low-pass *Wiener Filter* which is used to accomplish image enhancement during the binarization stage [44]. The authors of [55] improved image contrast and reduced background noise for better readability by adopting *histogram equalization* preprocessed with a generalized fuzzy operator. To flatten background gray levels, reduce noise, and smooth character parts by utilizing neighboring data information, Likforman-Sulem et al. [56] combined total variation *regularization* and *nonlocal means filtering* approaches. Additionally, Nomura et al. [57] proposed the adoption of *morphological operations* to eliminate critical shadows from the background caused by low contrast, poor quality, nonuniform illumination, and change in shape or size before performing binarization.

To overcome the challenges due to degradation factors such as *shining*, *bleed-through*, and *shadow-through*, existing methods in the literature are majorly classified into *blind* methods and *nonblind* methods respectively. Blind (nonregistration) methods are employed in document image enhancement to address challenges without relying on prior knowledge of the original document or its distortions. These methods utilize image processing algorithms, often grounded in statistical analysis or machine learning, to identify and mitigate issues such as noise, blur, and imperfections. Blind methods independently process each side of the page [38], [41], [58] without specific information about alignment or registration between the recto (front) and verso (back) images, and without prior knowledge of the degradation, making them more versatile but often less accurate. In contrast, nonblind (registration) methods [59], [60] require a verso image of the page, precisely registered with

TABLE I
TECHNIQUES USED IN DOCUMENT IMAGE BINARIZATION

Ref	Technique	Description	Pros	Limitations
[26], [37]	Global Thresholding	Converts a gray-scale or colored image to a binary image using a fixed threshold value.	Simple and computationally efficient. Suitable for images with clear foreground–background separation. Works well when lighting conditions are relatively consistent.	Struggles with images having overlapping foreground and background areas. Sensitive to variations in lighting conditions and intensity levels. May result in the loss of details in areas with low contrast.
[38], [39], [40], [41], [42], [43]	Local (Adaptive) Thresholding	Performs binarization with spatially uneven grey levels resulting from degradation. Adjusts the threshold value locally based on the characteristics of the surrounding pixels.	Can handle images with uneven illumination and nonuniform backgrounds. More robust to variations in lighting and contrast. Preserves details in locally varying regions.	Requires parameter tuning for different image types and conditions. Computationally more expensive than global thresholding. May produce noisy results in regions with complex textures.
[44], [45], [34], [46], [26]	Hybrid Thresholding	Combines global and local thresholding techniques to classify between text and background pixels. Utilizes both spatial and intensity information for accurate binarization.	Can handle images with varying illumination and complex backgrounds. Balances the strengths of global and local thresholding methods. Provides a tradeoff between adaptability and computational efficiency.	May require careful parameter selection for optimal performance. Performance can be sensitive to noise and artifacts in the input image.
[47], [2], [48]	Machine Learning	Utilizes machine learning techniques and training sets to distinguish between text and background pixels. Training data is used to build a model for automatic binarization.	Can adapt to various types of images and document conditions. Capable of learning complex patterns and features. Can improve accuracy with a well-trained model.	Requires a sufficient amount of high-quality training data. The training process can be time-consuming and computationally intensive. Performance may vary depending on the quality and representativeness of the training data.
[49], [50], [51]	Color Document Binarization	Deals with binarization of colored documents, utilizing information on colored pixels for document processing.	Preserves color information while converting to binary. Can handle documents with color variations and highlight important colored features.	Requires additional computational resources compared to gray-scale binarization. More complex processing steps may be involved.

the recto image. These methods rely on explicit information about the degradation process or characteristics of the input image, often requiring a priori knowledge of the distortion introduced during the image acquisition or degradation process, enabling targeted correction. Nonblind methods leverage external information about the document, such as fonts, layout patterns, or even reference versions. This prior knowledge enables them to tailor enhancement techniques to the specific document characteristics, leading to more accurate and targeted results. However, nonblind methods require access to additional resources or rely on accurate assumptions about the document, limiting their applicability in certain situations. After binarization, steps for noise removal, character enhancement, and border removal are generally implemented to enhance the quality of binarized images. For noise removal from the background and character enhancement operations such as *masking*, *morphological operations*, *shrink*, and *swell* is considered [61], [62], [63]. For border removal, methods based on *connected component analysis* [64], [65] and *projection-profile* [66], [67], [68] are employed to enhance the binary document of the image.

3) Document Image Normalization: This refers to the process of handling and restoring the document orientation, including rotated or upside–down text, by estimating and correcting the *slant*, *skew*, *dewarping*, and other alignment *distortions* present in the textual image. Document image normalization mainly includes *page orientation estimation*, *slant detection* (in the case of handwritten historical documents), and *skew correction*. Fig. 5 presents various techniques for detecting skew estimation and slant for historical document image normalization.

Page orientation techniques estimate the page layout and text alignment. The page skew for portrait/landscape orientation falls within 0° to 90°, while the skew can range from 0° to 180° in historical printed and handwritten document images for text alignment (up/down) [69], [70], [71], [72]. In the case of page orientations *(portrait/landscape)*, the page skew is between 0° and 90° [71], [72], whereas in the case of text alignments *(up/down)* the skew can vary from 0° to 180° in historical printed and handwritten document images [69], [70]. Techniques based on projection histograms and black-to-white transition counts are commonly employed to detect orientation.

Several research works have addressed the problem of *skew detection* and *slant estimation* in historical documents. The methods employed for *skew detection* are mainly grounded on the concepts of *projection-profile* [73], [74], [75], *Hough Transform* [76], [77], *nearest-neighbor clustering* [62], [78], [79], *cross-correlation* [80], [81], and *segmentation* techniques [82], [83], [84]. There are also some methods proposed for detecting skew in color documents [85], [86]. *Slant estimations*

TABLE II
TECHNIQUES USED IN DOCUMENT IMAGE ENHANCEMENT

Ref	Technique	Description	Pros	Limitations
[44]	Low-pass Wiener Filter	Enhances image quality during the binarization stage. Mitigates degradation factors such as bleed/shadow through, noise, uneven/low contrast, and damaged characters.	Improves image quality for better binarization results. Reduces degradation factors affecting readability.	Performance may vary depending on the specific degradation factors present in the image.
[55]	Histogram Equalization with Generalized Fuzzy Operator	Improves image contrast and reduces background noise for better readability.	Enhances image quality by adjusting contrast and reducing noise. Preserves important details and features.	May result in over-enhancement or loss of subtle image characteristics. Performance may vary depending on the image content and noise characteristics.
[56]	Total Variation Regularization and NonLocal Means Filtering	Flattens background gray levels, reduces noise, and smooths character parts by utilizing neighboring data information.	Suppresses noise and enhances character details. Improves image quality by reducing background variations.	Computationally intensive for large images. May introduce artifacts or blurring in certain cases.
[57]	Morphological Operations	Utilizes morphological operations to eliminate critical shadows from the background caused by low contrast, poor quality, nonuniform illumination, and changes in shape or size before binarization.	Removes shadows and improves foreground–background separation. Enhances image quality for better binarization results.	Performance may vary depending on the specific shadow characteristics and image conditions.
[58], [38], [41]	Blind Methods	Process each side of the page separately and independently to overcome challenges such as shining, bleed-through, or shadow-through.	Enables independent processing of each page side. Can address specific degradation factors on each side effectively.	May not fully address interferences between the recto and verso sides. Potential misalignment issues between the sides.
[59], [60]	Nonblind Methods	Require a verso image of the page along with a precisely registered recto image to overcome challenges such as bleed, shining, or shadow-through.	Allows for more accurate restoration by considering both recto and verso sides. Addresses interferences between the sides.	Requires the availability of both recto and verso images. Requires precise registration between the sides.
[61], [62], [63]	Noise Removal and Character Enhancement	Operations such as masking, morphological operations, shrink, and swell are used to remove noise from the background and enhance characters.	Improves the quality and readability of the binarized images. Enhances character sharpness and contrast.	May introduce artifacts or loss of fine details in the process. Performance can vary depending on the complexity of the noise and character structures.
[64], [65]	Border Removal (Connected Component Analysis)	Methods based on connected component analysis are employed to remove borders from the binary document images.	Helps in removing unwanted border artifacts and noise. Improves the overall aesthetics of the document.	May inadvertently remove important content near the document borders. Performance can be affected by the complexity of the border structures.
[66], [67], [68]	Border Removal (Projection Profile)	Methods based on projection profiles are used to remove borders from the binary document images.	Effective in removing border artifacts and preserving important content. Can handle documents with irregular border shapes.	Requires accurate estimation of projection profiles, which can be challenging. Performance can be affected by the complexity and variations in the border.

are mainly accomplished using methods based on *projection profile* [87], [88], and *chain code* [89], [90], [91]. There are also methods of slant estimation to analyze near-vertical strokes, as in [92] and [93].

There are numerous *dewarping* techniques available in the literature for image rectification, which are divided into two categories: shape reconstruction for 3-D documents and shape processing for 2-D documents. In shape reconstruction, a special setup or 3-D model is required to extract image features needed for reconstruction [94], [95], [96], while shape processing of 2-D documents does not require any prior information or additional hardware, but it does need 2-D information [97], [98], [99], [100], [101], [102]. Some dewarping techniques depend on text line detection to underline baseline findings [103], [104], while others aim to find the spatial transformations between warped and dewarped document images by analyzing the 2-D information [97], [99]. Table III summarizes various image normalization techniques existing in the literature.

C. Historical Document Structural Analysis

Document layout analysis (DLA) encompasses both document analysis (DA) and document understanding (DU). DA term refers to the *detection* and *extraction* of geometrical features such as blocks and regions from a document image, whereas DU corresponds to *aligning* and mapping the detected geometric blocks/regions to their logical structures. HDLA is essential however, not an easy task to perform as there are numerous arbitrary complex layouts of *printed* (official memos, technical writings, magazines, books, and newspapers) and *handwritten* historical documents. Besides the diversity in page layouts, the various degradation factors (refer Section III-A) and their evolution with modern technical printing tools pose a great challenge to processing these document images.

TABLE III
TECHNIQUES USED IN DOCUMENT IMAGE NORMALIZATION

Ref	Technique	Description	Pros	Limitations
[69], [70], [71], [72]	Page Orientation Estimation	Estimates page layout and text alignment.	Helps in determining the correct orientation of the document. Enables subsequent processing steps to be applied accurately.	Accuracy can be affected by complex page layouts or skewed content.
[73], [74], [75]	Skew Detection (Projection-Profile)	Detects skew using projection-profile techniques.	Allows for the correction of skew, improving document readability. Can handle various types of skew distortions.	May not perform well with heavily skewed documents or complex layouts.
[76], [77]	Skew Detection (Hough Transform)	Detects skew using the Hough Transform method.	Effective in detecting skew even with complex document layouts. Suitable for a wide range of document types.	Computationally intensive, especially for large documents.
[62], [78], [79]	Skew Detection (Nearest-Neighbor Clustering)	Detects skew using nearest-neighbor clustering techniques.	Can handle skewed documents with irregular patterns. Provides robust skew estimation.	Performance may vary depending on the clustering algorithm used.
[80], [81]	Skew Detection (Cross Correlation)	Detects skew using cross-correlation techniques.	Suitable for skewed documents with distinct patterns. Can handle different types of skew distortions.	Performance may degrade with low-quality or noisy documents.
[82], [83], [84]	Skew Detection (Segmentation)	Detects skew using segmentation techniques.	Effective in detecting skew in complex document regions. Can handle documents with multiple columns or irregular shapes.	Performance can be affected by the complexity of the document layout.
[85], [86]	Skew Detection (Color Documents)	Detects skew in color documents using color-based methods.	Can handle skew detection in documents with color information. Provides enhanced accuracy for color documents.	Limited to color document analysis only.
[87], [88]	Slant Estimation (Projection Profile)	Estimation of slant using methods based on projection profile.	Effective in estimating slant angles of text for improved readability. Utilizes projection profile analysis.	Performance can be affected by complex text layouts or nonuniform slant angles.
[89], [90], [91]	Slant Estimation (Chain Code)	Estimation of slant using methods based on chain code analysis.	Provides an alternative approach for slant estimation. Can handle different writing styles and stroke structures.	Performance can be sensitive to noise or irregularities in the chain code representation.
[92], [93]	Slant Estimation (Near-vertical Strokes)	Methods to analyze near-vertical strokes for slant estimation.	Suitable for cases where near-vertical strokes are predominant. Can handle variations in slant angles.	Performance may be limited to specific document types or writing styles.
[94], [95], [96]	Shape Reconstruction (3-D Documents)	Requires a special setup or 3-D model to extract image features for shape reconstruction.	Can handle 3-D documents and complex geometries. Provides accurate shape reconstruction.	Requires additional hardware or setup. May have limitations in handling certain document types or shapes.
[97], [98], [99], [100], [101], [102]	Shape Processing (2-D Documents)	Shape processing methods for 2-D documents without requiring additional hardware or setup.	Does not require special setup or 3-D information. Can handle 2-D documents and general shapes.	Performance can be affected by complex document structures or distortions. Accuracy may vary depending on the complexity of the shape-processing algorithm.
[103], [104]	Text Line Detection	Some dewarping techniques depend on text line detection to underline baseline findings.	Can assist in determining the baseline for dewarping. Helps in aligning the document text lines.	Performance may depend on the accuracy of the text line detection algorithm. Limited to cases where text lines are prominent and well defined.
[97], [99]	Spatial Transformation Analysis	Analysis of spatial transformations between warped and dewarped document images.	Provides insights into the spatial distortions of the document. Can assist in image rectification.	Requires accurate estimation of spatial transformations, which can be challenging. Performance may be affected by complex distortions or irregularities in the document images.

Developing a generic framework for HDLA is very strenuous, as each layout type requires specific treatment to fully exploit the informational content for its processing. The studies on HDLA have been classified into two main subcategories: 1) *document structural analysis*; and 2) *document semantic analysis* [18], as illustrated in Fig. 6. After the preprocessing of the document images, the initial significant subtask in the historical document processing pipeline (HDPP) is document structural analysis. *Document structural analysis* aims to extract the geometric structure of a document to recognize homogeneous region boundaries by labeling them known as geometrical/structural labels [105]. The key tasks involved in document structural analysis include *page segmentation*, *text line segmentation*, and *character segmentation*.

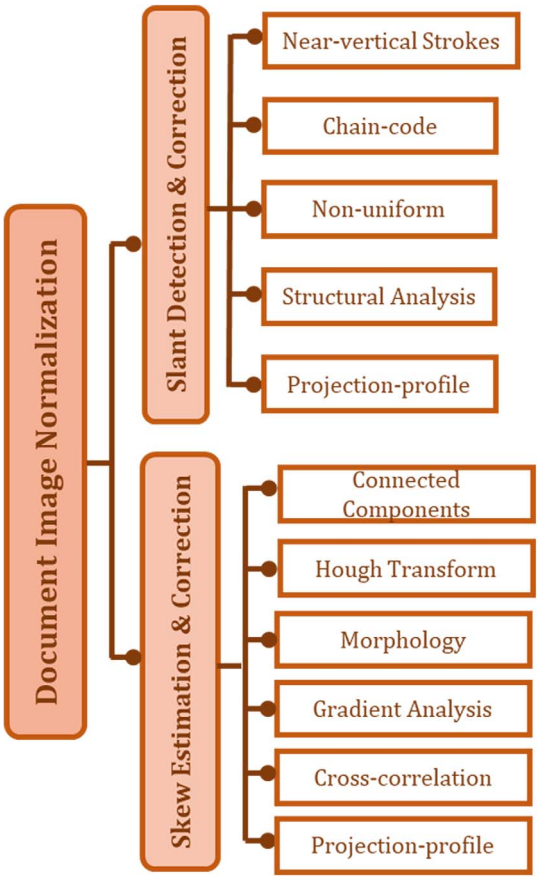

Fig. 5. Historical document image normalization techniques.

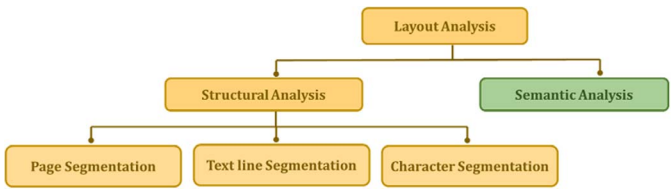

Fig. 6. HDLA.

Table IV summarizes state-of-the-art historical document structural analysis.

1) Page Segmentation: To process historical documents, it is necessary to identify the different parts, such as text, figures, tables, and text lines, which make up each page. The main textual component consists of text blocks, which are groups of text lines that contain characters. The process of separating and extracting these blocks from a page is known as *page segmentation*, while the task of detecting text lines and recognizing characters is referred to as *text line detection* and *character recognition*, respectively.

Page segmentation can be performed using various criteria, such as *object-based* (foreground/background), *primitive based* (pixel, super-pixel, and connected components), and *approach based* (top–down, bottom–up, and hybrid). Fig. 7 summarizes the methods proposed under each criterion.

In the field of historical document analysis, several techniques have been proposed for text region detection. One such approach was presented by Cohen et al. [113], which is a two-step process involving page segmentation and block classification. This method uses a super-pixel representation of images to reduce image-processing complexity, with the first step separating text from nontextual elements and then refining the nontextual regions into noise, drawing, and background by leveraging spatial and color features. Another approach was suggested by Asi et al. [114], which is a learning-free method that formulates text region detection as an optimization task to minimize energy using graph cuts. Additionally, Chen et al. [115] proposed an approach for page segmentation of handwritten historical documents that utilize various color and texture features for pixel classification. They applied a fast correlation-based Filter to select nonredundant and relevant features and improved the segmentation results in the postprocessing phase. Another method was presented by Chen et al. [116], which used deep learning for page segmentation of colored handwritten historical document images. They developed a convolutional auto encoder (CAE) architecture to learn low and high-level features from pixels automatically. The learned features were then utilized to train an SVM classifier for segmentation.

Hadjar and Ingold [106] compared complex layout page segmentation techniques and their adaptations for Arabic newspapers. Grana et al. [112] addressed the problem of automatic segmenting and extracting pictures from text, and floral and abstract decorations in historical manuscripts. They employed a gradient spatial dependency matrix for textual features and SVM for classification. Bukhari et al. [108] proposed machine learning-based layout analysis for Arabic historical document images. They suggested multilayer perceptron classification of connected components for side-notes text segmentation (text appearing in page margins) independent of block segmentation.

