# Unsupervised approach to text line extraction in Belfort civil registers of births

Wissam Alkendi, Franck Gechter, Laurent Heyberger, Christophe Guyeux

# Unsupervised approach to text line extraction in Belfort civil registers of births

Wissam AlKendi[1*], Franck Gechter[1,4†], Laurent Heyberger[2†], Christophe Guyeux[3]

[1*]CIAD, UTBM, UMR 7533, Belfort, F-90010, Belfort, France.
[2]FEMTO-ST Institute/RECITS, UTBM, UMR 6174 CNRS, Belfort, F-90010, Belfort, France.
[3]FEMTO-ST Institute/DISC, Université de Franche-Comté, UMR 6174 CNRS, Belfort, F-90016, Belfort, France.
[4]LORIA, SIMBIOT Team, Université de Lorraine, UMR 7503, Vandoeuvre-lès-Nancy, F-54506, Meurthe-et-Moselle, France.

*Corresponding author(s). E-mail(s): wissam.al-kendi@utbm.fr;
Contributing authors: franck.gechter@utbm.fr; laurent.heyberger@utbm.fr;
christophe.guyeux@univ-fcomte.fr;
†These authors contributed equally to this work.

**Abstract**

Historical documents are invaluable resources for understanding the development of civilizations and cultures. However, the transcription process of these documents comprises many challenges such as complex layouts, degradation, various handwritten styles, and skewed text. This paper presents an unsupervised approach for text line extraction in the Belfort Civil Registers of Births, a historical dataset containing a mix of printed and handwritten text with marginal annotations. The proposed method employs a series of image processing techniques to identify text line cores. The method also utilizes a dynamic gap identification and segment point localization strategy based on text density and histogram analysis to effectively identify the borders of the text lines in polygon shape. An XML file generation tool is then utilized to structure the resulting components and link them with their corresponding text. The method exhibits competitive accuracy in segmenting text lines on both the Belfort dataset and standard benchmarks such as the Saint Gall and READ Bozen datasets. This work contributes to the preservation and accessibility of historical documents by facilitating accurate transcription and structured data representation.

**Keywords:** Text Line Extraction, Historical Documents, Handwritten Text Recognition, Unsupervised Approach, Belfort Civil Registers of Birth, Structured Data Representation

## 1 Introduction

Historical documents are one of the most important sources that highlight the past of humanity and contribute to understanding the development of civilizations and cultures throughout the ages. Researchers increasingly rely on historical documents to extract and analyze valuable information, providing deeper insights into past events and exploring the evolving connection between

traditional records and modern digital interpretations [1].

Machine Learning (ML) techniques are important and central to the analysis and comprehension of these historical documents. These techniques are capable of processing large volumes of historical data very quickly and efficiently. Consequently, ML can easily recognize complex interrelations between various historical events and personalities thereby illuminating vague or hidden aspects of history [2]. However, the process of training machine learning models to recognize the text within the historical documents also requires extensive data due to several challenges including text style variation, complex document layout, text skew, and degradation.

The segmentation of historical document images into paragraphs [3], text lines [4], and word images is considered a crucial stage in the machine learning model training process [5, 6]. Many authors have proposed segmentation techniques for various text recognition applications, such as keyword spotting [7], writer identification [8, 9], and data extraction. The majority of these techniques rely on employing word and text line segmentation processes. Thus, it is an essential step in text recognition processing. It has been verified that employing a robust segmentation scheme leads to higher accuracy rates.

Historical document images may involve machine-printed text, handwritten text, or a combination of both. Segmentation of machine-printed text images into words or text line images can be performed relatively easily. However, segmenting handwritten text images or hybrid (containing both printed and handwritten text) images remains a challenging task due to several impediments. These include angular and spiky letters, and ornate flourishes that result in overlapped words and text lines within the image. Additionally, challenges arise from the diversity of writing styles, spots on the paper, noise, and text misalignment.

Unsupervised approaches such as clustering, projection profiles, and contour analysis do not require pre-labeled data to identify text regions within the document images. These methods are designed to infer structure and patterns without prior training [10], making them highly adaptable and capable of operating with minimal preparation.

However, these approaches encounter challenges such as attaining high accuracy in noisy or complex documents, especially in the presence of overlapping text lines, varying text orientations, or poor-quality documents [11]. Another significant challenge is the tuning of parameters and thresholds dynamically based on the input image. Despite these challenges, the flexibility and robustness of unsupervised methods make them highly valuable and versatile tools for text line extraction tasks.

Gap classification is one of the most common segmentation techniques. It is based on identifying gaps (distance) between text lines or word images, where a gap is labeled as inter or intra-word gap [12–14]. This process calculates the distribution of the text using various methods such as local/global thresholding [15] and the Gaussian Mixture Model [16]. Additionally, authors have proposed a supervised learning-based gap classification technique [17] to aid in the identification and separation of text data, facilitating better document understanding and analysis.