In the study conducted by Chen et al. [117], page segmentation for handwritten historical scripts was examined. They treated it as a problem of pixel labeling and proposed a straightforward solution using a convolutional neural network (CNN) with a single convolutional layer trained on labeled super-pixels. The super-pixels were generated using the SLIC algorithm, and the approach involved learning a set of feature detectors and training a nonlinear classifier on the extracted features without any prior knowledge of the document layout. Kaddas and Gatos [118] introduced a deep learning model for page segmentation of handwritten text into distinct text zones. The neural network-based architecture used in this study was based on the *convolutional encoder–decoder* (CED) and took advantage of prior information by incorporating spatial and texture information to label each pixel with a predefined class. They experimented with six different historical handwritten datasets, grouped them into two sets, and trained a model per set to demonstrate the generality of their approach. Hierarchical methods for text segmentation in historical documents have also been proposed, such as the bottom–up greedy approach based on HiTR presented in [119], which supports textual criticism and cultural studies.

TABLE IV
HISTORICAL DOCUMENT STRUCTURAL LAYOUT ANALYSIS

Ref	DT	Obj	Method	Approach	Dataset	Lang	Results (%)
[106]	HN	PS	CC	Visual	ANN:50 pages, ALH:50 pages	Arabic	Rec(ANN): 99.81(fr), 97.598(img), 96.51(tl) Rec(ALH): 90.98(fr), 91.17(img), 92.36(tl)
[107]	HN	PS	Rule based	Visual	100 pages	Greek	P: 75.20, R: 77.15
[108]	HN	PS	ML	Visual	38 documents	Arabic	F: 95.02, Acc: 95
[109]	HN	PS	ML	Visual	999 pages	Italian	Acc: 93.80
[6]	HN	LA	DL	Visual	4135 pages	Swiss	AvgDER: 13.78, AvgCS: 54.44
[110]	HN	LA	ML	Visual	JDG+GDL: 4000K articles	Swiss	Cov(JDG): 73, Cov(GDL): 76
[111]	HN	TBD, TLD	Rule based	Visual	420 pages	French	F: 0.96.2(tb), 97.9(tl), P: 95.9(tb), 96.9(tl), R: 96.6(tb), 99.1(tl)
[112]	HPD	LA	ML	Visual+Text	640 pages (3894 × 2792px)	Italian	P: 52.97(px), 73.36(rg), 85.80(bk), R: 83.49(px), 85.69(rg), 75.87(bk)
[113]	HPD	TS, DS	CC	Visual	252 pages	Arabic	Acc: 90(TS), 92(DS), P: 99.8(TS), 89.6(DS), R: 92.9(TS), 94.9(DS)
[114]	HPD	LA	Learning-free	Visual	38 documents	Arabic	F: 98.36
[3]	HPD	PS	ML	Visual	GW: 20 pages (550 × 850px), SG: 61 pages (500 × 752px), Parz: 39 pages (416 × 624px)	English, German Latin	Acc(GW): 89.5, Acc(SG): 95.7, Acc(Parz): 91.7
[15]	HHD	LA	ML	Visual	KDK: 300K	Dutch	Error: < 11
[115]	HHD	PS	Correlation-based	Visual	GW: 19 pages (138 × 213px), SG: 60 pages (104 × 156px) Parz: 40 pages (138 × 213px)	English, German Latin	Acc: 93
[4]	HHD	PS	DL	Visual	GW: 19 pages (2200 × 3400px), SG: 60 pages (1664 × 2496px), Parz: 40 pages (2000 × 3008px)	English, German Latin	Acc(GW): 88.7, Acc(SG): 97.1, Acc(Parz): 90.6

Ref: Reference, Obj: Objective, Lang: Language, PS: Page Segmentation, LA: Layout Analysis, MultL: Multilingual, TBD: Text Block Detection, TLS: Text Line Detection, TS: Text Segmentation, DS: Drawing Segmentation, HN: Historical Newspaper, HPD: Historical Printed Document, HHD: Historical Handwritten Document, MLT: Multiple Tasks (*page extraction*, *baseline detection*, *document segmentation*), ML: Machine Learning, DL: Deep Learning, SM: Statistical Model, CC: Connected Components, HMM: Hidden Markov Model, GW: George Washington, SG: Saint Gall, Parz: Parzival, JDG: Journal de Geneve, GDL: Gazette de Lausanne, JDR: Journal de Rouen, Blc: Balsac, Him: Himanis, KDK: Kabinet der Koningin, ANN: ANNAHAR, ALH: ALHAYAT, P: Precision, R: Recall, F: F-score, Acc: Accuracy, Rec: Recognition, AvgDER: Average Diarization Error Rate, AvgCS: Average Completeness Score, Cov: Covariance, mIoU: mean Intersection over Union, mAP: mean Average Precision, px: Pixels, rg: Region, bk: Block, fr: Frame, img: image, tl: Text line, tb: Text block.

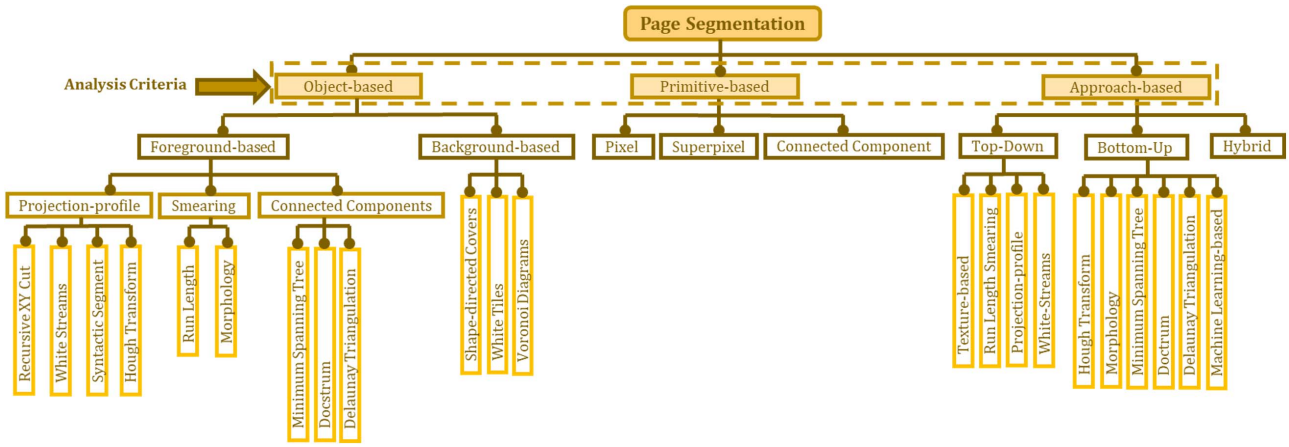

Fig. 7. Page segmentation methods and analysis criteria for historical documents.

Chen et al. [3] adopted an unsupervised feature learning approach for page segmentation in historical documents. The solution was crafted on a super-pixel approach generated using SLIC and a stacked convolutional auto-encoder convolutional model is trained. After the classification of super-pixels using an SVM-based classifier, a smoothing procedure based on the connected component approach was employed to refine the segmentation results. Similarly, Wick and Puppe [120] considered a fully convolutional network (FCN) consisting of convolutional–deconvolution layers for page segmentation of

historical document images. The model is learned directly from raw pixels and is designed to process one complete page per step without any preprocessing required.

Liebl and Burghardt [121] conducted a study to evaluate different deep neural network (DNN)-based page segmentation architectures for historical newspapers. The study aimed to assess the performance of these architectures while considering factors such as the size of training data, the number of labels, and different settings of tiling (the practice of dividing an image into smaller, overlapping or nonoverlapping regions (tiles) for analysis) and scaling (involves adjusting the size of the image or its components) for separating text, tables, and column lines in newspapers. They concluded that pixel-by-pixel segmentation approaches apply to newspapers and they preferred vertical tiling over other methods. However, detecting and annotating components in different page layouts of historical newspapers is a challenging and crucial task that requires statistical and deep learning methodologies based on pixel and super-pixel labeling. While there are feature-based approaches for digitized paper documents, the poor physical state of historical documents makes the task more challenging.

In summary, the optimal selection of a page segmentation technique relies on aligning algorithmic choices with the specific needs of HDP. Factors such as precision, computational efficiency, and the availability of training data play crucial roles in selection of the approach. Traditional methods such as the two-step page segmentation and optimization offer simplicity and learning-free advantages. Feature-based strategies, such as pixel-labeling approach, seamlessly integrate diverse elements by incorporating color, texture features, and unsupervised learning. Deep learning-based methods, including CAE, CNNs, and the convolutional encoder–decoder, excel in automatic feature extraction, while the fully convolutional neural network efficiently handles full pages without preprocessing. Complementing these approaches, the bottom–up greedy approach leverages hierarchical text segmentation, while the DNN-based page segmentation provides insightful perspectives, especially in historical newspaper segmentation. Table V presents an overview of the existing techniques of page segmentation in historical documents.

2) Text Line Segmentation: Detecting text lines is a critical preprocessing step in analyzing document structures and recognizing printed or handwritten text. While there are many techniques and approaches available for text line detection (see Fig. 8) in modern documents, there are few studies on ancient handwritten and printed historical document images [20], [122]. The purpose of text line detection is to divide the pixels of an image $I \in [0, 255]_{i \times j}$ into m subsets, which represent text lines, and one subset representing clutter pixels.

Zahour et al. [123] proposed a text line segmentation algorithm that utilized the *K*-means classifier and text block detection approach. This method was effective for both printed and handwritten Arabic documents; however, it was not accurate for historical documents with varying font types and sizes due to some parameters being dependent on text font size. On the other hand, Boussellaa et al. [124] proposed a two-step approach for text line detection in Arabic historical

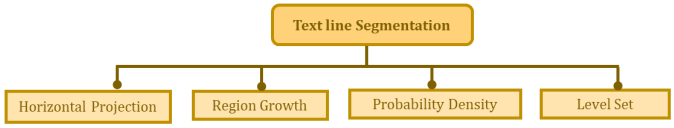

Fig. 8. Text line segmentation methods for historical documents.

handwritten script images that have multitouching, cursive, and overlapping text. For text line extraction, the approach uses projection profile-based iterative statistical block analysis followed by a fuzzy *C*-means clustering algorithm for text baseline detection. Additionally, Garz et al. [125] suggested an interest-point clustering-based approach for text line detection in historical documents. The method transforms document images into a sparse interest points domain, followed by word clustering using the DBSCAN algorithm, to achieve binarization-free line segmentation.

Jindal and Lehal [126] suggested a text block segmentation-based approach for text line detection. The results of this approach showed good accuracy for historical handwritten Gurumukhi scripts, in terms of line segmentation, but the incorrect zoning of characters remained a challenge. Grüning et al. [127] proposed a super-pixels-based binarization-free text line extraction method for both printed and handwritten historical document images. They used a clustering algorithm based on a greedy approach that takes advantage of the general characteristics of text lines without any predefined parametric information. Furthermore, Mechi et al. [128] proposed a variant of FCN, a deep-learning model for text line segmentation. Their method is based on an adaptive U-Net architecture that can handle variable image sizes and requires less training data.

In summary, the analysis of text line segmentation techniques in HDP reveals a diverse array of methodologies. Cluster-based approaches, as demonstrated by methods such as the *K*-means classifier and text block detection, effectively employ clustering for segmentation, catering to specific document types. The iterative statistical block analysis strategy with fuzzy *C*-means clustering proves particularly advantageous for specific script images. Notably, the text block segmentation-based approach showcases high accuracy for specific scripts. However, super-pixel-based methods, exemplified by the binarization-free method, offer adaptability but may encounter challenges in handling complex layouts. Moreover, the introduction of the adaptive U-Net with FCN introduces a notable deep learning-based avenue, showcasing flexibility in handling variable image sizes with reduced training data requirements. Table VI summarizes the existing studies on text line detection in historical documents.

3) Character Segmentation: Recently, there have been several deep learning-based models and methods proposed in the literature to address the problem of character segmentation and recognition in historical documents [129]. Table VII provides an overview of the existing literature on character segmentation in historical documents. Gao et al. [130] viewed character segmentation and recognition as a sequence labeling task for historical documents and proposed a deep learning-based approach

TABLE V
TECHNIQUES USED FOR PAGE SEGMENTATION

Ref	Technique	Description	Pros	Limitations	Results (%)
[113]	Two-Step Page Segmentation (Super-pixel Representation)	A two-step process involving page segmentation and block classification using super-pixel representation and spatial/color features.	Reduces image-processing complexity using super-pixel representation. Refines nontextual regions into noise, drawing, and background.	Performance may depend on the quality of the initial text block separation.	P: 99.8 R: 89.6 F: 92.92
[114]	Optimization-based Text Region Detection (Graph Cuts)	Formulates text region detection as an optimization task to minimize energy using graph cuts.	Energy-based approach for text region detection. Learning-free method.	May require parameter tuning for different document layouts.	F: 98.84
[115]	Page Segmentation for Handwritten Historical Documents (Color and Texture Features)	Utilizes various color and texture features for pixel classification in page segmentation of handwritten historical documents.	Considers color and texture features for segmentation. Applies feature selection using a Fast Correlation-based Filter.	Performance may vary depending on the complexity of the document layouts.	Acc: 97.9
[116]	Page Segmentation (Deep Learning Based)	Page segmentation of colored handwritten historical document images using a CAE architecture and SVM classifier.	Utilizes deep learning for feature extraction. Learns low and high-level features automatically.	Requires a large amount of training data for deep learning. Can be computationally intensive during training.	Acc: 97.66
[3]	Pixel-Labeling Approach (Unsupervised Learning)	Pixel-labeling approach incorporating unsupervised learning of stacked auto-encoders convolutional network for feature learning and conditional random field (CRF) modeling.	Provides quick and precise page segmentation results. Incorporates unsupervised learning for feature extraction.	Performance may vary depending on the quality of super-pixel generation and classification.- May require fine-tuning for different types of historical documents.	Acc: 95.7
[117]	CNN (Super-pixels)	Pixel labeling approach using a CNN with a single convolutional layer trained on labeled super-pixels.	Straightforward solution for page segmentation of handwritten historical scripts. Utilizes CNNs for pixel labeling.	Performance may vary depending on the complexity of the scripts and the quality of super-pixel generation. May not generalize well to different types of historical documents.	Acc: 91
[118]	Convolutional Encoder–Decoder (CED)	Neural network-based architecture using CED to segment handwritten text into distinct text zones.	Takes advantage of prior information by incorporating spatial and texture information. Demonstrates generality through experiments on different historical handwritten datasets.	Performance may depend on the quality and size of the training datasets. The complexity of the model may require significant computational resources.	Acc: 94
[119]	Bottom–up Greedy Approach	Hierarchical bottom–up approach for text segmentation in historical documents using HTR.	Supports textual criticism and cultural studies through hierarchical text segmentation. Leverages HTR techniques for text analysis.	Performance may depend on the accuracy of the underlying HTR system. May require additional preprocessing steps for nontextual elements.	Acc: 69.13
[120]	Fully FCN	Fully FCN consisting of convolutional–deconvolution layers for page segmentation of historical document images. Trained directly from raw pixels without preprocessing.	Enables page segmentation without preprocessing steps. Handles full pages in one step.	Performance may vary depending on the complexity and variability of historical document layouts.	Acc: 98.4
[121]	DNN-based Page Segmentation	Evaluation of DNN-based page segmentation architectures for historical newspapers.	Provides insights into the performance of DNN-based approaches for historical newspaper segmentation. Considers factors such as training data size and tiling/scaling settings.	Specific to historical newspaper segmentation. Performance may vary depending on the quality and variability of historical newspaper layouts.	Acc: 98.53

Ref: Reference, P: Precision, R: Recall, F: F-score, Acc: Accuracy.

TABLE VI
TECHNIQUES USED FOR TEXT LINE SEGMENTATION

Ref	Technique	Description	Pros	Limitations	Results (%)
[123]	*K*-means Classifier and Text Block Detection	Utilizes the *K*-means classifier and text block detection approach for text line segmentation in printed and handwritten Arabic documents.	Effective for printed and handwritten Arabic documents.	Accuracy may be affected by varying font types and sizes in historical documents.	Acc: 96
[124]	Iterative Statistical Block Analysis and Fuzzy *C*-means Clustering	Two-step approach for text line detection in Arabic historical handwritten script images. Uses iterative statistical block analysis and fuzzy *C*-means clustering for text baseline detection.	Suitable for Arabic historical handwritten script images with multitouching, cursive, and overlapping text.	Parameter dependency on text font size.	Acc: 95.1
[125]	Interest-Point Clustering	Interest-point clustering-based approach for text line detection in historical documents. Transforms document images into a sparse interest points domain and performs word clustering using the DBSCAN algorithm.	Binarization-free line segmentation.	Performance may vary depending on the document characteristics and density of interest points.	Acc: 97.97
[126]	Text Block Segmentation-based Approach	A text block segmentation-based approach for text line detection, specifically for historical handwritten Gurumukhi scripts.	Good accuracy for line segmentation in historical handwritten Gurumukhi scripts.	Challenges with incorrect zoning of characters.	Acc: 98.9
[127]	Super-pixels-based Binarization-free Method	A method based on super-pixels for binarization-free text line extraction in printed and handwritten historical document images.	Does not require binarization, making it applicable to documents with varying backgrounds and contrasts. General characteristics-based approach without predefined parametric information.	May have limitations with complex layouts and overlapping lines.	F: 98.24
[128]	Adaptive U-Net with FCN	Variant of FCN for text line segmentation, based on an adaptive U-Net architecture.	Can handle variable image sizes. Requires less training data.	Performance may vary depending on the dataset and text line complexities.	P: 75 R: 85 F: 79

Ref: Reference, P: Precision, R: Recall, F: F-score, Acc: Accuracy.

that combines a CNN and an LSTM network for feature extraction and sequence modeling respectively. This method further utilizes a CRF for labeling the sequence. In another work, Tang et al. [10] introduced an end-to-end anchorless approach called HRCenterNet for segmenting and detecting Chinese characters in historical documents. For this method, a joint approach was employed that combines high-resolution network (HRNet) proposed by [131] with the learning method of CenterNet, which is a center point-based object detector approach [132]. In an improved version of HRCenterNet for character segmentation and recognition, Tang et al. [11] proposed the HRRegionNet model, which adds an additional transposed convolutional module to the previously proposed HRCenterNet [10] to restore the output to a high resolution. The method also incorporates multiresolution aggregation to combine multiresolution features, resulting in sophisticated character segmentation results on Chinese historical document images.

Ma et al. [133] employed Fast R-CNN architecture for Chinese character recognition in historical documents, while Sihang et al. [134] proposed a deep reinforcement learning-based model that treats character detection as a Markov Decision Process. The latter approach used a FCN and implemented a variant of Q-learning to adjust bounding boxes for precise recognition of Chinese characters in historical text images. Similarly, Yang et al. [135] presented a CNN-based recognition guide detector (RGD) method for Chinese character recognition that utilized two recognition-guided proposal networks (RGPNs) to extract contextual information and accurately identify the characters. Zhang et al. [136] suggested a framework for character recognition in Uchen Tibetan Historical scripts, which involved text line shortening and character block database establishment using projection and syllable points-based methods. The proposed framework then performed character detection and segmentation using local baseline detection leveraging key feature information, followed by stroke attribution.

In summary, character segmentation techniques for HDP encompass a diverse range of approaches, each with its unique advantages. Deep learning models, such as the CNN-LSTM-CRF approach, offer a robust solution for general character segmentation and recognition. Anchorless methodologies, demonstrated by HRCenterNet and HRRegionNet, prove effective for end-to-end segmentation and detection, especially in Chinese historical documents. Models such as Fast R-CNN and deep reinforcement learning excel in precise character recognition. The RGD, utilizing CNN-based recognition, ensures accurate character identification. Nonetheless, these approaches may face challenges with intricate character variations and layouts. Tailored for Uchen Tibetan historical scripts, the projection and syllable points-based framework provides an effective, albeit language-specific solution. Fig. 9 presents examples of page, text line, and character segmentations in printed and handwritten historical documents.

TABLE VII
TECHNIQUES USED FOR CHARACTER SEGMENTATION

Ref	Technique	Description	Pros	Limitations	Results (%)
[130]	CNN-LSTM-CRF Approach	Deep learning-based approach combining a CNN and an LSTM network for feature extraction and sequence modeling, along with CRF for labeling the sequence.	Effective for character segmentation and recognition in historical documents.	Performance may vary depending on dataset characteristics and document complexities.	P: 93 R: 92 F: 93.6
[10]	HRCenterNet	Anchorless approach combining HRNet with the learning method of CenterNet for character segmentation and detection in Chinese historical documents.	End-to-end approach for segmenting and detecting Chinese characters.	May have limitations with complex character variations and layouts.	Acc: 81
[11]	HRRegionNet	Improved version of HRCenterNet with an additional transposed convolutional module and multiresolution aggregation for sophisticated character segmentation results.	Enhanced character segmentation results on Chinese historical document images.	Performance may be influenced by dataset characteristics and image resolutions.	Acc: 86.2
[133]	Fast R-CNN	Employed Fast R-CNN architecture for Chinese character recognition in historical documents.	Effective for Chinese character recognition.	Character segmentation performance may depend on the complexity of the document layout.	Acc: 94.12
[134]	Deep Reinforcement Learning	Deep reinforcement learning-based model treating character detection as a Markov Decision Process. Utilized a FCN and a variant of Q-learning for precise recognition of Chinese characters in historical text images.	Precise recognition of Chinese characters in historical text images.	Performance may vary based on training data and document complexities.	P: 90.15 R: 86.18
[135]	RGD	CNN-based method utilizing Recognition Guided Proposal Networks (RGPN) to extract contextual information and accurately identify Chinese characters.	Accurate identification of Chinese characters.	May have limitations with complex character variations and layouts.	P: 98.58 R: 96.96 F: 97.76
[136]	Projection and Syllable Points-based Framework	Framework for character recognition in Uchen Tibetan Historical scripts involving text line shortening, character block database establishment, character detection and segmentation using local baseline detection leveraging key features information, and stroke attribution.	Effective framework for character recognition in Uchen Tibetan Historical scripts.	Specific to Uchen Tibetan Historical scripts and may not be applicable to other languages/scripts.	P: 91.19 R: 94.1 F: 92.62

Ref: Reference, P: Precision, R: Recall, F: F-score, Acc: Accuracy.

D. Historical Document Semantic Analysis

Document semantic analysis aims at the logical labeling of the detected regions to its document elements (such as title, headings, paragraphs, and figures) termed *semantic labeling*. The prime objective of *document semantic analysis* is to assign logical labels to the identified regions in the historical documents, such as headings, paragraphs, figures, titles, and so on, which is also known as *semantic labeling* [137]. Various methods for semantic segmentation have been proposed, including *rule-mining*, *syntactic*, *perceptual*, *learning*, and *knowledge-based* approaches (as illustrated in Fig. 10).

Recently, visual and textual feature-based methods have been introduced for logical segmentation in historical documents [138]. Barman et al. [139] proposed a multimodal neural model for the semantic segmentation of historical newspapers and demonstrated that integrating pixel-level visual features with text embedding can result in a robust visual baseline approach. Palfray et al. [140] presented an article extraction method based on logical segmentation in digitized old newspapers, which combines pixel classification using CRFs and a set of predefined rules for article segregation. Gutehrlé and Atanassova [111] addressed the problem of the lack of logical structure and presented a rule-based layout segmentation method for historical documents, comparing its performance with machine learning-based RIPPER and Gradient Boosting models. Table VIII presents related semantic analysis-based studies on historical document analysis with their advantages and limitations.

One of the key tasks that affect the logical mapping of the segmented regions is *article separation*. The primary objective of article extraction and separation is to identify semantically related articles based on their content. For instance, in newspapers, articles are typically arranged in multicolumns, and text content may be interrupted by ads. The aim is to maintain the order of content that is suitable for human reading. Article separation finds applications in media monitoring.

In [19], the early offline document segmentation techniques for textual documents are examined. However, currently, there is growing interest in machine learning techniques, particularly deep learning, for the semantic segmentation of historical documents. Meier et al. [6] presented a visual feature-based semantic segmentation approach that employs a FCN to divide pages into articles that are semantically related. The goal is to process complex queries about a single article to retrieve information. The authors claim that their method is twice as efficient as a commercially available deep learning-based solution and outperforms it. Additionally, Oliveira et al. [7] introduced a dhSegment, which can handle pixelwise segmentation tasks on historical documents, including page and border extraction and layout analysis. The proposed approach incorporates a pixel-based trained model and task-dependent blocks for multiple

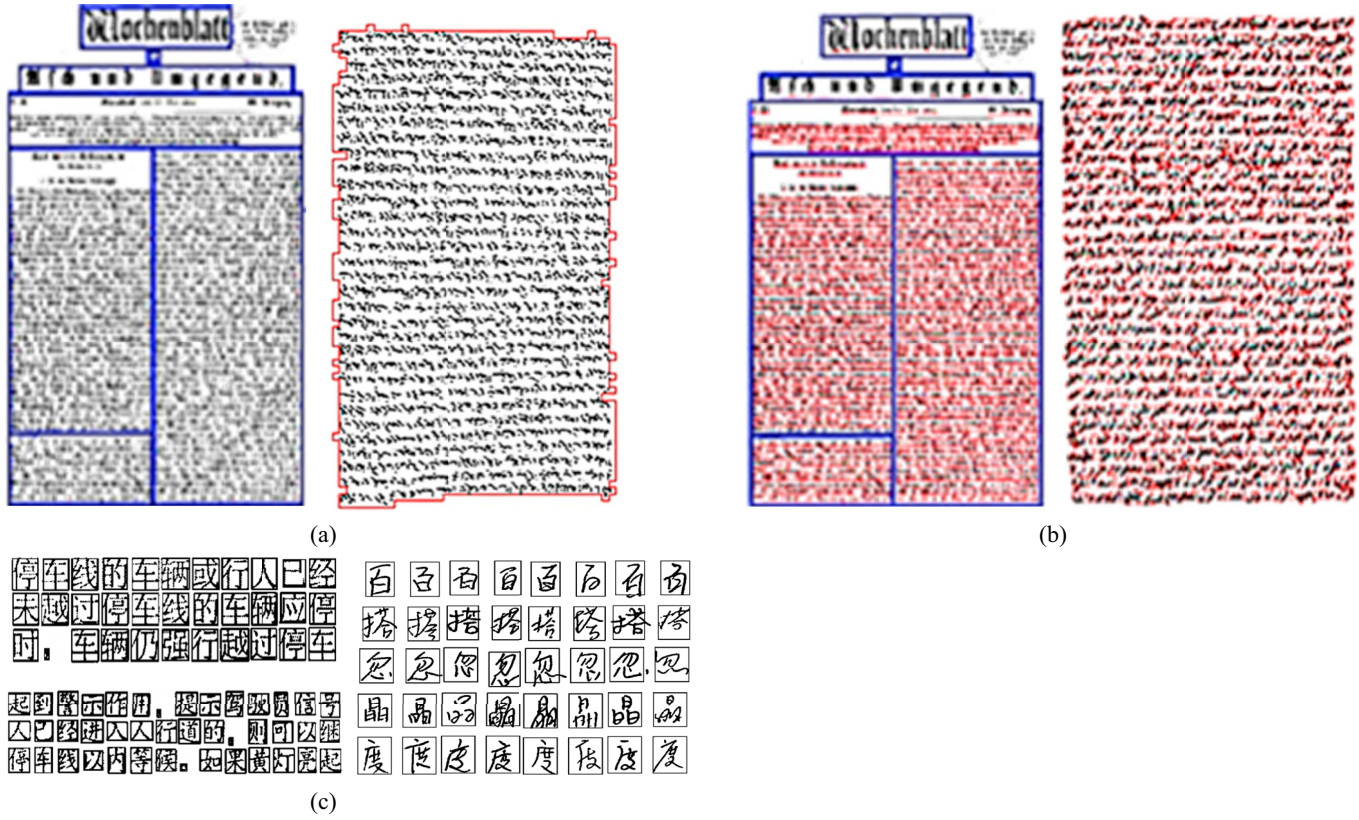

Fig. 9. Examples of (a) page [28], [29], (b) text line [28], [29], and (c) character [30], [31] segmentations in printed and handwritten historical documents.

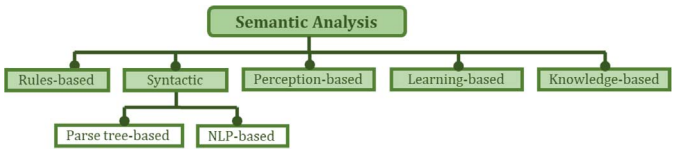

Fig. 10. Semantic analysis methods for historical documents.

segmentation activities. The CNN-based architecture retrains a single model for a specific segmentation task. However, the model struggles to learn various segmentation tasks simultaneously. Fig. 11 showcases an example of an article separation task in the historical newspaper document image.

In their work, Zhu et al. [141] developed a framework for layout segmentation of historical articles using multiple image segmentation models. They modified the FCN model proposed by [6] by adding two label settings to the last layer of the model. They also developed a model called SepDet based on FCN to distinguish between text and background regions within a document and partition them into semantic series blocks. To predict object masks, they used Mask R-CNN, which is an extended version of Fast R-CNN that has an additional branch for object prediction. This approach is more convenient because it eliminates the need to obtain the bounding box by the OCR engine. Finally, they implemented a semantic segmentation system called SETR [142], which treats the problem as a sequence-to-sequence prediction task. SETR encodes an image as a sequence of patches using a Transformer encoder that is then combined with a decoder to produce a segmentation model. This approach overcomes the limitations of the FCN model by down-sampling the spatial resolution of the input resulting in a lower resolution of mapped features which are useful for discriminating semantic classes. The proposed framework is also capable of end-to-end OCR after applying some postprocessing steps.

Despite the advancements in techniques and methodologies for digitizing historical documents over the past decade, there is still a need for more research to overcome the remaining challenges. These challenges include semantic analysis of historical documents, where the task of article separation needs more attention to improve the experience of information search and retrieval from historical repositories [143]. Moreover, labeling historical documents for page and text line segmentation, as well as semantic information retrieval, presents formidable challenges owing to the intricate nature of historical texts. Recent advancements in deep learning techniques offer promising avenues to mitigate manual labeling efforts [25]. However, addressing variability, degradation, and semantic ambiguity remains a pressing research area for automating historical document labeling. Understanding these challenges is crucial for the effective application of these techniques in this domain. For instance, weakly supervised learning methods struggle to accurately find settlement features on historical maps due to variations in map quality and symbol representation [144].

TABLE VIII
TECHNIQUES USED FOR HISTORICAL DOCUMENT SEMANTIC ANALYSIS

Ref	**Technique**	**Description**	**Pros**	**Limitations**	**Results (%)**
[138]	Visual and Textual Feature-based Methods	Methods utilizing visual and textual features for logical segmentation in historical documents.	Integration of visual and textual features for robust segmentation.	Performance may vary based on document types and complexities.	Acc: 74
[139]	Multimodal Neural Model	Integration of pixel-level visual features with text embedding for robust visual baseline approach in semantic segmentation of historical newspapers.	Robust semantic segmentation of historical newspapers.	Performance may vary depending on document complexities and the quality of visual features.	Acc: 70.8
[140]	Pixel Classification and Rule-based Method	Article extraction method based on logical segmentation in digitized old newspapers. Combines pixel classification using CRFs and a set of predefined rules for article segregation.	Effective article extraction and segregation in digitized old newspapers.	Performance may depend on the accuracy of pixel classification and predefined rules.	Acc: 85.84
[111]	Rule-based Layout Segmentation	Rule-based layout segmentation method for historical documents, comparing its performance with machine learning-based models.	Effective layout segmentation of historical documents.	May have limitations with complex document layouts and variations.	P: 96.9 R: 99.1 F: 97.9
[6]	FCN-based Semantic Segmentation	Visual feature-based semantic segmentation approach employing a FCN to divide pages into semantically related articles.	Efficient semantic segmentation of historical documents.	Performance may depend on dataset characteristics and the complexity of semantic queries.	DER: 2.36
[7]	dhSegment Framework	Deep learning-based document segmentation framework (dhSegment) handling pixelwise segmentation tasks on historical documents, including page and border extraction, and layout analysis.	Capable of handling various segmentation tasks on historical documents.	Model may struggle to learn multiple segmentation tasks simultaneously.	P: 96 R: 92 F: 94
[141]	FCN, SepDet, Mask R-CNN, SETR	Framework for layout segmentation of historical articles using multiple image segmentation models, including modifications to FCN, Separator Detector (SepDet), Mask R-CNN, and implementation of SEgmentation TRansformer (SETR).	Improved semantic segmentation accuracy with multiple segmentation models.	Performance may depend on the complexity and variability of historical document layouts.	P: 90.68 R: 86.78 Acc: 84.92

Ref: Reference, P: Precision, R: Recall, F: F-score, Acc: Accuracy, DER: Diarization Error Rate.

Similarly, self-supervised learning approaches for pretraining neural networks find difficulties in adapting to the diverse visual characteristics and languages inherent in historical documents [145]. Furthermore, the task of transcription in low-resource settings remains challenging despite significant enhancements in recognition accuracy achieved through self-supervised pretraining. Even in keyword spotting tasks, where self-supervised learning offers a potential solution to data scarcity, ensuring robustness and generalizability across diverse historical document collections poses ongoing challenges [146]. Table IX provides an overview of the current state of the art in historical document semantic analysis.

E. Historical Document Panoptic Analysis

To improve the performance of HDP beyond the limitations of *syntactic* and *semantic*-based analysis, researchers have turned their attention to *panoptic*-based analysis. The *panoptic segmentation* technique, which combines *semantic* and *instance* segmentation into a single network architecture, is a popular computer vision method for object detection. *Semantic segmentation* identifies object categories in an image, while *instance segmentation* separates individual instances of each semantic category [149], [150].

Cao et al. [151] introduced DPS, a document panoptic segmentation framework for recognizing layouts in digital

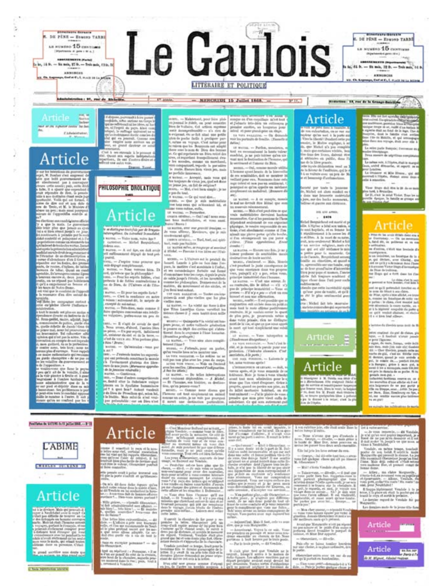

Fig. 11. Article segmentation in historical document (https://www.newseye.eu/).

documents. Their approach partitions the page into column regions and then groups token into line regions, assigning a class label and instance id to each line region. The authors also proposed a new evaluation metric, token-IoU, which provides

TABLE IX
HISTORICAL DOCUMENT SEMANTIC LAYOUT ANALYSIS

Ref	DT	Obj	Method	Approach	Dataset	Lang	Results(%)
[147]	HN	LLA	RLSA	Visual	100+ pages	Arabic	Rec: 91.90
[148]	HN	LLA	Rule based	Visual	54 documents	French	P: 72.6 R: 29.8 F: 40.6
[140]	HN	AS	SM	Visual	JDR:42 images	French	Acc: 85.84
[139]	HN	LS	DL	Visual+Text	2200 pages	MultL	P: 89.43 R: 92.39 mIoU: 82
[7]	HPD	MLT	DL	Visual	885 pages	MultL	IoU: 97.4
[138]	HHD	AcS	DL+HMM	Visual+Text	Blc: 911 pages, Him: 562 pages	English, French	mAP: 76

Ref: Reference, Obj: Objective, Lang: Language, LA: Layout Analysis, LS: Logical Segmentation, DT: Document Type, HN: Historical Newspaper, HPD: Historical Printed Document, HHD: Historical Handwritten Document, LLA: Logical Layout Analysis, CC: Connected Components, AS: Article Separation, AcS: Act Segmentation, DL: Deep Learning, SM: Statistical Model, MLT: Multiple Tasks (*page extraction*, *baseline detection*, *document segmentation*), HMM: Hidden Markov Model, RLSA: Run Length Smoothing Algorithm, JDR: Journal de Rouen, Blc: Balsac, Him: Himanis, P: Precision, R: Recall, F: F-score, mIoU: mean Intersection over Union, Acc: Accuracy, Rec: Recognition, mAP: mean Average Precision, MultL: Multilingual.

TABLE X
TECHNIQUES USED IN HISTORICAL DOCUMENT PANOPTIC ANALYSIS

Ref	Technique	Description	Pros	Limitations	Results (%)
[151]	DPS (Document Panoptic Segmentation)	A framework for recognizing layouts in digital documents by partitioning the page into column regions and grouping tokens into line regions. Assign class labels and instance IDs to each line region. Introduces the token-IoU evaluation metric.	Outperforms Faster R-CNN and Mask R-CNN methods on multiple datasets in terms of token-IoU scores.	Performance may vary depending on the complexity and variability of document layouts.	mAP: 92.05
[152]	Panoptic-DLA	A proposal-free panoptic segmentation method for historical newspaper layout analysis. Includes branches for semantic and instance segmentation, grouping predicted "things" into their respective instance IDs. Evaluated on a dataset of historical European newspapers with competitive performance.	Proposal-free approach reduces complexity and computational cost.	Applicability may be limited to specific types of historical newspaper layouts.	P: 99.75 R: 99.54 F: 99.64

Ref: Reference, mAP: mean Average Precision, P: Precision, R: Recall, F: F-score.

pinpoint accuracy for measuring intersection over the union between tokens in predicted and ground-truth objects. They concluded that DPS outperformed Faster R-CNN and Mask R-CNN methods on three datasets, *PubLayNet*, *ArXiv*, and *Financial* datasets in terms of token-IoU scores. Lu et al. [152] presented a proposal-free panoptic segmentation method, called *Panoptic-DLA*, for historical newspaper layout analysis. The method includes two branches for semantic and instance segmentation, which group the predicted "things" into their respective instance IDs. The model was evaluated on a dataset of historical European newspapers and demonstrated to have competitive performance. Table X provides an overview of the studies based on panoptic analysis on DLA.