In the 1911 census, Belfort was only the fifty-third-largest city in France. However, despite its modest size, the département of Territoire de Belfort was one of the most industrialized regions in France at the beginning of the 20th century, with the fourth-highest population density (166.6 inhabitants/km$^2$). This placed it behind only the départements of Seine (Paris), Nord (Lille), and Rhône (Lyon), but ahead of Pas-de-Calais. Between 1881 and 1911, following the establishment of SACM (mechanical engineering, the precursor to Alstom, which today produces TGV high-speed trains) and DMC (a leading global textile manufacturer) in 1879, Belfort experienced the fourth-highest increase in population density among French départements, trailing only the same three major industrial regions. This city also ranked second to Seine (Paris, 0.4%) for the low proportion of sparsely populated areas (5.8%), underscoring its concentrated industrial workforce.

Belfort's strong industrial growth, coupled with its location near the borders of Germany and Switzerland, made it a destination for foreign immigrants. Despite its relatively small population, it was one of 17 French départements with the highest number of foreign residents (10,778) at the start of the 20th century. This influx was

partly due to a strong migration of Alsatians after 1871, as well as a wave of Italian immigrants before World War I. By 1911, Belfort also led the nation in naturalized immigrants per 10,000 inhabitants (349 for men and 451 for women). Moreover, Belfort's military presence contributed to another demographic peculiarity: it had the second-highest number of single men of marriageable and childbearing age (18-59) in France, largely due to the numerous barracks established in the city in preparation for a potential conflict with Germany [18–20].

These demographic and social factors likely contributed to a high frequency of out-of-wedlock births, particularly among the working-class population. This could be explained by several factors: the traditional historiographical view that working-class men and women were less influenced by religious precepts, housing shortages that led to cohabitation and subletting (increasing the likelihood of out-of-wedlock births), and the prevalence of exploitative practices such as droit de cuissage in factories at the time, particularly towards domestic servants. Such practices were exacerbated by the extreme social inequalities in France during this period. Additionally, the presence of a large male population, primarily soldiers, fueled the growth of prostitution in Belfort from the 1880s onward. By 1912, according to a report by the Belfort municipality (which should be treated cautiously), the town had as many as 700 clandestine prostitutes for a population of 32,000. All of these factors likely contributed to the rise of out-of-wedlock births, which can be traced through civil birth registers—an essential source for historians studying this phenomenon in Belfort.

In fact, the rate of out-of-wedlock births in Belfort was relatively low (2-7%) before 1870, similar to other small French towns at the time. However, by the 1870s, this rate surged, reaching 17.6%, significantly higher than the rate in Roubaix (13.3%), a city often referred to as the "French Manchester," despite Roubaix having three times the population of Belfort. In general, the rate of out-of-wedlock births tended to correlate with city size: it peaked in Paris (26.3% in 1876) and reached 21% in Lyon (1885). Belfort, however, presented a unique case. This exceptional situation makes a thorough examination of Belfort's civil birth registers indispensable, as it allows for cross-referencing with the DMC personnel records—Belfort's largest employer of women. By doing so, we can better understand the life trajectories of unmarried mothers and assess their degree of agency within the intersections of sexual, social, and religious pressures [21]. Figure 1 depicts an example document from these civil registers of birth.

To examine and address these birth declarations concerns, it is essential to establish a comprehensive knowledge database enabling complete investigation and resolution. Text recognition of these declaration images presents numerous challenges, including document layouts, reading orders, hybrid (printed and handwritten text), marginal mentions, skewness, degradation, and diverse text styles. Despite notable progress in Optical Character Recognition (OCR) techniques [22, 23], these attempts have yet to adequately address the complexities associated with transcribing such declaration images. Thus, a robust segmentation approach is necessary to ensure precise transcription.

In this paper, we propose an unsupervised approach based on gaps to segment French Belfort civil registers of birth (BCRB) document images into text line images. Additionally, we develop structured data representations linking the segmented text line images with their corresponding transcriptions using generated XML files. The segmentation process comprises four main stages: first, text line detection. Second, gaps identification. Third, gaps analysis and segment points determination. Finally, identification of text line image borders. This aims to address the challenges of the segmentation phase associated with transcribing the BCRB, thus enhancing the preservation and accessibility of these historical registers. Figure 2 illustrates the general pipeline for the segmentation process.

The proposed unsupervised method is simple and does not require an extensive preprocessing phase or a machine learning model. Additionally, there is no need for text line image skew correction. Moreover, many parameters values are calculated dynamically based on the input image, eliminating the need for page annotation or a training dataset. Furthermore, it is robust and independent of language or font-specific dependencies. Thus, it can be employed with French and other languages.