F. ATR

One way to convert paper-based documents into digital format is by scanning the image, but this form is not useful for searching and information retrieval. Therefore, additional processing is required to convert these scanned images into a machine-readable format, such as alphanumeric code, so that the computer can recognize the text and facilitate the searching and information mining process. However, text extraction is a common problem regardless of whether the document is printed or handwritten. To address this problem, ATR is one of the prime solutions that is intensively researched which allows reforming a pictorial document into an electronic-text document. There are primarily two solutions to perform ATR: HOCR and HTR. The fundamental difference between HOCR and HTR is that the efficiency of HOCR relies on the predictable accurate classification of characters based on the uniformity of spacing among characters and words. On the other hand, this is not the case with HTR given the human handwriting idiosyncrasies and peculiarities, thus making character-level segmentation-based methods unfit for HTR.

1) HOCR: The history of OCR can be traced back to the retina scanner invented by Carey in 1870. The first data processing OCR (first generation) for business applications emerged in 1986 after the advent of digital computers. These early OCRs could recognize only a limited number of specially designed

TABLE XI
SUMMARY OF VARIOUS PROPOSED SYSTEMS FOR LAYOUT ANALYSIS IN ICDAR COMPETITIONS

Ref	Method(s)	Succ_Rate (%)		F (%)	R (%)	IoU_LA (%)	IoU_LS (%)	Dataset	Eval (#img)
		Seg	OCR						
[162]	Besus	-	-	-	35.85	-	-	PRimA	32
	Tsinghua1	-	-	-	67.59	-	-	PRimA	32
	Tsinghua2	-	-	-	35.70	-	-	PRimA	32
[163]	DICE	-	-	90.09	39.12	-	-	PRimA	32
	Fraunhofer'09	-	-	93.14	76.45	-	-	PRimA	32
	REGIM-ENIS	-	-	87.82	16.08	-	-	PRimA	32
	Tesseract	-	-	91.04	68.44	-	-	PRimA	32
	FineReader	-	-	91.90	54.90	-	-	PRimA	32
	OCRopus	-	-	78.35	32.98	-	-	PRimA	32
[164]	FR9	80.4	78.4	-	-	-	-	PRimA	100
	MEPhI	63.8	62.7	-	-	-	-	PRimA	100
	Jouve'11	85.2	83.3	-	-	-	-	PRimA	100
	Fraunhofer'11	79.8	78.6	-	-	-	-	PRimA	100
	EPITA'11	82.6	81.6	-	-	-	-	PRimA	100
[165]	Tesseract 3	58.3	53.3	-	-	-	-	PRimA	50
	FRE10	83.7	80.3	-	-	-	-	PRimA	50
	EPITA	83.2	81.4	-	-	-	-	PRimA	50
	Jouve'13	85.6	82.3	-	-	-	-	PRimA	50
	PAL	86.4	85.2	-	-	-	-	PRimA	50
	Fraunhofer'11	85.5	82.4	-	-	-	-	PRimA	50
	Fraunhofer'13	86.9	85.2	-	-	-	-	PRimA	50
[166]	Fraunhofer'13	84.1	79.9	-	-	-	-	PRimA	70
	ISPL	83.1	82.0	-	-	-	-	PRimA	70
	MHS	90.5	88.3	-	-	-	-	PRimA	70
	PAL	83.2	79.5	-	-	-	-	PRimA	70
	FRE10	71.6	68.5	-	-	-	-	PRimA	70
	FRE11	72.0	69.3	-	-	-	-	PRimA	70
	Tesseract 3.02	74.1	70.8	-	-	-	-	PRimA	70
	Tesseract 3.03	74.3	70.0	-	-	-	-	PRimA	70
[167]	KFUPM	-	-	-	-	65.36	-	DIVA-HisDB	150
	BYU	-	-	97.68	-	90.18	81.50	DIVA-HisDB	150
	Demokritos	-	-	-	-	92.27	-	DIVA-HisDB	150
	MindGarage 4.1	-	-	-	-	93.95	-	DIVA-HisDB	150
	MindGarage 4.2	-	-	-	-	89.98	-	DIVA-HisDB	150
	NLPR	-	-	-	-	94.90	-	DIVA-HisDB	150
	Fraunhofer'17	-	-	-	-	74.07	39.17	DIVA-HisDB	150
	CVML	-	-	93.40	-	-	-	DIVA-HisDB	150
	CitLab	-	-	98.22	-	-	96.99	DIVA-HisDB	150
	DIVA+MindGarage 4.1	-	-	-	-	-	97.86	DIVA-HisDB	150

Ref: Reference, img: images, Succ_Rate: Success Rate, Seg: Segmentation, F: F-score, R: Recall, IoU: Intersection over Union, LA: Layout Analysis, LS: Line Segmentation, Eval: Evaluation.

characters. However, modern OCRs (fourth generation) have come a long way since then and can now recognize complex artifacts such as printed/handwritten text, images, tables, symbols, characters, and more [153]. Historical documents, in particular, pose a significant challenge due to their abundance of font designs, variability in typography, and the limited precision of early typing [154]. Traditional OCR methods, designed for modern printed text, may struggle to accurately interpret the intricate details of historical manuscripts. Overcoming these challenges is crucial for successful digitization and subsequent information extraction. HOCR techniques focus on adapting OCR algorithms to address the peculiarities of historical documents. By incorporating features such as adaptive thresholding, layout analysis, and historical font recognition, HOCR enhances the accuracy of text extraction from aged manuscripts. This facilitates the creation of digital repositories that preserve historical content while enabling advanced search and retrieval capabilities. To address these challenges, machine learning and deep learning-based approaches have proliferated in the last decade. These techniques have substantially improved the accuracy and robustness of OCR systems when dealing with historical documents. In the realm of HOCR, researchers and developers have introduced innovative tools and applications. Notable among these are *ABBYY FineReader*, *Tesseract*, *Kraken*, and *Alethia*. These applications utilize advanced algorithms and models to enhance the recognition of characters in historical documents, offering improved accuracy and efficiency.

2) HTR: Although HOCR tools and applications have made some progress in recognizing machine-printed historical documents, there are still numerous challenges to overcome, particularly when it comes to digitizing historical handwritten text document images. Due to the high variability in handwriting fonts, styles, and types of historical documents compared to machine-printed documents, they present a variety of inherent challenges [22] detailed earlier. Various machine learning and deep learning-based methods have been developed to address these challenges, leading to a proliferation

TABLE XII
ATR TOOLS AND SOFTWARE SYSTEMS FOR HISTORICAL DOCUMENTS PROCESSING

System	ATR	Resource	Link	Reference
OCRopus/OCRopy	HOCR	Tool	https://github.com/ocropus/ocropy	[13], [14], [168], [169]
OCRoRact	HOCR	Tool	-	[170], [171]
anyOCR	HOCR	Tool	https://github.com/syedsaqibbukhari/docanalysis	[171], [172]
Ocular	HOCR	Tool	https://github.com/tberg12/ocular	[173], [174]
ABBYY FineReader	HOCR	Software	https://www.abbyy.com	[12], [174], [175], [176]
Aletheia	HOCR	Software	https://www.primaresearch.org/tools/Aletheia, https://github.com/daniellerch/aletheia	[177], [178], [179], [180]
Tesseract	HOCR	Software	https://github.com/tesseract-ocr/tesseract	[174], [181], [182], [183], [184]
Kraken	HOCR	Package	https://kraken.re/master/index.html, https://github.com/mittagessen/kraken	[28], [185], [186], [187]
FromThePage	HTR	Software	https://fromthepage.com/, https://github.com/benwbrum/fromthepage	[16], [188], [189], [190]
Transkribus	HTR	Platform	https://readcoop.eu/transkribus/, https://github.com/transkribus/	[191], [192], [193], [194], [195]
eScriptorium	HTR	Platform	https://escriptorium.fr/, https://github.com/UB-Mannheim/escriptorium	[9], [185], [196]
iDocChip	HOCR	Platform	-	[197], [198]

of the literature based on logistic regression, support vector machines (SVMs), hidden Markov models (HMMs), and other approaches. Deep-learning-based models, including long short-term memory (LSTM) models, recurrent neural networks (RNNs) [155], and encoder–decoder networks with an attention mechanism, have been proposed to recognize text in historical documents [156], [157], [158]. Furthermore, transformer-based solutions have been proposed to recognize handwritten historical text [17], [159], [160]. Generative adversarial networks (GANs) [5], [161]-based techniques are gaining popularity for generating synthetic handwritten text images. Additionally, user-friendly graphical platforms have been developed for automatic historic text recognition, such as eScriptorium, FromThePage, and Transkribus. Table XI summarizes the results of recognition systems presented at various competitions held under the umbrella of the eminent *ICDAR* conference series [162], [163], [164], [165], [166], [167]. Table XII spotlights various ATR (HOCR and HTR) tools used in historical document analysis. Furthermore, there are numerous efforts that have been made by researchers to introduce diverse recognition systems for the analysis of document layouts.

IV. DATASETS AND EVALUATION METRICS

This section provides an overview of various datasets and the evaluation metrics adopted for the performance evaluation of HDP research.

A. Datasets

There are various databases offering diverse datasets for historical document analysis, encompassing a range of languages, layouts (such as *Manhattan*, *non-Manhattan*, *multicolumn*, *overlapped*, and *arbitrary*), and document types (including *books*, *journals*, *articles*, *letters*, *memos*, and *newspapers*). These datasets are typically categorized into two main groups: Handwritten datasets (e.g., *Diva-hisdb*, *Saint-Gall*, and *Parzival*.) and Printed datasets (e.g., *IlluHisDoc*, *BIR*, and *READ-BAD*). Some datasets contain both printed and handwritten image samples, such as *HJDataset* and *GloSAT*. Table XIII provides a summary of historical document image datasets, presenting details such as the year of publication, references, dataset size, dataset types (*handwritten*, *printed*, and *both*), language, tasks, and corresponding access links. For a comprehensive review of prevalent handwritten and printed historical image datasets, interested readers can refer to [32].

B. Evaluation Metrics

Since historical documents come in various types and exhibit unique characteristics, it is challenging to establish a universal set of criteria to evaluate and standardize the assessment of ongoing research. Thus, due to the absence of standardized evaluation metrics, earlier works were *subjectively quantified* [264]. However, subjective evaluation has a significant drawback in that it cannot be used to compare studies and measure the overall progress made by the research toward achieving efficient HDP. *Objective computations*, on the other hand, offer a means of quantifying performance that is easy to compute and enables comparison among methods, thus evaluating algorithm performance.

A study conducted by Nikolaidou et al. [32] to analyze performance measures has categorized metrics into three different levels of evaluation (see Table XIV). The *first* level of evaluation is based on the *pixel-level*, which calculates the number of matched pixels between the predicted and ground-truth data to measure the accuracy of the algorithm. The *second* level is *region-level* based, which focuses on abstract performance. The *jaccard coefficient*, also known as intersection over union (IoU), is an example of a region-level evaluation metric. In contrast to pixel-level evaluation based on recognition rate and region-level that focused on abstract performance, the *third* level, *customizable* evaluation, provides detailed insights into the performance analysis of the proposed algorithm or methodology. Table XV summarizes the various performance evaluation metrics and their formulations.

TABLE XIII
HISTORICAL DOCUMENT IMAGE DATASETS

Year	Ref	Dataset	Size	Type			Language	Task	Link
				HW	Printed	Both			
2009	[199]	GERMANA	Pages: 764	✓			Catalan, French German, Italian Latin, Spanish	HWR, TBD BLD	https://bv.gva.es/va/
2009	[200]	IAM-HistDB (Parzival)	Pages: 47	✓			German	HWR, WS LA	https://fki.tic.heia-fr.ch/databases/parzival-database
2010	[201]	RODRIGO	Pages: 853	✓			Castilian	HWR, TBD BLD	https://zenodo.org/records/1490009
2011	[202]	PHTD	Pages: 140	✓			Persian	TLS, SR/U WS/R, CS	https://users.encs.concordia.ca/j_sadri/PersianDatabase.htm
2011	[203]	IAM-HistDB (Saint Gall)	Pages: 60	✓			Latin	HWR, LA	https://fki.tic.heia-fr.ch/databases/saint-gall-database
2012	[204]	IAM-HistDB (George Washington)	Pages: 20	✓			English	HWR WS	http://memory.loc.gov/ammem/gwhtml/gwseries2.html
2012	[205]	PBOK	Pages: 707	✓			Persina, Bangla Oriya, Kannada	TLS WS/R	https://link.springer.com/article/10.1007/s10032-010-0122-8
2013	[180]	IMPACT	Pages: 600K			✓	MultL (18)	LA	www.digitisation.eu
2013	[206]	ELSPOSALLES LICENCES/ INDEX	Pages: 173/29	✓			Spanish	OHWR	http://www.cvc.uab.es/5cofm/groundtruth
2014	[207]	BH2M	Pages: 174	✓			Catalan	LA/R/U	http://dag.cvc.uab.es/the-historical-marriages-database/
2014	[208]	HADARA80P	Pages: 80	✓			Arabic	WS	https://www.tu-braunschweig.de/ifn//data/HADARA80P
2015	[209]	ENP	Pages: 528		✓		Dutch, English Estonian, Finnish French, German Latvian, Polish Russian, Serbian Swedish, Yidish Ukrainian	OCR TR, LA	https://www.primaresearch.org/datasets/ENP
2015	[210]	GRPOLY-DB	Pages: 399			✓	Greek	WS TLS	https://www.iit.demokritos.gr/-nstamlGRPOLY-DB
2016	[167]	DIVA-HisDB	Images: 150	✓			Latin Italian	LA	https://diuf.unifr.ch/main/hisdoc/diva-hisdb.html
2017	[211]	HBA	Images: 4K			✓	Latin Italian	LA	http://hba.litislab.eu/index.php/dataset/
2017	[212]	ICDAR17 REID2017	Images: 31		✓		Bengali, English	LA TR	https://www.primaresearch.org/datasets/REID2017
2018	[213]	READ-BAD	Images: 2036		✓		Latin	BLD	https://zenodo.org/records/835441
2018	[214]	Warped Arabic	Images: 200			✓	Arabic	TLS	https://sites.google.com/site/arpangarai/research-area/sample-images-of-wdid
2018	[135]	Tripitaka TKH/ MTH	Images: 1K/ 500	✓			Chinese	CD/R	https://github.com/HCIILAB/TKH_MTH_Datasets_Release
2018	[215]	ICFHR18 RASM2018	Images: 95	✓			Arabic	PS, TLD OCR	https://www.primaresearch.org/RASM2018/
2018	[216]	OHG	Pages: 596	✓			Spanish	LA, HTR	https://zenodo.org/record/1322666#.Ypi6Ty8RoUE
2019	[217]	Pinkas	Pages: 30	✓			Hebrew	WS, PS	https://zenodo.org/records/3569694
2019	[218]	BADAM	Pages: 400	✓			Arabic, Persian	BLD	https://zenodo.org/records/3274428
2019	[219]	HORAE	Pages: 557	✓			Latin	LA	https://figshare.com/articles/dataset/Horse_dataset/12336773

(Continued)

TABLE XIII
(Continued.) HISTORICAL DOCUMENT IMAGE DATASETS

Year	Ref	Dataset	Size	Type			Language	Task	Link
				HW	Printed	Both			
2019	[220]	ICDAR19 DMAS2019	Pages: 100		✓		-	AS, PS TR	https://www.primaresearch.org/DMAS2019/
2019	[221]	CASIA-AHCDB	Pages: 11937	✓			Chinese	CD/R	http://www.nlpr.ia.ac.cn/pal/CASIA-AHCDB.html
2019	[222]	ICDAR19 HDRC Chinese DB	Pages: 12840	✓			Chinese	LA, TR	https://tc11.cvc.uab.es/datasets/ICDAR2019HDRC_1
2019	[223]	ICDAR19 REID	Pages: 81		✓		Bengali, English	LA, TR	https://www.primaresearch.org/REID2019/
2019	[224]	ICDAR19 RASM2019	Images: 120	✓			Arabic	PS, TLD	https://www.primaresearch.org/RASM2019/
2020	[225]	FCR	Pages: 500	✓			Swedish	LA, HTR	https://zenodo.org/records/3945088
2020	[226]	ScribbleLens	Pages: 1K	✓			Dutch	HWR	https://www.openslr.org/84/
2020	[227]	HJDataset	Images: 271			✓	Japanese	LA	https://dell-research-harvard.github.io/HJDataset/
2021	[176]	BIR-database	Pages: 285		✓		French, Latin	WS	https://github.com/asciusb/BIR-database
2021	[228]	BiblIA	Pages: 202	✓			Hebrew, Aramaic	TLS	https://zenodo.org/records/5167263
2022	[229]	Hugin-Munin	Pages: 828	✓			Norwegian	HTR	https://hugin-munin-project.github.io/
2023	[230]	NAS	Pages: 200 Pages: 183 Pages: 230	✓			Finnish French German	AS	https://zenodo.org/records/5654858 https://zenodo.org/records/5654841 https://zenodo.org/records/5654907

Ref: Reference, HW: Handwritten, HWR: Handwriting Recognition, OHWR: Offline Handwriting Recognition, WS: Word Spotting, LA: Layout Analysis, TBD: Textblock Detection, BLD: Baseline Detection, TLS: Text Line Segmentation, SR/U: Sentence Recognition/Understanding, WS/R: Word Segmentation/Recognition, CS: Character Segmentation, MultL: Multiple, LA/R/U: Layout Analysis/Recognition/Understanding, TR: Text Recognition, PHTD: Persian Handwritten Text Dataset, PBOK: Persian, Bangla, Oriya and Kannada, TR: Text Recognition, BH2M: Barcelona Historical Handwritten Marriages, GRPOLY-DB: Greek Polytonic Database, HBA: Historical Book Analysis, CD/R: Character Detection/Recognition, OHG: Oficio de Hipotecas de Girona, HTR: Handwritten Text Recognition, PS: Page Segmentation, AS: Article Segmentation, BIR: Bold Italic Regular, NAS: NewsEye Article Separation.

TABLE XIV
HISTORICAL DOCUMENT PROCESSING PERFORMANCE EVALUATION METRICS

Metric	Level	Reference(s)
Benchmarked and Element Recognition Rate-Based		
Precision	Pixel	[231], [232], [233]
Recall	Pixel	[18], [234], [235]
Exact Match	Pixel	[18], [235]
Hamming Score	Pixel	[235]
Matching Score	Pixel	[120], [235], [236], [237], [238]
Completeness Score	-	[108], [239], [240]
F-Score	Pixel	[3], [113], [114], [241], [242], [243], [244], [245], [246], [247], [248], [249], [250], [251]
Success Rate	Pixel	[18], [115], [116], [117], [120], [231], [252]
Error Rate	Pixel	[8], [18], [235]
Connected Components	Pixel	[18], [125], [232], [253], [254], [255], [256], [257]
Text/Nontext Classification	Pixel	[258], [259]
Intersection over Union (IoU)	Region	[7], [260], [261]
mean Intersection over Union (mIoU)	Region	[18], [117]
Frequency Weighted IoU (f.w.IoU)	Region	[18], [117]
mean Diarization Error Rate (mDER)	-	[6], [262]
Customizable		
Merge	-	[18], [162], [164], [263]
Split	-	[18], [163], [165]
Miss	-	[18], [166]
Partial Miss	-	[18], [117]
False Detection	-	[18], [162]

TABLE XV
EVALUATION METRICS FORMULATION

Metric	Description	Formulation	Denotation
Exact Match (EM)	Precise count of the matched predicted labels with the ground-truth labels.	$EM = \frac{\lvert\{i \lvert P_i = A_i\}\rvert}{N}$	N : # Total pixels. A_i : Actual label for i^{th} pixel.
Hamming Score (HS)	Low-key variant of the Exact Match metric referring to the fraction of correctly predicted labels.	$HS = 1 - \frac{1}{N \cdot \lvert L \rvert} \cdot \sum \mid P_i \oplus A_i \mid$	P_i : Predicted label.
Precision	Signifies the count of entities predicted with label l is actually of label l.	$P_l = \frac{n_{ll}}{n_{ll}+n_{lj}}$	n_{lj} : # Pixels of label l predicted to be of label j.
Recall	Signifies the count of entities of label l that are actually predicted with label l.	$R_l = \frac{n_{ll}}{n_{ll}+n_{jl}}$	n_{ll} : # True positives.
F-score	Harmonic mean of Precision and Recall metrics	$F - score = \frac{2 \times P_l \times R_l}{P_l + R_l}$	n_{lj} : # False positives.
Accuracy (Acc)	Measures the observational error between observed and actual results.	$Acc = \frac{\sum_l n_{ll}}{\sum_l t_l}$	n_{jl} : # False negatives.
Mean Accuracy (mAcc)	Average of accuracy.	$mAcc = \frac{1}{L} \times \sum_l \frac{n_{ll}}{t_l}$	t_l : # Total pixels with label l.
Intersection over Union (IoU)	A statistic to estimate similarity and diversity in sample sets.	$IoU_l = \frac{n_{ll}}{n_{ll}+n_{lj}+n_{jl}}$	L : # Total labels.
mean IoU (mIoU)	Average of Intersection over Union.	$mIoU = \frac{1}{L} \times \sum_l \frac{n_{ll}}{t_l + \sum_j n_{jl} - n_{ll}}$	
Frequency Weighted IoU (f.w.IoU)	Frequency weighted Intersection over Union.	$f.w.IoU = \frac{1}{\sum_k t_k} \times \sum_l \frac{t_l \times n_{ll}}{t_l + \sum_j n_{jl} - n_{ll}}$	
Missed Detection Rate (MDR)	The ratio of the count of missed (not detected) entities to the count of ground truth entities.	$MDR = \frac{\#Misses}{N}$	*Misses* : # Undetected entities by the recognition system. N : # Actual Entities.
False Alarm Rate (FAR)	The percentage of false alarms reported by the recognition system.	$FAR = \frac{\#False\_alarms}{M}$	*False_alarms* : # Unmatched entities generated by the recognition system. M : # Detected Entities.
Error Weight (EW)	The ratio of the error rate factor to the number of types of errors considered in the evaluation scenario.	$EW = \frac{(N-1)ER_i+1}{N}$	$ER_i = \alpha_i \cdot err_i$, where, α_i is the predefined error weight for i^{th} error.
Success Rate (SR)	Correctness of recognition system computed on overall error weights and error types considered at the time of evaluation.	$SR = \frac{\sum_{i=1}^{N} w_i}{\sum_{i=1}^{N} \frac{w_i}{1-ER_i}}$	w_i : Predefined weights. $o2o$: # Complete matched entities with $(MS >= AT)$, where, AT is acceptance threshold and MS is matching score.
Detection Rate (DR)	The percentage of detected ground-truth entities by the recognition system.	$DR = w_1 \cdot \frac{o2o}{N} + w_2 \cdot \frac{g\_o2m}{N} + w_3 \cdot \frac{g\_m2o}{N}$	g_o2m : # Ground-truth entities partially matched with 2 or more detected entities. g_m2o : # Two or more ground-truth entities partially matched with one detected entity.

V. CONCLUSION

Historical documents mirror past happenings and are a vital link that connects the present generation with past civilizations, symbolizing cultural heritage and are documents of a nation's interests. The preservation of these invaluable abundant resources necessitates transforming them from *fugitive physical form* to *forever digital format*. In the digitization process, *structural* and *semantic* layout analysis are fundamental, however, challenging tasks since there is a myriad of historical archives interlaced with diverse and complex layouts. Machine and deep learning methods have proven to give better results yet there are shortcomings to these models that require further investigations, for instance, representative features generation and addressing the imbalance in data and development of fast robust segmentation techniques.

The current survey presents a broad overview of the different essential tasks involved in the HDPP and offers a general framework that connects related tasks that are critical to efficient document processing. This study compiles different task-specific approaches and techniques suggested by researchers to provide readers with a clear understanding of the ongoing work and the progress achieved so far, as well as the unaddressed challenges in the field.

The key contributions of the current work are summarized as follows.

1) The presented work delineates and categorizes various image degradation factors and acquisition challenges inherent in digitizing historical documents, as detailed in Section III-A.
2) Our work presents a comprehensive generic workflow pipeline for HDP and digitization, elaborated in Section III.
3) The current work overviews various document preprocessing techniques, including image binarization, enhancement, and normalization, as discussed in Section III-B.
4) Our work provides an in-depth exploration of syntactic tasks encompassing page, text line, and character segmentation; semantic techniques; and panoptic methods utilized in historical document analysis, detailed in Sections III-C, III-D, and III-E, respectively.
5) The work overviews ATR approaches, encompassing optical character recognition and optical text recognition, alongside insights into relevant software, tools, and recognition systems for HDP, as outlined in Section III-F.
6) We further explored various datasets and evaluation metrics employed in the state-of-the-art historical document digitization techniques, providing a comprehensive overview of the assessment methods, mentioned in Section IV.

Some of the key prospects for future research and opportunities in HDP include exploring the following areas.

1) *Exploration of Advanced Segmentation Techniques*: Investigating and developing advanced segmentation techniques, with a particular focus on addressing the challenges presented by diverse and complex layouts in historical archives.
2) *Handling Imbalanced Data*: Further research can tackle imbalances in datasets of historical documents, exploring techniques to ensure robust performance across various document types.
3) *Representative Feature Generation*: Additional research into novel approaches for generating representative features in machine and deep learning models can significantly enhance the accuracy and efficiency of HDP.
4) *Integration of Panoptic Methods*: It would be interesting to explore the integration and enhancement of panoptic methods for a more comprehensive analysis of historical documents, considering both semantic and syntactic aspects.

Some of the key implications that the discussion on HDP and digitization presents for researchers, practitioners, and policymakers are as follows.

1) *Researchers*: This survey paper provides researchers with a comprehensive overview of the advancements, challenges, and trends in HDP and digitization. It highlights various techniques, methods, and approaches employed in this field, offering a roadmap for future research directions.
2) *Practitioners*: For practitioners involved in HDP and digitization, the survey paper serves as a valuable resource for staying up-to-date with the latest methodologies and best practices. It outlines practical challenges and methodologies that can guide practitioners in their efforts to preserve and digitize historical documents. Moreover, it sheds light on potential tools and technologies that can be integrated into their workflows to improve efficiency and accuracy.
3) *Policymakers*: The paper underscores the importance of funding and supporting initiatives that aim to preserve valuable cultural artifacts through digitization. It provides policymakers with a clearer understanding of the technical challenges and opportunities in this domain, aiding them in making informed decisions regarding resource allocation and policy formulation.

Authors' Contribution Statement

The authors conceptualized and conducted the survey, analyzed the literature, and contributed to manuscript writing. Nancy Girdhar drafted the initial version with assistance from Mickaël Coustaty. All authors participated in editing and revising the manuscript.

References

[1] J. Philips and N. Tabrizi, "Historical document processing: A survey of techniques, tools, and trends," in *Proc. 12th Int. Joint Conf. Knowl. Discovery, Knowl. Eng. Knowl. Manag. (KDIR)*, 2020, pp. 341–349.
[2] C.-H. Chou, W.-H. Lin, and F. Chang, "A binarization method with learning-built rules for document images produced by cameras," *Pattern Recognit.*, vol. 43, no. 4, pp. 1518–1530, 2010.
[3] K. Chen, C.-L. Liu, M. Seuret, M. Liwicki, J. Hennebert, and R. Ingold, "Page segmentation for historical document images based on superpixel classification with unsupervised feature learning," in *Proc. 12th IAPR Workshop Document Anal. Syst. (DAS)*, Piscataway, NJ, USA: IEEE Press, 2016, pp. 299–304.
[4] K. Chen, M. Seuret, M. Liwicki, J. Hennebert, C.-L. Liu, and R. Ingold, "Page segmentation for historical handwritten document images using conditional random fields," in *Proc. 15th Int. Conf. Frontiers Handwriting Recognit. (ICFHR)*, Piscataway, NJ, USA: IEEE Press, 2016, pp. 90–95.
[5] S. Fogel, H. Averbuch-Elor, S. Cohen, S. Mazor, and R. Litman, "ScrabbleGAN: Semi-supervised varying length handwritten text generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 4324–4333.
[6] B. Meier, T. Stadelmann, J. Stampfli, M. Arnold, and M. Cieliebak, "Fully convolutional neural networks for newspaper article segmentation," in *Proc. 14th IAPR Int. Conf. Document Anal. Recognit. (ICDAR)*, vol. 1, Piscataway, NJ, USA: IEEE Press, 2017, pp. 414–419.
[7] S. A. Oliveira, B. Seguin, and F. Kaplan, "dhSegment: A generic deep-learning approach for document segmentation," in *Proc. 16th Int. Conf. Frontiers Handwriting Recognit. (ICFHR)*, Piscataway, NJ, USA: IEEE Press, 2018, pp. 7–12.
[8] M. Seuret, M. Alberti, M. Liwicki, and R. Ingold, "PCA-initialized deep neural networks applied to document image analysis," in *Proc. 14th IAPR Int. Conf. Document Anal. Recognit. (ICDAR)*, vol. 1, Piscataway, NJ, USA: IEEE Press, 2017, pp. 877–882.
[9] F. Lombardi and S. Marinai, "Deep learning for historical document analysis and recognition—A survey," *J. Imag.*, vol. 6, no. 10, 2020, Art. no. 110.
[10] C.-W. Tang, C.-L. Liu, and P.-S. Chiu, "HRCenterNet: An anchorless approach to Chinese character segmentation in historical documents," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Piscataway, NJ, USA: IEEE Press, 2020, pp. 1924–1930.
[11] C.-W. Tang, C.-L. Liu, and P.-S. Chiu, "HRRegionNet: Chinese character segmentation in historical documents with regional awareness,"

in *Proc. 16th Int. Conf. Document Anal. Recognit. (ICDAR)*, in Part IV 16, Lausanne, Switzerland. New York, NY, USA: Springer-Verlag, 2021, pp. 3–17.

[12] J. M. O. Koistinen, K. T. Kettunen, and T. A. Pääkkönen, "Improving optical character recognition of Finnish historical newspapers with a combination of Fraktur & Antiqua models and image preprocessing," in *Proc. 21st Nordic Conf. Comput. Linguistics, NoDaLiDa*, Gothenburg, Sweden: Linköping Univ. Electronic Press, 2017.

[13] S. Drobac, P. Kauppinen, and K. Lindén, "OCR and post-correction of historical Finnish texts," in *Proc. 21st Nordic Conf. Comput. Linguistics*, 2017, pp. 70–76.

[14] S. Drobac and K. Lindén, "Optical character recognition with neural networks and post-correction with finite state methods," *Proc. Int. J. Document Anal. Recognit. (IJDAR)*, 2020, vol. 23, no. 4, pp. 279–295.

[15] M. Bulacu, R. Van Koert, L. Schomaker, and T. van der Zant, "Layout analysis of handwritten historical documents for searching the archive of the cabinet of the Dutch queen," in *Proc. 9th Int. Conf. Document Anal. Recognit. (ICDAR)*, vol. 1, Piscataway, NJ, USA: IEEE Press, 2007, pp. 357–361.

[16] J. Andrés, J. R. Prieto, E. Granell, V. Romero, J. A. Sánchez, and E. Vidal, "Information extraction from handwritten tables in historical documents," in *Proc. Int. Workshop Document Anal. Syst.*, New York, NY, USA: Springer-Verlag, 2022, pp. 184–198.

[17] K. Barrere, Y. Soullard, A. Lemaitre, and B. Coüasnon, "A light transformer-based architecture for handwritten text recognition," in *Proc. Int. Workshop Document Anal. Syst.*, New York, NY, USA: Springer-Verlag, 2022, pp. 275–290.

[18] G. M. Binmakhashen and S. A. Mahmoud, "Document layout analysis: A comprehensive survey," *ACM Comput. Surveys (CSUR)*, vol. 52, no. 6, pp. 1–36, 2019.

[19] S. Eskenazi, P. Gomez-Krämer, and J.-M. Ogier, "A comprehensive survey of mostly textual document segmentation algorithms since 2008," *Pattern Recognit.*, vol. 64, pp. 1–14, Apr. 2017.

[20] L. Likforman-Sulem, A. Zahour, and B. Taconet, "Text line segmentation of historical documents: A survey," *Int. J. Document Anal. Recognit. (IJDAR)*, vol. 9, no. 2, pp. 123–138, 2007.

[21] C. Tensmeyer and T. Martinez, "Historical document image binarization: A review," *SN Comput. Sci.*, vol. 1, no. 3, 2020, Art. no. 173.

[22] M. T. Parvez and S. A. Mahmoud, "Offline Arabic handwritten text recognition: A survey," *ACM Comput. Surveys (CSUR)*, vol. 45, no. 2, pp. 1–35, 2013.

[23] J. A. Sanchez, V. Romero, A. H. Toselli, M. Villegas, and E. Vidal, "ICDAR2017 competition on handwritten text recognition on the read dataset," in *Proc. 14th IAPR Int. Conf. Document Anal. Recognit. (ICDAR)*, vol. 1, Piscataway, NJ, USA: IEEE Press, 2017, pp. 1383–1388.

[24] A. Zoizou, A. Zarghili, and I. Chaker, "Skew correction and text line extraction of Arabic historical documents," in *Proc. 7th Int. Conf. Arabic Lang. Process. Theory Pract. (ICALP)*, Nancy, France. New York, NY, USA: Springer-Verlag, 2019, pp. 181–193.

[25] J. H. Uhl, S. Leyk, Y.-Y. Chiang, W. Duan, and C. A. Knoblock, "Automated extraction of human settlement patterns from historical topographic map series using weakly supervised convolutional neural networks," *IEEE Access*, vol. 8, pp. 6978–6996, 2019.

[26] M. Sezgin and B. Sankur, "Survey over image thresholding techniques and quantitative performance evaluation," *J. Electron. Imag.*, vol. 13, no. 1, pp. 146–165, 2004.

[27] D. Sinwar, V. S. Dhaka, N. Pradhan, and S. Pandey, "Offline script recognition from handwritten and printed multilingual documents: A survey," *Int. J. Document Anal. Recognit. (IJDAR)*, vol. 24, pp. 97–121, Jun. 2021.

[28] J. Martínek, L. Lenc, and P. Král, "Building an efficient OCR system for historical documents with little training data," *Neural Comput. Appl.*, vol. 32, no. 23, pp. 17209–17227, 2020.

[29] O. Mechi, M. Mehri, R. Ingold, and N. Essoukri Ben Amara, "A two-step framework for text line segmentation in historical Arabic and Latin document images," *Int. J. Document Anal. Recognit. (IJDAR)*, vol. 24, no. 3, pp. 197–218, 2021.

[30] Y. Mei, X. Wang, and J. Wang, "A Chinese character segmentation algorithm for complicated printed documents," *Int. J. Signal Process., Image Process. Pattern Recognit.*, vol. 6, no. 3, pp. 91–100, 2013.

[31] M.-K. Zhou, X.-Y. Zhang, F. Yin, and C.-L. Liu, "Discriminative quadratic feature learning for handwritten Chinese character recognition," *Pattern Recognit.*, vol. 49, pp. 7–18, Jan. 2016.

[32] K. Nikolaidou, M. Seuret, H. Mokayed, and M. Liwicki, "A survey of historical document image datasets," *Int. J. Document Anal. Recognit. (IJDAR)*, vol. 25, no. 4, pp. 305–338, 2022.

[33] D. Doermann and K. Tombre, *Handbook of Document Image Processing and Recognition*, vol. 1. New York, NY, USA: Springer-Verlag, 2014.

[34] W. Niblack, *An Introduction to Digital Image Processing*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1985.

[35] W. K. Pratt, *Digital Image Processing: PIKS Scientific Inside*, vol. 4. Hoboken, NJ, USA: Wiley, 2007.

[36] F. M. Wahl, K. Y. Wong, and R. G. Casey, "Block segmentation and text extraction in mixed text/image documents," *Comput. Graph. Image Process.*, vol. 20, no. 4, pp. 375–390, 1982.

[37] O. D. Trier and T. Taxt, "Evaluation of binarization methods for document images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 17, no. 3, pp. 312–315, Mar. 1995.

[38] I.-K. Kim, D.-W. Jung, and R.-H. Park, "Document image binarization based on topographic analysis using a water flow model," *Pattern Recognit.*, vol. 35, no. 1, pp. 265–277, 2002.