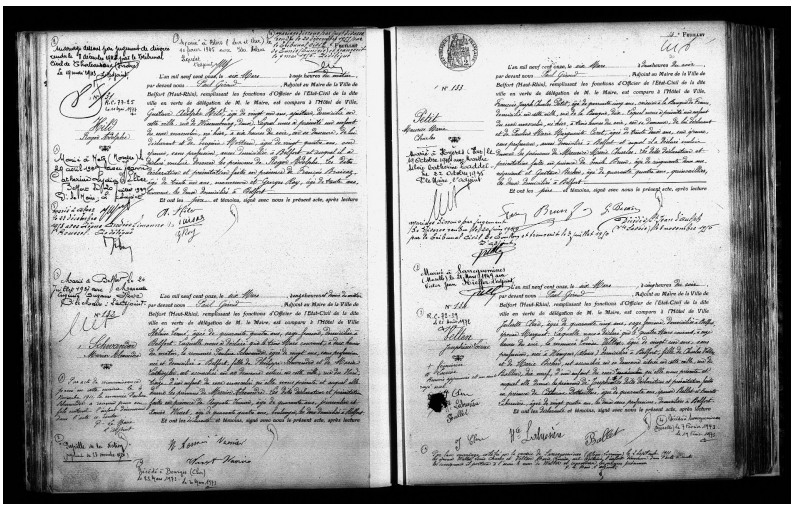


**Fig. 1** Sample of Belfort civil registers of births

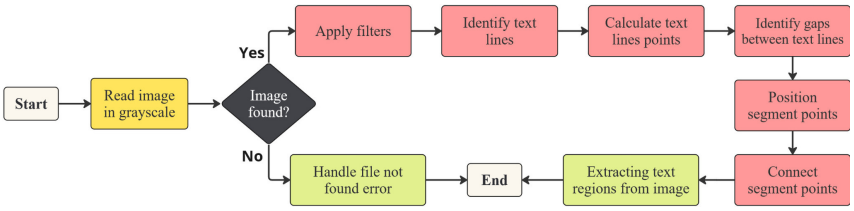



**Fig. 2** General pipeline of the proposed segmentation process, illustrating the steps from image acquisition to text extraction.

The structure of the paper is as follows: Section 2 highlights recent research in the field. Section 3 discusses the characteristics and challenges of the BCRB and the stages of the proposed method. Section 4 summarizes the experimental results. Lastly, Section 5 concludes the paper and proposes future research avenues.

## 2 State of the Art Overview

Segmentation involves dividing the text image into letters, words, lines, or paragraphs, often using the image's pixel characteristics. Numerous research works on document image segmentation have been proposed in recent years. These comprise text line and word image level segmentation on both machine-printed and handwritten text document images. These works include employing various methods to facilitate the segmentation process, including:

- **Thresholding:** This technique involves separating the text image into foreground (text) and background based on intensity thresholding [24–26]. Several thresholding techniques have been proposed and evaluated using a variety of challenging document images, including historical documents, newspapers, forms, and cheques, exhibiting the strengths and limitations of global versus local thresholding techniques in document image analysis [27]. Authors of [28] utilized an adaptive thresholding technique to enhance the document image, with affine-invariant texture analysis for text area segmentation. The method involves iteratively adjusting light intensities by employing iterative gamma correction followed by contrast stretching. This resulted in text areas being distinctly highlighted against background clutter, setting a robust basis for further analysis.
- **Edge Detection:** Detecting edges can assist in identifying text regions. Authors proposed several techniques for edge detection such as Sobel, Prewitt [29], Canny, Laplacian of Gaussian (LoG) [30], and Zero Crossing edge detection [31]. Authors of [32] presented a novel

approach to scene text detection based on the original Canny edge detector. This approach utilizes double thresholding and hysteresis tracking to identify texts with various confidence levels and classify them as strong, weak, or non-text. Strong candidates are selected, while weak ones are filtered through hysteresis tracking based on features such as proximity, size, and color, allowing for effective localization of a wide array of text types across different languages and image conditions.