[39] E. Badekas and N. Papamarkos, "Optimal combination of document binarization techniques using a self-organizing map neural network," *Eng. Appl. Artif. Intell.*, vol. 20, no. 1, pp. 11–24, 2007.

[40] B. Gatos, I. Pratikakis, and S. J. Perantonis, "Efficient binarization of historical and degraded document images," in *Proc. 8th IAPR Int. Workshop Document Anal. Syst.*, Piscataway, NJ, USA: IEEE Press, 2008, pp. 447–454.

[41] Y.-H. Tseng and H.-J. Lee, "Document image binarization by two-stage block extraction and background intensity determination," *Pattern Anal. Appl.*, vol. 11, no. 1, pp. 33–44, 2008.

[42] V. Vonikakis, I. Andreadis, and N. Papamarkos, "Robust document binarization with off center-surround cells," *Pattern Anal. Appl.*, vol. 14, no. 3, pp. 219–234, 2011.

[43] B. Su, S. Lu, and C. L. Tan, "Combination of document image binarization techniques," in *Proc. Int. Conf. Document Anal. Recognit.*, Piscataway, NJ, USA: IEEE Press, 2011, pp. 22–26.

[44] B. Gatos, I. Pratikakis, and S. J. Perantonis, "Adaptive degraded document image binarization," *Pattern Recognit.*, vol. 39, no. 3, pp. 317–327, 2006.

[45] M. Kamel and A. Zhao, "Extraction of binary character/graphics images from grayscale document images," *CVGIP Graph. Models Image Process.*, vol. 55, no. 3, pp. 203–217, 1993.

[46] J. Sauvola and M. Pietikäinen, "Adaptive document image binarization," *Pattern Recognit.*, vol. 33, no. 2, pp. 225–236, 2000.

[47] Y. Chen and G. Leedham, "Decompose algorithm for thresholding degraded historical document images," *IEE Proc.-Vis., Image Signal Process.*, vol. 152, no. 6, pp. 702–714, 2005.

[48] S. Huang, M. Ahmadi, and M. A. Sid-Ahmed, "A hidden Markov model-based character extraction method," *Pattern Recognit.*, vol. 41, no. 9, pp. 2890–2900, 2008.

[49] E. Badekas, N. Nikolaou, and N. Papamarkos, "Text binarization in color documents," *Int. J. Imag. Syst. Technol.*, vol. 16, no. 6, pp. 262–274, 2006.

[50] C.-M. Tsai and H.-J. Lee, "Binarization of color document images via luminance and saturation color features," *IEEE Trans. Image Process.*, vol. 11, no. 4, pp. 434–451, Apr. 2002.

[51] B. Wang, X.-F. Li, F. Liu, and F.-Q. Hu, "Color text image binarization based on binary texture analysis," *Pattern Recognit. Lett.*, vol. 26, no. 10, pp. 1568–1576, 2005.

[52] M. Cheriet, J. N. Said, and C. Y. Suen, "A recursive thresholding technique for image segmentation," *IEEE Trans. Image Process.*, vol. 7, no. 6, pp. 918–921, Jun. 1998.

[53] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst., Man, Cybern.*, vol. 9, no. 1, pp. 62–66, Jan. 1979.

[54] Y. Solihin and C. Leedham, "Integral ratio: A new class of global thresholding techniques for handwriting images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, no. 8, pp. 761–768, Aug. 1999.

[55] C.-C. Leung, K.-S. Chan, H.-M. Chan, and W.-K. Tsui, "A new approach for image enhancement applied to low-contrast–low-illumination IC and document images," *Pattern Recognit. Lett.*, vol. 26, no. 6, pp. 769–778, 2005.

[56] L. Likforman-Sulem, J. Darbon, and E. H. B. Smith, "Enhancement of historical printed document images by combining total variation regularization and non-local means filtering," *Image Vis. Comput.*, vol. 29, no. 5, pp. 351–363, 2011.

[57] S. Nomura, K. Yamanaka, T. Shiose, H. Kawakami, and O. Katai, "Morphological preprocessing method to thresholding degraded word images," *Pattern Recognit. Lett.*, vol. 30, no. 8, pp. 729–744, 2009.

[58] D. Fadoua, F. L. Bourgeois, and H. Emptoz, "Restoring ink bleed-through degraded document images using a recursive unsupervised classification technique," in *Proc. Int. Workshop Document Anal. Syst.*, New York, NY, USA: Springer-Verlag, 2006, pp. 38–49.

[59] E. Dubois and A. Pathak, "Reduction of bleed-through in scanned manuscript documents," in *Proc. Image Capture Syst. (PICS)*, vol. 1. Citeseer, 2001, pp. 177–180.

[60] R. F. Moghaddam and M. Cheriet, "A variational approach to degraded document enhancement," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 8, pp. 1347–1361, Aug. 2009.

[61] Y. Yang and H. Yan, "An adaptive logical method for binarization of degraded document images," *Pattern Recognit.*, vol. 33, no. 5, pp. 787–807, 2000.

[62] T. Obafemi-Ajayi, G. Agam, and O. Frieder, "Historical document enhancement using LUT classification," *Int. J. Document Anal. Recognit. (IJDAR)*, vol. 13, no. 1, pp. 1–17, 2010.

[63] F. Drira, F. LeBourgeois, and H. Emptoz, "A new PDE-based approach for singularity-preserving regularization: Application to degraded characters restoration," *Int. J. Document Anal. Recognit. (IJDAR)*, vol. 15, no. 3, pp. 183–212, 2012.

[64] L. O'Gorman, "The document spectrum for page layout analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 15, no. 11, pp. 1162–1173, Nov. 1993.

[65] F. Shafait and T. M. Breuel, "A simple and effective approach for border noise removal from document images," in *Proc. IEEE 13th Int. Multitopic Conf.*, Piscataway, NJ, USA: IEEE Press, 2009, pp. 1–5.

[66] M. M. Haji, T. D. Bui, and C. Y. Suen, "Simultaneous document margin removal and skew correction based on corner detection in projection profiles," in *Proc. Int. Conf. Image Anal. Process.*, New York, NY, USA: Springer-Verlag, 2009, pp. 1025–1034.

[67] D. X. Le, G. R. Thoma, and H. Wechsler, "Automated borders detection and adaptive segmentation for binary document images," in *Proc. 13th Int. Conf. Pattern Recognit.*, vol. 3. Piscataway, NJ, USA: IEEE Press, 1996, pp. 737–741.

[68] N. Stamatopoulos, B. Gatos, and A. Kesidis, "Automatic borders detection of camera document images," in *Proc. 2nd Int. Workshop Camera-Based Document Anal. Recognit.*, Curitiba, Brazil, 2007, pp. 71–78.

[69] R. S. Caprari, "Algorithm for text page up/down orientation determination," *Pattern Recognit. Lett.*, vol. 21, no. 4, pp. 311–317, 2000.

[70] H. B. Aradhye, "A generic method for determining up/down orientation of text in roman and non-roman scripts," *Pattern Recognit.*, vol. 38, no. 11, pp. 2114–2131, 2005.

[71] D. S. Le, G. R. Thoma, and H. Wechsler, "Automated page orientation and skew angle detection for binary document images," *Pattern Recognit.*, vol. 27, no. 10, pp. 1325–1344, 1994.

[72] P.-Y. Yin, "Skew detection and block classification of printed documents," *Image Vis. Comput.*, vol. 19, no. 8, pp. 567–579, 2001.

[73] H. S. Baird, "The skew angle of printed documents," in *Proc. SPSE's 40th Annu. Conf. Symp. Hybrid Imag. Syst.*, 1987, p. 21.

[74] A. Teruo and H. Norihiro, "Automated entry system for printed documents," *Pattern Recognit.*, vol. 23, no. 11, pp. 1141–1154, 1990.

[75] S. Li, Q. Shen, and J. Sun, "Skew detection using wavelet decomposition and projection profile analysis," *Pattern Recognit. Lett.*, vol. 28, no. 5, pp. 555–562, 2007.

[76] A. Amin and S. Fischer, "A document skew detection method using the Hough transform," *Pattern Anal. Appl.*, vol. 3, no. 3, pp. 243–253, 2000.

[77] C. Singh, N. Bhatia, and A. Kaur, "Hough transform based fast skew detection and accurate skew correction methods," *Pattern Recognit.*, vol. 41, no. 12, pp. 3528–3546, 2008.

[78] Y. Lu and C. L. Tan, "A nearest-neighbor chain based approach to skew estimation in document images," *Pattern Recognit. Lett.*, vol. 24, no. 14, pp. 2315–2323, 2003.

[79] Y. Cao, S. Wang, and H. Li, "Skew detection and correction in document images based on straight-line fitting," *Pattern Recognit. Lett.*, vol. 24, no. 12, pp. 1871–1879, 2003.

[80] B. Gatos, N. Papamarkos, and C. Chamzas, "Skew detection and text line position determination in digitized documents," *Pattern Recognit.*, vol. 30, no. 9, pp. 1505–1519, 1997.

[81] H. Yan, "Skew correction of document images using interline cross-correlation," *CVGIP Graph. Models Image Process.*, vol. 55, no. 6, pp. 538–543, 1993.

[82] H. Liu, Q. Wu, H. Zha, and X. Liu, "Skew detection for complex document images using robust borderlines in both text and non-text regions," *Pattern Recognit. Lett.*, vol. 29, no. 13, pp. 1893–1900, 2008.

[83] H. Fan, L. Zhu, and Y. Tang, "Skew detection in document images based on rectangular active contour," *Int. J. Document Anal. Recognit. (IJDAR)*, vol. 13, no. 4, pp. 261–269, 2010.

[84] C.-H. Chou, S.-Y. Chu, and F. Chang, "Estimation of skew angles for scanned documents based on piecewise covering by parallelograms," *Pattern Recognit.*, vol. 40, no. 2, pp. 443–455, 2007.

[85] Y.-K. Chen and J.-F. Wang, "Skew detection and reconstruction of color-printed document images," *IEICE Trans. Inf. Syst.*, vol. 84, no. 8, pp. 1018–1024, 2001.

[86] M. Makridis, N. Nikolaou, and N. Papamarkos, "An adaptive technique for global and local skew correction in color documents," *Expert Syst. Appl.*, vol. 37, no. 10, pp. 6832–6843, 2010.

[87] E. Kavallieratou, N. Fakotakis, and G. Kokkinakis, "Slant estimation algorithm for OCR systems," *Pattern Recognit.*, vol. 34, no. 12, pp. 2515–2522, 2001.

[88] A. Vinciarelli and J. Luettin, "A new normalization technique for cursive handwritten words," *Pattern Recognit. Lett.*, vol. 22, no. 9, pp. 1043–1050, 2001.

[89] F. Kimura, M. Shridhar, and Z. Chen, "Improvements of a lexicon directed algorithm for recognition of unconstrained handwritten words," in *Proc. 2nd Int. Conf. Document Anal. Recognit. (ICDAR'93)*, Piscataway, NJ, USA: IEEE Press, 1993, pp. 18–22.

[90] Y. Ding, F. Kimura, Y. Miyake, and M. Shridhar, "Slant estimation for handwritten words by directionally refined chain code," in *Proc. 7th Int. Workshop Frontiers Handwriting Recognit.*, 2000, pp. 53–61.

[91] Y. Ding, W. Ohyama, F. Kimura, and M. Shridhar, "Local slant estimation for handwritten English words," in *Proc. 9th Int. Workshop Frontiers Handwriting Recognit.*, Piscataway, NJ, USA: IEEE Press, 2004, pp. 328–333.

[92] R. M. Bozinovic and S. N. Srihari, "Off-line cursive script word recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 11, no. 1, pp. 68–83, Jan. 1989.

[93] G. Kim and V. Govindaraju, "A lexicon driven approach to handwritten word recognition for real-time applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 4, pp. 366–379, 1997.

[94] H. Cao, X. Ding, and C. Liu, "Rectifying the bound document image captured by the camera: A model based approach," in *Proc. 7th Int. Conf. Document Anal. Recognit.,* Piscataway, NJ, USA: IEEE Press, 2003, pp. 71–75.

[95] J. Liang, D. DeMenthon, and D. Doermann, "Geometric rectification of camera-captured document images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 4, pp. 591–605, Apr. 2008.

[96] L. Zhang, A. M. Yip, M. S. Brown, and C. L. Tan, "A unified framework for document restoration using inpainting and shape-from-shading," *Pattern Recognit.*, vol. 42, no. 11, pp. 2961–2978, 2009.

[97] M. S. Brown and Y.-C. Tsoi, "Geometric and shading correction for images of printed materials using boundary," *IEEE Trans. Image Process.*, vol. 15, no. 6, pp. 1544–1554, Jun. 2006.

[98] O. Lavialle, X. Molines, F. Angella, and P. Baylou, "Active contours network to straighten distorted text lines," in *Proc. Int. Conf. Image Process. (Cat. No. 01CH37205)*, vol. 3, Piscataway, NJ, USA: IEEE Press, 2001, pp. 748–751.

[99] A. Masalovitch and L. Mestetskiy, "Usage of continuous skeletal image representation for document images de-warping," in *Proc. Int. Workshop Camera-Based Document Anal. Recognit.*, Curitiba, 2007, pp. 45–53.

[100] N. Stamatopoulos, B. Gatos, I. Pratikakis, and S. J. Perantonis, "Goal-oriented rectification of camera-based document images," *IEEE Trans. Image Process.*, vol. 20, no. 4, pp. 910–920, Apr. 2011.

[101] C. Wu and G. Agam, "Document image de-warping for text/graphics recognition," in *Proc. Joint IAPR Int. Workshops Statist. Techn. Pattern Recognit. (SPR) Struct. Syntactic Pattern Recognit. (SSPR)*, New York, NY, USA: Springer-Verlag, 2002, pp. 348–357.

[102] L. Zhang and C. L. Tan, "Warped image restoration with applications to digital libraries," in *Proc. 8th Int. Conf. Document Anal. Recognit. (ICDAR'05)*, Piscataway, NJ, USA: IEEE Press, 2005, pp. 192–196.

[103] A. Ulges, C. H. Lampert, and T. M. Breuel, "Document image dewarping using robust estimation of curled text lines," in *Proc. 8th Int. Conf. Document Anal. Recognit. (ICDAR'05)*, Piscataway, NJ, USA: IEEE Press, 2005, pp. 1001–1005.

[104] S. Lu, B. M. Chen, and C. C. Ko, "A partition approach for the restoration of camera images of planar and curled document," *Image Vis. Comput.*, vol. 24, no. 8, pp. 837–848, 2006.

[105] G. M. BinMakhashen and S. A. Mahmoud, "Historical document layout analysis using anisotropic diffusion and geometric features," *Int. J. Digit. Libraries*, vol. 21, no. 3, pp. 329–342, 2020.
[106] K. Hadjar and R. Ingold, "Arabic newspaper page segmentation," in *Proc. 7th Int. Conf. Document Anal. Recognit.* (*ICDAR*), vol. 3, Citeseer, 2003, pp. 895–899.
[107] B. Gatos, S. Mantzaris, K. Chandrinos, A. Tsigris, and S. J. Perantonis, "Integrated algorithms for newspaper page decomposition and article tracking," in *Proc. 5th Int. Conf. Document Anal. Recognit. (ICDAR'99) (Cat. No. PR00318)*, Piscataway, NJ, USA: IEEE Press, 1999, pp. 559–562.
[108] S. S. Bukhari, T. M. Breuel, A. Asi, and J. El-Sana, "Layout analysis for Arabic historical document images using machine learning," in *Proc. Int. Conf. Frontiers Handwriting Recognit.*, Piscataway, NJ, USA: IEEE Press, 2012, pp. 639–644.
[109] A. Corbelli, L. Baraldi, F. Balducci, C. Grana, and R. Cucchiara, "Layout analysis and content classification in digitized books," in *Proc. Italian Res. Conf. Digit. Libraries*, New York, NY, USA: Springer-Verlag, 2016, pp. 153–165.
[110] V. Buntinx, F. Kaplan, and A. Xanthos, "Layout analysis on newspaper archives," in *Digital* Humanities, 2017, pp. 409–412.
[111] N. Gutehrlé and I. Atanassova, "Processing the structure of documents: Logical layout analysis of historical newspapers in french," 2022, *arXiv:2202.08125*.
[112] C. Grana, D. Borghesani, and R. Cucchiara, "Automatic segmentation of digitalized historical manuscripts," *Multimedia Tools Appl.*, vol. 55, no. 3, pp. 483–506, 2011.
[113] R. Cohen, A. Asi, K. Kedem, J. El-Sana, and I. Dinstein, "Robust text and drawing segmentation algorithm for historical documents," in *Proc. 2nd Int. Workshop Historical Document Imag. Process.*, 2013, pp. 110–117.
[114] A. Asi, R. Cohen, K. Kedem, J. El-Sana, and I. Dinstein, "A coarse-to-fine approach for layout analysis of ancient manuscripts," in *Proc. 14th Int. Conf. Frontiers Handwriting Recognit.*, Piscataway, NJ, USA: IEEE Press, 2014, pp. 140–145.
[115] K. Chen, H. Wei, J. Hennebert, R. Ingold, and M. Liwicki, "Page segmentation for historical handwritten document images using color and texture features," in *Proc. 14th Int. Conf. Frontiers Handwriting Recognit.*, Piscataway, NJ, USA: IEEE Press, 2014, pp. 488–493.
[116] K. Chen, M. Seuret, M. Liwicki, J. Hennebert, and R. Ingold, "Page segmentation of historical document images with convolutional autoencoders," in *Proc. 13th Int. Conf. Document Anal. Recognit. (ICDAR)*, Piscataway, NJ, USA: IEEE Press, 2015, pp. 1011–1015.
[117] K. Chen, M. Seuret, J. Hennebert, and R. Ingold, "Convolutional neural networks for page segmentation of historical document images," in *Proc. 14th IAPR Int. Conf. Document Anal. Recognit. (ICDAR)*, vol. 1, Piscataway, NJ, USA: IEEE Press, 2017, pp. 965–970.
[118] P. Kaddas and B. Gatos, "A deep convolutional encoder-decoder network for page segmentation of historical handwritten documents into text zones," in *Proc. 16th Int. Conf. Frontiers Handwriting Recognit. (ICFHR)*, Piscataway, NJ, USA: IEEE Press, 2018, pp. 259–264.
[119] A. Hazem, B. Daille, D. Stutzmann, C. Kermorvant, and L. Chevalier, "Hierarchical text segmentation for medieval manuscripts," in *Proc. 28th Int. Conf. Comput. Linguistics*, 2020, pp. 6240–6251.
[120] C. Wick and F. Puppe, "Fully convolutional neural networks for page segmentation of historical document images," in *Proc. 13th IAPR Int. Workshop Document Anal. Syst. (DAS)*, Piscataway, NJ, USA: IEEE Press, 2018, pp. 287–292.
[121] B. Liebl and M. Burghardt, "An evaluation of DNN architectures for page segmentation of historical newspapers," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Piscataway, NJ, USA: IEEE Press, 2021, pp. 5153–5160.
[122] N. Fischer, A. Hartelt, and F. Puppe, "Line-level layout recognition of historical documents with background knowledge," *Algorithms*, vol. 16, no. 3, p. 136, 2023.
[123] A. Zahour, L. Likforman-Sulem, W. Boussalaa, and B. Taconet, "Text line segmentation of historical arabic documents," in *Proc. 9th Int. Conf. Document Anal. Recognit. (ICDAR)*, vol. 1, Piscataway, NJ, USA: IEEE Press, 2007, pp. 138–142.
[124] W. Boussellaa, A. Zahour, H. Elabed, A. Benabdelhafid, and A. M. Alimi, "Unsupervised block covering analysis for text-line segmentation of Arabic ancient handwritten document images," in *Proc. 20th Int. Conf. Pattern Recognit.*, Piscataway, NJ, USA: IEEE Press, 2010, pp. 1929–1932.
[125] A. Garz, A. Fischer, R. Sablatnig, and H. Bunke, "Binarization-free text line segmentation for historical documents based on interest point clustering," in *Proc. 10th IAPR Int. Workshop Document Anal. Syst.*, Piscataway, NJ, USA: IEEE Press, 2012, pp. 95–99.
[126] S. Jindal and G. S. Lehal, "Line segmentation of handwritten Gurmukhi manuscripts," in *Proc. Workshop document Anal. Recognit.*, 2012, pp. 74–78.
[127] T. Grüning, G. Leifert, T. Strauss, and R. Labahn, "A robust and binarization-free approach for text line detection in historical documents," in *Proc. 14th IAPR Int. Conf. Document Anal. Recognit. (ICDAR)*, vol. 1, Piscataway, NJ, USA: IEEE Press, 2017, pp. 236–241.
[128] O. Mechi, M. Mehri, R. Ingold, and N. E. B. Amara, "Text line segmentation in historical document images using an adaptive U-Net architecture," in *Proc. Int. Conf. Document Anal. Recognit. (ICDAR)*, Piscataway, NJ, USA: IEEE Press, 2019, pp. 369–374.
[129] E. F. Bilgin Tasdemir, "Printed Ottoman text recognition using synthetic data and data augmentation," *Int. J. Document Anal. Recognit. (IJDAR)*, pp. 1–15, 2023.
[130] L. Gao, X. Zhang, Z. Tang, Y. Huang, and L. Jin, "A sequence labeling based approach for character segmentation of historical documents," in *Proc. 13th IAPR Int. Workshop Document Anal. Syst. (DAS)*, Piscataway, NJ, USA: IEEE Press, 2018, pp. 305–310.
[131] J. Wang et al., "Deep high-resolution representation learning for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3349–3364, Oct. 2021.
[132] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," 2019, *arXiv:1904.07850*.
[133] W. Ma, H. Zhang, L. Jin, S. Wu, J. Wang, and Y. Wang, "Joint layout analysis, character detection and recognition for historical document digitization," in *Proc. 17th Int. Conf. Frontiers Handwriting Recognit. (ICFHR)*, Piscataway, NJ, USA: IEEE Press, 2020, pp. 31–36.
[134] W. Sihang, W. Jiapeng, M. Weihong, and J. Lianwen, "Precise detection of Chinese characters in historical documents with deep reinforcement learning," *Pattern Recognit.*, vol. 107, 2020, Art. no. 107503.
[135] H. Yang, L. Jin, W. Huang, Z. Yang, S. Lai, and J. Sun, "Dense and tight detection of chinese characters in historical documents: Datasets and a recognition guided detector," *IEEE Access*, vol. 6, pp. 30174–30183, 2018.
[136] C. Zhang, W. Wang, H. Liu, G. Zhang, and Q. Lin, "Character detection and segmentation of historical Uchen Tibetan documents in complex situations," *IEEE Access*, vol. 10, pp. 25376–25391, 2022.
[137] N. Rahal, L. Vögtlin, and R. Ingold, "Historical document image analysis using controlled data for pre-training," *Int. J. Document Anal. Recognit. (IJDAR)*, vol. 26, no. 3, Sep. 2023, pp. 241–254.
[138] B. Melodie, M. Martin, P. Thierry, and K. Christopher, "Including keyword position in image-based models for act segmentation of historical registers," in *Proc. 6th Int. Workshop Historical Document Imag. Process.*, 2021, pp. 31–36.
[139] R. Barman, M. Ehrmann, S. Clematide, S. Ares Oliveira, and F. Kaplan, "Combining visual and textual features for semantic segmentation of historical newspapers," *J. Data Mining Digit. Humanities*, 2021 pp. 1–26, Jan. 2021.
[140] T. Palfray, D. Hébert, S. Nicolas, P. Tranouez, and T. Paquet, "Logical segmentation for article extraction in digitized old newspapers," in *Proc. ACM Symp. Document Eng.*, 2012, pp. 129–132.
[141] W. Zhu, N. Sokhandan, G. Yang, S. Martin, and S. Sathyanarayana, "DocBed: A multi-stage OCR solution for documents with complex layouts," 2022, *arXiv:2202.01414*.
[142] S. Zheng et al., "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 6881–6890.
[143] N. Girdhar, M. Coustaty, and A. Doucet, "STRAS: A semantic textual-cues leveraged rule-based approach for article separation in historical newspapers," in *Proc. Int. Conf. Asian Digit. Libraries*, New York, NY, USA: Springer-Verlag, 2023, pp. 89–105.
[144] J. Zenk, F. Kordon, M. Mayr, M. Seuret, and V. Christlein, "Investigations on self-supervised learning for script-, font-type, and location classification on historical documents," in *Proc. 7th Int. Workshop Historical Document Imag. Process.*, 2023, pp. 97–102.
[145] N. Vogler, J. P. Allen, M. T. Miller, and T. Berg-Kirkpatrick, "Lacuna reconstruction: Self-supervised pre-training for low-resource historical document transcription," 2021, *arXiv:2112.08692*.
[146] S. K. Jemni, S. Ammar, M. A. Souibgui, Y. Kessentini, and A. Cheddad, "ST-KeyS: Self-supervised transformer for keyword spotting in historical handwritten documents," 2023, *arXiv:2303.03127*.

[147] H. Bouressace and J. Csirik, "Recognition of the logical structure of arabic newspaper pages," in *Proc. Int. Conf. Text, Speech, Dialogue*, New York, NY, USA: Springer-Verlag, 2018, pp. 251–258.

[148] N. Gutehrlé and I. Atanassova, "Logical layout analysis applied to historical newspapers," in *Proc. Workshop Natural Lang. Process. Digit. Humanities (NLP4DH)*, NLP Assoc. India (NLPAI), 2021, pp. 85–94.

[149] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollár, "Panoptic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 9404–9413.

[150] O. Elharrouss, S. Al-Maadeed, N. Subramanian, N. Ottakath, N. Almaadeed, and Y. Himeur, "Panoptic segmentation: A review," 2021, *arXiv:2111.10250*.

[151] R. Cao, H. Li, G. Zhou, and P. Luo, "Towards document panoptic segmentation with pinpoint accuracy: Method and evaluation," in *Proc. 16th Int. Conf. Document Anal. Recognit. (ICDAR)*, in Part II 16, Lausanne, Switzerland. New York, NY, USA: Springer-Verlag, 2021, pp. 3–18.

[152] M. Lu, F. Bao, and G. Gao, "Panoptic-DLA: Document layout analysis of historical newspapers based on proposal-free panoptic segmentation model," in *Proc. 14th Int. Conf. Knowl. Sci., Eng. Manage. (KSEM)*, in Part II, Tokyo, Japan. New York, NY, USA: Springer-Verlag, 2021, pp. 176–190.

[153] S. R. Narang, M. K. Jindal, and M. Kumar, "Ancient text recognition: A review," *Artif. Intell. Rev.*, vol. 53, no. 8, pp. 5517–5558, 2020.

[154] C. Reul, S. Tomasek, F. Langhanki, and U. Springmann, "Open source handwritten text recognition on medieval manuscripts using mixed models and document-specific finetuning," in *Proc. Int. Workshop Document Anal. Syst.*, New York, NY, USA: Springer-Verlag, 2022, pp. 414–428.

[155] G. Leifert, R. Labahn, and T. Strauß, "CITlab ARGUS for Arabic handwriting," 2014, *arXiv:1412.6061*.

[156] T. Bluche, J. Louradour, and R. Messina, "Scan, attend and read: End-to-end handwritten paragraph recognition with mdlstm attention," in *Proc. 14th IAPR Int. Conf. Document Anal. Recognit. (ICDAR)*, vol. 1, Piscataway, NJ, USA: IEEE Press, 2017, pp. 1050–1055.

[157] D. Kass and E. Vats, "AttentionHTR: Handwritten text recognition based on attention encoder-decoder networks," in *Proc. Int. Workshop Document Anal. Syst.*, New York, NY, USA: Springer-Verlag, 2022, pp. 507–522.

[158] S. Tarride, A. Lemaitre, B. Coüasnon, and S. Tardivel, "A comparative study of information extraction strategies using an attention-based neural network," in *Proc. Int. Workshop Document Anal. Syst.*, New York, NY, USA: Springer-Verlag, 2022, pp. 644–658.

[159] L. Kang, P. Riba, M. Rusiñol, A. Fornés, and M. Villegas, "Pay attention to what you read: Non-recurrent handwritten text-line recognition," *Pattern Recognit.*, vol. 129, 2022, Art. no. 108766.

[160] Y. Nagata, J. Otao, D. Haraguchi, and S. Uchida, "Truetype transformer: Character and font style recognition in outline format," in *Proc. Int. Workshop Document Anal. Syst.*, New York, NY, USA: Springer-Verlag, 2022, pp. 18–32.

[161] B. Madi, R. Alaasam, A. Droby, and J. El-Sana, "HST-GAN: Historical style transfer GAN for generating historical text images," in *Proc. Int. Workshop Document Anal. Syst.*, New York, NY, USA: Springer-Verlag, 2022, pp. 523–537.

[162] A. Antonacopoulos and D. Bridson, "Performance analysis framework for layout analysis methods," in *Proc. 9th Int. Conf. Document Anal. Recognit. (ICDAR)*, vol. 2, Piscataway, NJ, USA: IEEE Press, 2007, pp. 1258–1262.

[163] A. Antonacopoulos, S. Pletschacher, D. Bridson, and C. Papadopoulos, "ICDAR 2009 page segmentation competition," in *Proc. 10th Int. Conf. Document Anal. Recognit.*, Piscataway, NJ, USA: IEEE Press, 2009, pp. 1370–1374.

[164] A. Antonacopoulos, C. Clausner, C. Papadopoulos, and S. Pletschacher, "Historical document layout analysis competition," in *Proc. Int. Conf. Document Anal. Recognit.*, Piscataway, NJ, USA: IEEE Press, 2011, pp. 1516–1520.

[165] A. Antonacopoulos, C. Clausner, C. Papadopoulos, and S. Pletschacher, "ICDAR 2013 competition on historical newspaper layout analysis (HNLA 2013)," in *Proc. 12th Int. Conf. Document Anal. Recognit.*, Piscataway, NJ, USA: IEEE Press, 2013, pp. 1454–1458.

[166] A. Antonacopoulos, C. Clausner, C. Papadopoulos, and S. Pletschacher, "ICDAR2015 competition on recognition of documents with complex layouts-RDCL2015," in *Proc. 13th Int. Conf. Document Anal. Recognit. (ICDAR)*, Piscataway, NJ, USA: IEEE Press, 2015, pp. 1151–1155.

[167] F. Simistira, M. Bouillon, M. Seuret, M. Würsch, M. Alberti, R. Ingold, and M. Liwicki, "ICDAR2017 competition on layout analysis for challenging medieval manuscripts," in *Proc. 14th IAPR Int. Conf. Document Anal. Recognit. (ICDAR)*, vol. 1, Piscataway, NJ, USA: IEEE Press, 2017, pp. 1361–1370.

[168] M. R. Yousefi, M. R. Soheili, T. M. Breuel, E. Kabir, and D. Stricker, "Binarization-free OCR for historical documents using LSTM networks," in *Proc. 13th Int. Conf. Document Anal. Recognit. (ICDAR)*, Piscataway, NJ, USA: IEEE Press, 2015, pp. 1121–1125.

[169] U. Springmann, D. Najock, H. Morgenroth, H. Schmid, A. Gotscharek, and F. Fink, "OCR of historical printings of Latin texts: Problems, prospects, progress," in *Proc. 1st Int. Conf. Digit. Access Textual Cultural Heritage*, 2014, pp. 71–75.

[170] A. Ul-Hasan, S. S. Bukhari, and A. Dengel, "OCRoRACT: A sequence learning OCR system trained on isolated characters," in *Proc. 12th IAPR Workshop Document Anal. Syst. (DAS)*, Los Alamitos, CA, USA: IEEE Comput. Soc. Press, 2016, pp. 174–179.

[171] M. Jenckel, S. S. Bukhari, and A. Dengel, "anyOCR: A sequence learning based OCR system for unlabeled historical documents," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Piscataway, NJ, USA: IEEE Press, 2016, pp. 4035–4040.

[172] S. S. Bukhari, A. Kadi, M. A. Jouneh, F. M. Mir, and A. Dengel, "anyOCR: An open-source OCR system for historical archives," in *Proc. 14th IAPR Int. Conf. Document Anal. Recognit. (ICDAR)*, vol. 1, Piscataway, NJ, USA: IEEE Press, 2017, pp. 305–310.

[173] T. Berg-Kirkpatrick and D. Klein, "Improved typesetting models for historical OCR," in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics (Short Papers)*, vol. 2, 2014, pp. 118–123.

[174] T. Berg-Kirkpatrick, G. Durrett, and D. Klein, "Unsupervised transcription of historical documents," in *Proc. 51st Annu. Meeting Assoc. Comput. Linguistics (Long Papers)*, vol. 1, 2013, pp. 207–217.

[175] M. R. Gupta, N. P. Jacobson, and E. K. Garcia, "OCR binarization and image pre-processing for searching historical documents," *Pattern Recognit.*, vol. 40, no. 2, pp. 389–397, 2007.

[176] A. Scius-Bertrand, S. Gabay, J. Janes, L. Petkovic, C. Corbieres, and T. Clérice, "The BIR database—Identifying typographic emphasis in list-like historical documents," in *Proc. 6th Int. Workshop Historical Document Imag. Process.*, 2021, pp. 37–42.

[177] C. Clausner, S. Pletschacher, and A. Antonacopoulos, "Aletheia—An advanced document layout and text ground-truthing system for production environments," in *Proc. Int. Conf. Document Anal. Recognit.*, Piscataway, NJ, USA: IEEE Press, 2011, pp. 48–52.

[178] C. Clausner, S. Pletschacher, and A. Antonacopoulos, "Efficient OCR training data generation with aletheia," in *Proc. Int. Assoc. Pattern Recognit. (IAPR)*, Tours, France, 2014, pp. 7–10.

[179] A. Dulla, "Data gathering and analysis for Arabic historical documents," *Int. J. Comput. Inf. Eng.*, vol. 13, no. 4, pp. 185–191, 2019.

[180] C. Papadopoulos, S. Pletschacher, C. Clausner, and A. Antonacopoulos, "The impact dataset of historical document images," in *Proc. 2nd Int. Workshop Historical Document Imag. Process.*, 2013, pp. 123–130.

[181] R. Smith, "An overview of the Tesseract OCR engine," in *Proc. 9th Int. Conf. Document Anal. Recognit. (ICDAR)*, vol. 2, Piscataway, NJ, USA: IEEE Press, 2007, pp. 629–633.

[182] S. Akhil, "An overview of Tesseract OCR engine," in A Seminar Report. Dept. Comput. Sci. Eng., Nat. Inst. Technol., Calicut Monsoon, 2016.

[183] N. Pawar, Z. Shaikh, P. Shinde, and Y. Warke, "Image to text conversion using Tesseract," *Image*, vol. 6, no. 2, pp. 516–519, 2019.

[184] B. Nunamaker, S. S. Bukhari, D. Borth, and A. Dengel, "A Tesseract-based OCR framework for historical documents lacking ground-truth text," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Piscataway, NJ, USA: IEEE Press, 2016, pp. 3269–3273.

[185] B. Kiessling, R. Tissot, P. Stokes, and D. S. B. Ezra, "eScriptorium: An open source platform for historical document analysis," in *Proc. Int. Conf. Document Anal. Recognit. Workshops (ICDARW)*, vol. 2, Piscataway, NJ, USA: IEEE Press, 2019, pp. 19–19.

[186] J. Martínek, L. Lenc, and P. Král, "Building an efficient OCR system for historical documents with little training data." *Neural Comput. Appli.*, vol. 32, pp. 17209–17227, Dec. 2020.

[187] T. Clérice, "You Actually Look Twice At it (YALTAi): Using an object detection approach instead of region segmentation within the Kraken engine," 2022, *arXiv:2207.11230*.

[188] N. Okinina, L. Nicolas, and V. Lyding, "Transc&Anno: A graphical tool for the transcription and on-the-fly annotation of handwritten documents," in *Proc. 11th Int. Conf. Lang. Resour. Eval. (LREC)*, 2018, pp. 701–705.

[189] K. Mika, J. De Veer, and C. Rinaldo, "Crowdsourcing natural history archives: Tools for extracting transcriptions and data," *Biodiversity Inform.*, vol. 12, pp. 58–75, Nov. 2017.

[190] M. Jarnecke, "Crowdsource indexing: Virtual volunteering means big index projects keep on moving at the Indiana state archives," *Hoosier Genealogist Connections*, vol. 60, no. 2, pp. 62–68, 2020.

[191] P. Kahle, S. Colutto, G. Hackl, and G. Mühlberger, "Transkribus—A service platform for transcription, recognition and retrieval of historical documents," in *Proc. 14th IAPR Int. Conf. Document Anal. Recognit. (ICDAR)*, vol. 4, Piscataway, NJ, USA: IEEE Press, 2017, pp. 19–24.

[192] G. Muehlberger et al., "Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study," *J Documentation*, vol. 75, no. 5, Jul. 2019, pp. 954–976.

[193] S. Colutto, P. Kahle, H. Guenter, and G. Mühlberger, "Transkribus. A platform for automated text recognition and searching of historical documents," in *Proc. 15th Int. Conf. eScience (eScience)*, Piscataway, NJ, USA: IEEE Press, 2019, pp. 463–466.

[194] A. Kokaze, "Using Transkribus to transcribe eighteenth-century French historical manuscripts," *Historical Stud. Western World*, vol. 1, 2022, Art. no. 9.