- **Connected Component Analysis (CCA):** This technique determines connected regions in an image and considers them as potential text regions [33]. It is beneficial for both printed and handwritten text. Authors of [34] presented a method for text line-level segmentation in historical Tibetan manuscripts based on a connected components analysis technique. After a pre-processing phase for skew correction and baseline identification, a text location analysis process is employed to identify text line boundaries and classify connected components based on their interaction with an optimal segmentation line. Lastly, uncertain components are analyzed by their shape and location to correctly merge and allocate them within the text region boundaries, forming complete text lines.
- **Clustering:** Clustering techniques such as K-means or DBSCAN [35] can be utilized to allocate pixels into text and non-text clusters based on features such as color, texture, or gradient information.
  Authors of [36] presented a hybrid clustering algorithm comprising two stages to attain improved accuracy rates with reduced time complexity. In the first stage, the K-means algorithm is utilized to split text resources into smaller clusters. Additionally, the Canopy algorithm [37] is employed to determine the optimal selection of K-means parameters, such as the number of clusters (k) and initial clustering centers, to overcome the randomness and difficulties associated with predefined k values. In the second stage, the algorithm employs a hierarchical agglomeration clustering algorithm [38, 39] to combine the previously generated clusters into one cluster tree, providing an effective solution to handle large text images with complex structures.
- **Machine Learning-Based Segmentation:** In recent years, many authors have proposed training machine learning models such as Support Vector Machine (SVM) [14], Random Forest [40], and Convolutional Neural Networks (CNNs) to automatically segment text regions within documents. However, this approach requires appropriate annotated training data to achieve high-accuracy segmentation results.
  In [41], authors proposed a novel approach for extracting text lines from historical documents based on Convolutional Neural Networks (CNNs). The method applies layout analysis to classify each pixel as part of text blocks, background, or graphics. Subsequently, a second CNN filters these text blocks to generate the Main Body Area (MBA) map, which is then segmented to extract text lines using a region-based technique.
- **Deep Learning Approaches:** Deep learning architectures such as Fully Convolutional Networks (FCNs) or U-Net [4] can be employed for end-to-end text segmentation. These models are capable of learning complex representations from data and have reported highly beneficial accuracy results for both printed and handwritten text segmentation. In [42], authors proposed a technique for text line segmentation of historical document images based on line masks. These masks are predicted using a Fully Convolutional Network (FCN) that analyzes connected components on the same text line. After preprocessing steps, which include adaptive binarization and manual annotation of line masks on the document images, the FCN is trained on binarized document images labeled with line masks. The architecture of the FCN is based on the VGG 16-layer network, modified to suit the characteristics of the documents. This includes adjustments to the input size and output channels to account for the number of classes (text line or background), leading to promising results within the field.
- **Hybrid Approaches:** Authors also attempt to combine multiple techniques or models to enhance the accuracy of segmentation. For example, they use a combination of thresholding and Connected Component Analysis (CCA), or integrate traditional methods with deep learning for enhanced performance [43–46]

### 2.1 Segmentation-based gap classification

Several authors proposed numerous methods for extracting text images based on gaps around the text. In [47], two innovative segmentation approaches employing the Viterbi algorithm and Support Vector Machines were proposed for effective text line and word segmentation in handwritten documents. The approach of text line segmentation involves dividing the document image into vertical zones to identify text and gap areas using smoothed projection profiles. A significant innovation is the use of the Viterbi algorithm to improve the initial set of text and gap areas, followed by a method for drawing text-line separators across the document. Connected components (CCs) are subsequently assigned to text lines, with special consideration for CCs that span multiple lines or comprise ascenders/descenders, ensuring appropriate segmentation even in complex layouts.

Word segmentation is carried out utilizing a gap metric based on the objective function of a soft-margin linear SVM, which distinguishes between consecutive connected components. This metric is combined with a threshold defined by the probability density function of gap metric values across the document page to categorize gaps as "within" or "between" words. The work on word segmentation is expanded by [12] to enhance the method by employing local spatial features, leveraging a gap metric derived from the objective function of a soft-margin linear Support Vector Machine that separates consecutively connected components. The gap metrics are initially classified as "within" or "between" word classes utilizing a global threshold, and by examining local features for each pair of CCs in a text line, including the margin and slope of the linear classifier, leading to a more precise classification.

Other authors divided the document images into subsets of connected components and employed the Hough transform for text line segmentation, enhancing the outcomes with post-processing skeletonization. Additionally, word segmentation was also carried out using a fusion of convex and Euclidean distance metrics within a Gaussian mixture modeling framework to differentiate between intra-word and inter-word distances [16]. Table 1 provides a detailed exposition of State-of-the-Art approaches.

## 3 Methodology

### 3.1 Belfort Civil Registers of Births

The civil registers of the Belfort commune comprise 39,627 birth declarations written in French and scanned at a resolution of 300 dpi each. They consist of two types of declarations: first, handwritten declarations; second, hybrid declarations (partially printed), with blank spaces left for filling in specific information about the child statement. Each declaration provides information such as the child's name, parent's name, date of birth, and other details. Table 2 illustrates the structure and content of these declarations.

The registers have Gregorian dates ranging from 1807 to 1919. They are available online up to 1902 through the following link: https://archives.belfort.fr/search/form/e5a0c07e-9607-42b0-9772-f19d7bfa180e (accessed on May 08, 2024). Moreover, we have clearance from the municipal archives to review the registers up to 1919. Figure 3 shows an example declaration from the civil birth registers.

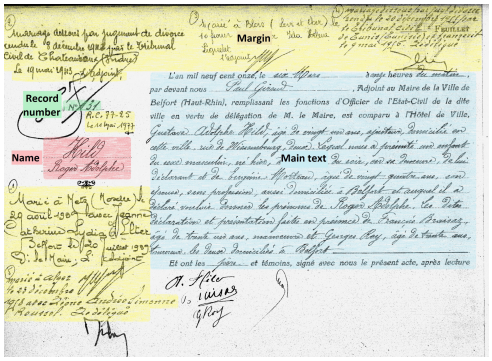



**Fig. 3** A declaration from the Belfort civil registers of birth with annotations indicating different components. The main paragraph of the text is highlighted in blue. The marginal annotations are highlighted in yellow. The declaration number is highlighted in green. The declaration name is highlighted in red. Both the declaration number and name represent the header margin of the declaration.