[195] N. Milioni, "Automatic transcription of historical documents: Transkribus as a tool for libraries, archives and scholars," dissertation, 2020. [Online]. Available: https://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-412565

[196] W. Thompson et al., "Using handwritten text recognition (HTR) tools to transcribe historical multilingual lexica," *Scripta e-Scripta*, vol. 21, pp. 217–231, 2021.

[197] T. M. Kina, R. Vladimir, M. M. Ghaffar, J. A. Varela, W. Norbert, and A. Dengel, "iDocChip: A configurable hardware architecture for historical document image processing," *Int. J. Parallel Program.*, vol. 49, no. 2, pp. 253–284, 2021.

[198] V. Rybalkin, S. S. Bukhari, M. M. Ghaffar, A. Ghafoor, N. Wehn, and A. Dengel, "iDocChip: A configurable hardware architecture for historical document image processing: Percentile based binarization," in *Proc. ACM Symp. Document Eng.*, 2018, pp. 1–8.

[199] D. Pérez, L. Tarazón, N. Serrano, F. Castro, O. R. Terrades, and A. Juan, "The GERMANA database," in *Proc. 10th Int. Conf. Document Anal. Recognit.*, 2009, pp. 301–305.

[200] A. Fischer et al., "Automatic transcription of handwritten medieval documents," in *Proc. 15th Int. Conf. Virtual Syst. Multimedia*, Piscataway, NJ, USA: IEEE Press, 2009, pp. 137–142.

[201] N. Serrano, F. Castro, and A. Juan, "The RODRIGO database," in *Proc. 7th Int. Conf. Lang. Resour. Eval. (LREC)*, 2010, pp. 19–21.

[202] A. Alaei, U. Pal, and P. Nagabhushan, "A new scheme for unconstrained handwritten text-line segmentation," *Pattern Recognit.*, vol. 44, no. 4, pp. 917–928, 2011.

[203] A. Fischer, V. Frinken, A. Fornés, and H. Bunke, "Transcription alignment of Latin manuscripts using hidden Markov models," in *Proc. Workshop Historical Document Imag. Process.*, 2011, pp. 29–36.

[204] A. Fischer, A. Keller, V. Frinken, and H. Bunke, "Lexicon-free handwritten word spotting using character HMMs," *Pattern Recognit. Lett.*, vol. 33, no. 7, pp. 934–942, 2012.

[205] A. Alaei, U. Pal, and P. Nagabhushan, "Dataset and ground truth for handwritten text in four different scripts," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 26, no. 4, 2012, Art. no. 1253001.

[206] V. Romero et al., "The ESPOSALLES database: An ancient marriage license corpus for off-line handwriting recognition," *Pattern Recognit.*, vol. 46, no. 6, pp. 1658–1669, 2013.

[207] D. Fernández-Mota, J. Almazán, N. Cirera, A. Fornés, and J. Lladós, "BH2M: The Barcelona historical, handwritten marriages database," in *Proc. 22nd Int. Conf. Pattern Recognit.*, Piscataway, NJ, USA: IEEE Press, 2014, pp. 256–261.

[208] W. Pantke, M. Dennhardt, D. Fecker, V. Märgner, and T. Fingscheidt, "An historical handwritten Arabic dataset for segmentation-free word spotting-HADARA80P," in *Proc. 14th Int. Conf. Frontiers Handwriting Recognit.*, Piscataway, NJ, USA: IEEE Press, 2014, pp. 15–20.

[209] C. Clausner, C. Papadopoulos, S. Pletschacher, and A. Antonacopoulos, "The ENP image and ground truth dataset of historical newspapers," in *Proc. 13th Int. Conf. Document Anal. Recognit. (ICDAR)*, Piscataway, NJ, USA: IEEE Press, 2015, pp. 931–935.

[210] B. Gatos et al., "GRPOLY-DB: An old Greek polytonic document image database," in *Proc. 13th Int. Conf. Document Anal. Recognit. (ICDAR)*, Piscataway, NJ, USA: IEEE Press, 2015, pp. 646–650.

[211] M. Mehri, P. Héroux, R. Mullot, J.-P. Moreux, B. Coüasnon, and B. Barrett, "HBA 1.0: A pixel-based annotated dataset for historical book analysis," in *Proc. 4th Int. Workshop Historical Document Imag. Process.*, 2017, pp. 107–112.

[212] C. Clausner, A. Antonacopoulos, T. Derrick, and S. Pletschacher, "ICDAR2017 competition on recognition of early Indian printed documents—REID2017," in *Proc. 14th IAPR Int. Conf. Document Anal. Recognit. (ICDAR)*, vol. 1, 2017, pp. 1411–1416.

[213] T. Grüning, R. Labahn, M. Diem, F. Kleber, and S. Fiel, "READ-BAD: A new dataset and evaluation scheme for baseline detection in archival documents," in *Proc. 13th IAPR Int. Workshop Document Anal. Syst. (DAS)*, Piscataway, NJ, USA: IEEE Press, 2018, pp. 351–356.

[214] A. Dulla, "A dataset of warped historical Arabic documents," in *9th Int. Conf. Pattern Recognit. Syst. (ICPRS 2018)*, 2018, p. 6.

[215] C. Clausner, A. Antonacopoulos, N. Mcgregor, and D. Wilson-Nunn, "ICFHR 2018 competition on recognition of historical Arabic scientific manuscripts—RASM2018," in *Proc. 16th Int. Conf. Frontiers Handwriting Recognit. (ICFHR)*, Piscataway, NJ, USA: IEEE Press, 2018, pp. 471–476.

[216] L. Quirós et al., "Oficio de Hipotecas de Girona. A dataset of Spanish notarial deeds (18th century) for handwritten text recognition and layout analysis of historical documents." Zenodo. Accessed: Jan. 18, 2024. [Online]. Available: https://zenodo.org/records/1322666#.Ypi6Ty8RoUE

[217] B. K. Barakat, J. El-Sana, and I. Rabaev, "The Pinkas dataset," in *Proc. Int. Conf. Document Anal. Recognit. (ICDAR)*, Piscataway, NJ, USA: IEEE Press, 2019, pp. 732–737.

[218] B. Kiessling, D. S. B. Ezra, and M. T. Miller, "BADAM: A public dataset for baseline detection in Arabic-script manuscripts," in *Proc. 5th Int. Workshop Historical Document Imag. Process.*, 2019, pp. 13–18.

[219] M. Boillet, M.-L. Bonhomme, D. Stutzmann, and C. Kermorvant, "HORAE: An annotated dataset of books of hours," in *Proc. 5th Int. Workshop Historical Document Imag. Process.*, 2019, pp. 7–12.

[220] C. Clausner, A. Antonacopoulos, and S. Pletschacher, "PRImA—DMAS2019." PRImA Research Lab. Accessed: Jan. 18, 2024. [Online]. Available: https://www.primaresearch.org/DMAS2019/

[221] Y. Xu, F. Yin, D.-H. Wang, X.-Y. Zhang, Z. Zhang, and C.-L. Liu, "CASIA-AHCDB: A large-scale Chinese ancient handwritten characters database," in *Proc. Int. Conf. Document Anal. Recognit. (ICDAR)*, Piscataway, NJ, USA: IEEE Press, 2019, pp. 793–798.

[222] R. Saini, D. Dobson, J. Morrey, M. Liwicki, and F. S. Liwicki, "ICDAR 2019 historical document reading challenge on large structured Chinese family records," in *Proc. Int. Conf. Document Anal. Recognit. (ICDAR)*, Piscataway, NJ, USA: IEEE Press, 2019, pp. 1499–1504.

[223] C. Clausner, A. Antonacopoulos, T. Derrick, and S. Pletschacher, "ICDAR2019 competition on recognition of early Indian printed documents—REID2019," in *Proc. Int. Conf. Document Anal. Recognit. (ICDAR)*, Piscataway, NJ, USA: IEEE Press, 2019, pp. 1527–1532.

[224] C. Clausner, A. Antonacopoulos, and S. Pletschacher, "PRImA—RASM2019." PRImA Research Lab. Accessed: Jan. 18, 2024. [Online]. Available: https://www.primaresearch.org/RASM2019/

[225] L. Quirós, M. Kallio, and E. Vidal, "Finnish court records-sub500. A dataset of Finnish notarial records (19th century)." Zenodo. Accessed: Jan. 18, 2024. [Online]. Available: https://zenodo.org/records/3945088

[226] H. J. Dolfing, J. Bellegarda, J. Chorowski, R. Marxer, and A. Laurent, "The "ScribbleLens" Dutch historical handwriting corpus," in *Proc. 17th Int. Conf. Frontiers Handwriting Recognit. (ICFHR)*, Piscataway, NJ, USA: IEEE Press, 2020, pp. 67–72.

[227] Z. Shen, K. Zhang, and M. Dell, "A large dataset of historical Japanese documents with complex layouts," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2020, pp. 548–549.

[228] D. Stoekl Ben Ezra, B. Brown-DeVost, P. Jablonski, H. Lapin, B. Kiessling, and E. Lolli, "BiblIA—A general model for medieval Hebrew manuscripts and an open annotated dataset," in *Proc. 6th Int. Workshop Historical Document Imag. Process.*, 2021, pp. 61–66.

[229] M. Maarand, Y. Beyer, A. Kåsen, K. T. Fosseide, and C. Kermorvant, "A comprehensive comparison of open-source libraries for handwritten text recognition in Norwegian," in *Proc. Int. Workshop Document Anal. Syst.*, New York, NY, USA: Springer-Verlag, 2022, pp. 399–413.

[230] N. Girdhar, M. Coustaty, and A. Doucet, "Benchmarking NAS for article separation in historical newspapers," in *Proc. Int. Conf. Asian Digit. Libraries*, New York, NY, USA: Springer-Verlag, 2023, pp. 76–88.

[231] M. Mehri, N. Nayef, P. Héroux, P. Gomez-Krämer, and R. Mullot, "Learning texture features for enhancement and segmentation of historical document images," in *Proc. 3rd Int. Workshop Historical Document Imag. Process.*, 2015, pp. 47–54.

[232] R. Saabni, A. Asi, and J. El-Sana, "Text line extraction for historical document images," *Pattern Recognit. Lett.*, vol. 35, pp. 23–33, 2014.

[233] F. Shafait, D. Keysers, and T. M. Breuel, "Pixel-accurate representation and evaluation of page segmentation in document images," in *Proc. 18th Int. Conf. Pattern Recognit. (ICPR'06)*, vol. 1, Piscataway, NJ, USA: IEEE Press, 2006, pp. 872–875.

[234] T. A. Tran, I. S. Na, and S. H. Kim, "Page segmentation using minimum homogeneity algorithm and adaptive mathematical morphology," *Int. J. Document Anal. Recognit. (IJDAR)*, vol. 19, no. 3, pp. 191–209, 2016.

[235] M. Alberti, M. Bouillon, R. Ingold, and M. Liwicki, "Open evaluation tool for layout analysis of document images," in *Proc. 14th IAPR Int. Conf. Document Anal. Recognit. (ICDAR)*, vol. 4, Piscataway, NJ, USA: IEEE Press, 2017, pp. 43–47.

[236] A. Antonacopoulos, B. Gatos, and D. Karatzas, "ICDAR2003 page segmentation competition," in *Proc. Seventh Int. Conf. Doc. Anal. Recognit. 2003*, Edinburgh, UK: IEEE Computer Society Press, Aug 2003, pp. 688–692.

[237] B. Gatos, N. Stamatopoulos, and G. Louloudis, "ICFHR 2010 handwriting segmentation contest," in *Proc. 12th Int. Conf. Frontiers Handwriting Recognit.*, Piscataway, NJ, USA: IEEE Press, 2010, pp. 737–742.

[238] G. Louloudis, B. Gatos, I. Pratikakis, and C. Halatsis, "Text line and word segmentation of handwritten documents," *Pattern Recognit.*, vol. 42, no. 12, pp. 3169–3183, 2009.

[239] S. S. Bukhari, F. Shafait, and T. M. Breuel, "Script-independent handwritten textlines segmentation using active contours," in *Proc. 10th Int. Conf. Document Anal. Recognit.*, Piscataway, NJ, USA: IEEE Press, 2009, pp. 446–450.

[240] I. T. Phillips and A. K. Chhabra, "Empirical performance evaluation of graphics recognition systems," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, no. 9, pp. 849–870, Sep. 1999.

[241] A. Garz and R. Sablatnig, "Multi-scale texture-based text recognition in ancient manuscripts," in *Proc. 16th Int. Conf. Virtual Syst. Multimedia*, Piscataway, NJ, USA: IEEE Press, 2010, pp. 336–339.

[242] A. Garz, R. Sablatnig, and M. Diem, "Layout analysis for historical manuscripts using sift features," in *Proc. Int. Conf. Document Anal. Recognit.*, Piscataway, NJ, USA: IEEE Press, 2011, pp. 508–512.

[243] T. Grüning, G. Leifert, T. Strauß, J. Michael, and R. Labahn, "A two-stage method for text line detection in historical documents," *Int. J. Document Anal. Recognit. (IJDAR)*, vol. 22, no. 3, pp. 285–302, 2019.

[244] M. Mehri, P. Héroux, P. Gomez-Krämer, and R. Mullot, "Texture feature benchmarking and evaluation for historical document image analysis," *Int. J. Document Anal. Recognit. (IJDAR)*, vol. 20, no. 1, pp. 1–35, 2017.

[245] N. Nikolaou, M. Makridis, B. Gatos, N. Stamatopoulos, and N. Papamarkos, "Segmentation of historical machine-printed documents using adaptive run length smoothing and skeleton segmentation paths," *Image Vis. Comput.*, vol. 28, no. 4, pp. 590–604, 2010.

[246] N. Arvanitopoulos and S. Süsstrunk, "Seam carving for text line extraction on color and grayscale historical manuscripts," in *Proc. 14th Int. Conf. Frontiers Handwriting Recognit.*, Piscataway, NJ, USA: IEEE Press, 2014, pp. 726–731.

[247] A. Asi, R. Cohen, K. Kedem, and J. El-Sana, "Simplifying the reading of historical manuscripts," in *Proc. 13th Int. Conf. Document Anal. Recognit. (ICDAR)*, Piscataway, NJ, USA: IEEE Press, 2015, pp. 826–830.

[248] R. Cohen, I. Dinstein, J. El-Sana, and K. Kedem, "Using scale-space anisotropic smoothing for text line extraction in historical documents," in *Proc. Int. Conf. Image Anal. Recognit.*, New York, NY, USA: Springer-Verlag, 2014, pp. 349–358.

[249] I. Rabaev, O. Biller, J. El-Sana, K. Kedem, and I. Dinstein, "Text line detection in corrupted and damaged historical manuscripts," in *Proc. 12th Int. Conf. Document Anal. Recognit.*, Piscataway, NJ, USA: IEEE Press, 2013, pp. 812–816.

[250] W. Swaileh, K. A. Mohand, and T. Paquet, "Multi-script iterative steerable directional filtering for handwritten text line extraction," in *Proc. 13th Int. Conf. Document Anal. Recognit. (ICDAR)*, Piscataway, NJ, USA: IEEE Press, 2015, pp. 1241–1245.

[251] T. A. Tran, I.-S. Na, and S.-H. Kim, "Hybrid page segmentation using multilevel homogeneity structure," in *Proc. 9th Int. Conf. Ubiquitous Inf. Manage. Commun.*, 2015, pp. 1–6.

[252] A. Fischer, M. Baechler, A. Garz, M. Liwicki, and R. Ingold, "A combined system for text line extraction and handwriting recognition in historical documents," in *Proc. 11th IAPR Int. Workshop Document Anal. Syst.*, Piscataway, NJ, USA: IEEE Press, 2014, pp. 71–75.

[253] Z. Shi, S. Setlur, and V. Govindaraju, "A steerable directional local profile technique for extraction of handwritten Arabic text lines," in *Proc. 10th Int. Conf. Document Anal. Recognit.*. Piscataway, NJ, USA: IEEE Press, 2009, pp. 176–180.

[254] Y. Lu, Z. Wang, and C. L. Tan, "Word grouping in document images based on Voronoi tessellation," in *Proc. Int. Workshop Document Anal. Syst.*, New York, NY, USA: Springer-Verlag, 2004, pp. 147–157.

[255] F. Shafait, J. Van Beusekom, D. Keysers, and T. M. Breuel, "Background variability modeling for statistical layout analysis," in *Proc. 19th Int. Conf. Pattern Recognit.*, Piscataway, NJ, USA: IEEE Press, 2008, pp. 1–4.

[256] A. Asi, R. Saabni, and J. El-Sana, "Text line segmentation for gray scale historical document images," in *Proc. Workshop Historical Document Imag. Process.*, 2011, pp. 120–126.

[257] J.-Y. Ramel, S. Leriche, M.-L. Demonet, and S. Busson, "User-driven page layout analysis of historical printed books," *Int. J. Document Anal. Recognit. (IJDAR)*, vol. 9, no. 2, pp. 243–261, 2007.

[258] S. S. Bukhari, M. I. A. Al Azawi, F. Shafait, and T. M. Breuel, "Document image segmentation using discriminative learning over connected components," in *Proc. 9th IAPR Int. Workshop Document Anal. Syst.*, 2010, pp. 183–190.

[259] S. S. Bukhari, F. Shafait, and T. M. Breuel, "Improved document image segmentation algorithm using multiresolution morphology," in *Proc. Document Recognit. Retrieval XVIII*, vol. 7874, San Francisco Airport, CA, United States: SPIE, 2011, pp. 109–116.

[260] M. Alberti, M. Seuret, V. Pondenkandath, R. Ingold, and M. Liwicki, "Historical document image segmentation with LDA-initialized deep neural networks," in *Proc. 4th Int. Workshop Historical Document Imag. Process.*, 2017, pp. 95–100.

[261] S. Capobianco, L. Scommegna, and S. Marinai, "Historical handwritten document segmentation by using a weighted loss," in *Proc. IAPR Workshop Artif. Neural Netw. Pattern Recognit.*, New York, NY, USA: Springer-Verlag, 2018, pp. 395–406.

[262] M. Kotti, V. Moschou, and C. Kotropoulos, "Speaker segmentation and clustering," *Signal Process.*, vol. 88, no. 5, pp. 1091–1124, 2008.

[263] A. M. Namboodiri and A. K. Jain, "Document structure and layout analysis," in *Digital Document Processing.* New York, NY, USA: Springer-Verlag, 2007, pp. 29–48.

[264] F. Kleber, R. Sablatnig, M. Gau, and H. Miklas, "Ancient document analysis based on text line extraction," in *Proc. 19th Int. Conf. Pattern Recognit.*, Piscataway, NJ, USA: IEEE Press, 2008, pp. 1–4.