#### 3.1.1 Challenges in Transcribing the declarations

The transcription of Belfort declarations poses a variety of challenges, as outlined below.

- **Document layout:** The Belfort birth documents comprise double pages with two different

**Table 1** Summary of state-of-the-art approaches proposed in text image segmentation

| Ref. | Method | Dataset | Seg. Level | Accuracy |
|---|---|---|---|---|
| [4] | Adaptive U-Net architecture. | Tunisian national archives, READ, cBAD and DIVA-HisDB3 | Text line | 79.00% |
| [41] | Combines CNNs for layout analysis and the estimation of the Main Body Area (MBA) of text lines, followed by watershed transform for text line extraction. | IAM-HisDB dataset | Text line | 98.76% |
| [42] | Fully Convolutional Network (FCN). | Arabic Islamic Heritage Project (IHP), Harvard | Text line | 80.00% |
| [48] | Generative Adversarial Networks (GANs) with a U-Net architecture. | Handwritten Chinese text dataset HIT-MW and the ICDAR 2013 | Text line | 98.67% |
| [49] | Deep learning and attention (AR2U-Net model). | Arabic BADAM | Text line | 93.7% |
| [13] | A Structured Support Vector Machine (SSVM)-based method for binary quadratic problem, considering gap correlations. | ICDAR 2009 and ICDAR 2013 | Text line and word | 92.82% |
| [12] | Enhanced ILSP-LWseg method [47] employing local spatial features and gap metrics derived from SVM. | ICDAR07, ICDAR09, and ICFHR10 | Word level | 91.78% |
| [14] | A SSVM for binary classification task, distinguishing between inter-word and intra-word gaps. | ICDAR 2009 and ICDAR 2013 | Word level | 96.48% |
| [50] | Thresholding approach incorporating skew correction, baseline detection, and connected component analysis. | PHDIndic_11 | Word and character | 85.12% |
| [51] | Segmentation Facilitate Feature (SFF) technique identifies seed pixels to find junction paths, segregating touching characters in handwritten images. | 1840 legal amount words containing touching components | Character | 89.90% |

**Table 2** The structure and contents of a declaration in the Belfort civil registers of births as presented in [52].

| Structure | Content |
|---|---|
| Head margin | Registration number.<br>First and last name of the person born. |
| Main paragraph | Time and date of declaration.<br>Surname, first name and position of the official registering.<br>Surname, first name, age, profession and address of declarant.<br>Sex of the newborn.<br>Time and date of birth.<br>First and last name of the father (if different of the declarant).<br>Surname, first name, status (married or other), profession (sometimes) and address (sometimes) of the mother.<br>Surnames of the newborn.<br>Surnames, first names, ages, professions and addresses (city) of the 2 witnesses.<br>Mention of absence of signature or illiteracy of the declarant (very rarely). |
| Margins (annotations) | Mention of official recognition of paternity/maternity (by father or/and mother): surname, name of the declarant, date of recognition (by marriage or declaration).<br>Mention of marriage: date of marriage, wedding location, surname and name of spouse.<br>Mention of divorce: date of divorce, divorce location.<br>Mention of death: date and place of death, date of the declaration of death. |

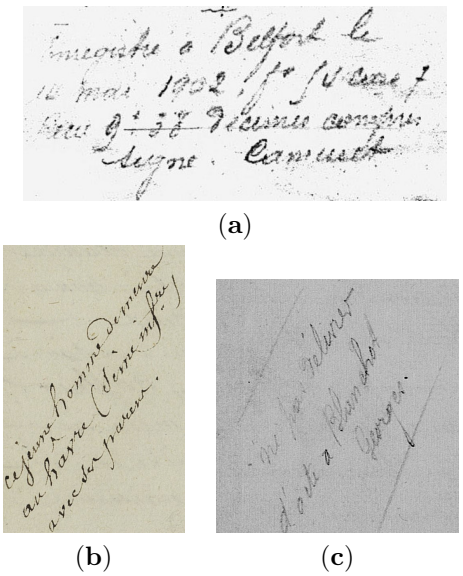
declaration layout distributions. The first type contains one complete declaration per page, while the second type contains two complete declarations per page. In specific cases, some declarations begin on the first page and extend to the second page.

- **Reading order:** Reading the components of the declaration entails following a left-to-right, top-to-bottom approach, starting with the declaration number, followed by the name, the main declaration, and the marginal annotations. However, since most annotations are written randomly along the margins of the declaration, a specific technique is required to identify the correct sequence of annotations.

- **Hybrid format:** Most of the registers consist of declarations with both printed and handwritten texts. The writing style and ink density of this handwritten text may differ significantly, which requires a seamless switch between OCR and handwriting recognition modes during transcription. Moreover, handwritten text may appear between printed lines, or even overlaid on the printed content, resulting in a complicated spatial arrangement between printed and handwritten text.
- **Marginal mentions:** These mentions contain additional information about the individual born but written thereafter, frequently with different handwriting styles. Additionally, they are positioned randomly along the margins of the primary paragraph of the declaration.
- **Text styles:** The declarations are written in a hybrid format, combining printed and handwritten text. Furthermore, the styles of handwritten text vary from one commune to another, and sometimes even within the same declaration. These variations include spiky, angular letters, varying character sizes, ornate flourishes and irregular spacing between text. As a result, overlapping word and text lines can occur within both the marginal annotations and the primary paragraph. Furthermore, mistakes were common in handwritten texts, and writers often corrected errors by scratching out or crossing through words and replacing them with corrections between text lines. Figure 3 illustrates many of these challenges found within the BCRB.
- **Skewness:** The declarations show numerous misalignment of handwritten text lines in the primary text and marginal annotations, including vertical text (rotated 90 degrees). Effective methods are necessary to correct text alignment for any degree of rotation.
- **Degradation:** The declarations also contain degraded text caused by fading ink, smearing (ink stains), and yellowing of the pages, resulting in loss of valuable content. Figure 4 demonstrates more challenges present within the BCRB.

### 3.1.2 Manual Transcription

Samples of declarations were selected for different time periods by authors. These declarations have been transcribed manually and structured by employing tags similar to XML tags, ensuring proper identification of the declaration components for further processes. Table 3 illustrates the types of tags employed in the manual transcription process.



(a) (b) (c)

**Fig. 4** Text degradation and skewness in the Belfort civil registers of births. (a) Text degradation. (b) Text skew. (c) Both challenges.

**Table 3** The set of tags utilized in the manual transcription process

| Tag | Description |
|---|---|
| <begin> | Begin of the declaration. |
| <text>...<\text> | Primary paragraph of the declaration. |
| <margin>...<\margin> | Marginal annotations. |
| <ptext>...<\ptext> | Printed text. |
| <striped>...<\striped> | Striped text. |
| <unreadable>...<\unreadable> | Unreadable text. |
| <added above>...<\added above> | Small text added above the text line. |
| <added below>...<\added below> | Small text added below the text line. |
| <page> | Start new page. |

To date, a total of 319 .txt files representing 1,010 declarations have been transcribed. Each file consists of approximately 4 declarations. Moreover, the files contain 984 margin texts. In total, there are 21,939 text lines in all the declaration components, comprising 189,976 words and 1,177,354 characters. Figure 5 provides an example of transcriptions with tags.

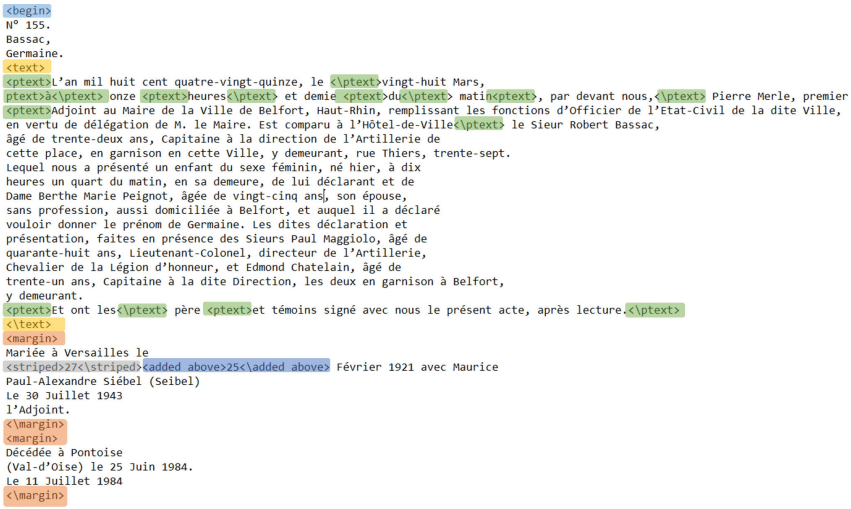



**Fig. 5** Examples of the tags used in the manual transcription process of Belfort civil registers of births.

## 3.2 Segmentation Method

The Birth declarations comprise four main components: number, name, primary paragraph, and marginal annotations. Moreover, the primary paragraph and marginal annotations are composed of several text lines. Additionally, these components comprise spaces or gaps between each two consecutive text lines. These gaps are preserved in the case of the printed text. However, it requires intensive processes to be identified in the case of the handwritten text due to numerous skewed text lines.

Manual document layout analysis methodology is employed to identify the components of the declarations within the images. This entails calculating the coordinates of these components utilizing a special interface tool. This facilitates the automatic extraction of the text line images within the primary paragraph and marginal annotations.

The proposed segmentation technique relies on calculating text distribution within the gaps between text lines for automatic extraction, which can be represented through the vertical and horizontal histograms. The crucial steps of the technique are listed below.

### 3.2.1 Filters

Gaussian blur [53] is utilized with a large blur kernel to reduce text image noise. Moreover, Otsu's thresholding [54] is applied to convert the image to binary format where pixels are either black (0) or white (255), separating foreground (text) from background and eliminating the overlap between the text lines. Furthermore, a morphological closing operation [55] is applied to connect nearby pixels in the binary image to smooth the text. Table 4 shows the parameters utilized during the filtration process. Finally, erosion operation [56] is applied to further refine the binary image and make subsequent processing of detecting the core of the text lines more robust. Figure 6 illustrates examples of the filters applied to a primary paragraph of a birth declaration.

**Table 4** Parameters values used in the image filtration process

| Parameter | Value |
|---|---|
| Gaussian blur Kernel Size | $(255, 1)$ |
| Otsu's Thresholding | – |
| Morphological Closing Kernel | $(2, 250)$ |
| Erosion Kernel | $(4, 4)$ |
| Erosion Iterations | 1 |

### 3.2.2 Core text line identification

In this step, contour detection [57] is utilized to identify the potential core of the text lines within the binary image by retrieving only the external contours, ignoring any contours nested within others. The employed contour approximation method effectively reduces the number of points representing each contour, resulting in a more concise

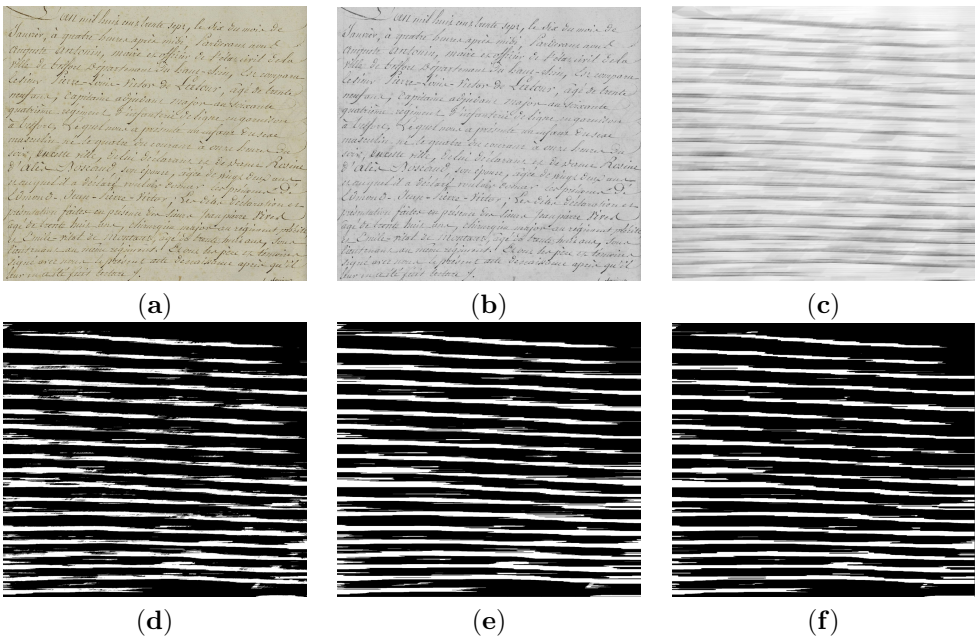


**Fig. 6** Examples of the filters applied to a primary paragraph of a declaration from Belfort civil registers of births. (a) Original declaration. (b) Grayscale conversion. (c) Gaussian blur. (d) Otsu's thresholding. (e) Morph close operation. (f) Erosion operation.

representation of the shape. Additionally, the process involves applying size and position filters to exclude non-text regions and refine the list of the detected contours as follows:

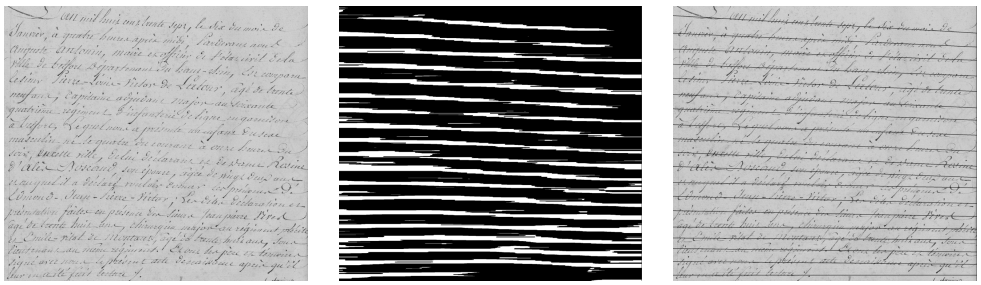
- Size filters: Contours with heights less than a certain threshold (1.5% of image height) and widths less than 30 pixels are filtered out. These filters support removing artifacts from the binary image.
- Position filters: contours lying on the right side of the image (beyond 50% of the image width) are eliminated. This filter assists in avoiding non-text elements such as the writer's signature.

Additionally, due to the skewness of the text, some lines may be detected as two separate, closely spaced, or overlapping contours. A merging method is employed to merge those portions of contours into a single contour based on proximity and angle thresholds. The proximity threshold (or distance threshold) ensures that only nearby contours (50% of the image width) are considered for merging, whereas the angle threshold permits similar angles (within 90 degrees) to account for contours that may be slightly tilted or skewed. Finally, the detected contours are expanded to span the entire width of the image, ensuring that text regions are identified as continuous lines. Figure 7 depicts the identification of the core of the text lines in the binary image.

### 3.2.3 Gaps identification

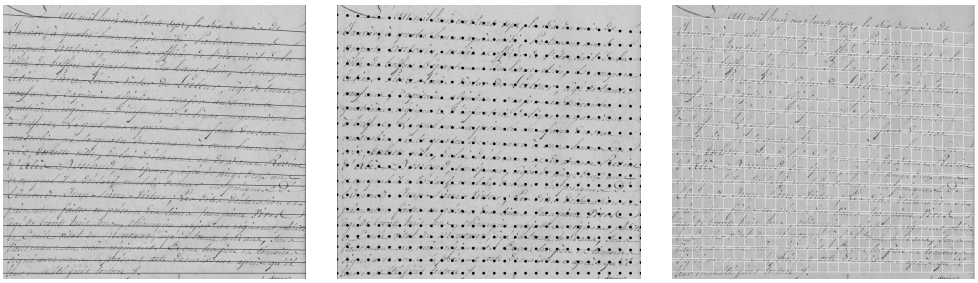
To facilitate the process of identifying gaps between consecutive text lines, points are interpolated along each line segment defined by the detected contours. This step calculates the midpoints between the start and end coordinates of each text line. These interpolated points are then associated with their respective contours, taking into account the skewness of the text line if applicable. The number of applied points is determined based on the width of the input image (primary paragraph or marginal annotation). Additionally, it can be set manually based on image characteristics.

Gaps are identified as rectangles (windows) by comparing the interpolated points of adjacent lines. The top left and top right points are taken from the first text line, while the bottom left and bottom right points are from the second line. Additionally, a height threshold is established to determine the region of interest within the gaps for further processing. Figure 8 depicts the process of points interpolation and gaps identification.



(a) (b) (c)

**Fig. 7** The process of identifying the core of the text lines in the Belfort civil registers of births. (a) Grayscale image. (b) Binary image. (c) Text lines core identification.



(a) (b) (c)

**Fig. 8** The process of identifying the gaps (windows) between each two consecutive text lines within the Belfort civil registers of births documents, taking into account the skewness of the text lines. (a) Text lines core identification. (b) Gaps points interpolation. (c) Gaps generation direction. (d) Gaps identification.

### 3.2.4 Segment points localization

In this step, the method analyzes the gaps to identify the most suitable locations for potential segment points based on the absence of significant text presence. This analysis first splits the gap horizontally into three regions and examines the density of black pixels within the middle region of the gap. If the text density value is lower than a predefined threshold (5%), exhibiting an absence of substantial text content, a single segment point is strategically placed at the midpoint of the gap, ensuring the segment points inserted in areas devoid of significant text, and minimizing the risk of disrupting actual text regions. Figure 9 shows the identified regions within gaps area.

The process can be defined mathematically by the following formulas.

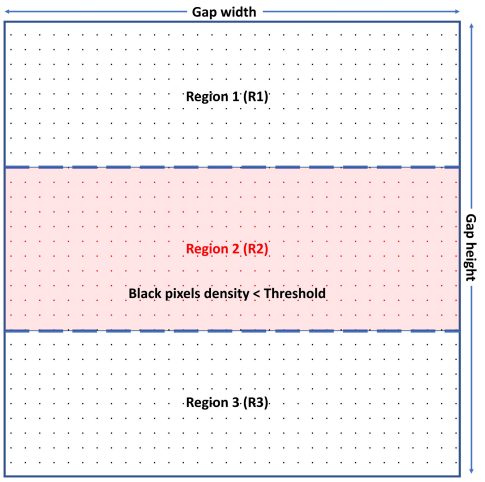



**Fig. 9** The process of dividing the gaps into three regions to localize the segment points based on the black pixels density.

$$D_{R_2} = \frac{S_{R_2}}{N_{R_2}} \tag{1}$$

$$S_{R_2} = \sum_{(x,y)\in R_2} I(x,y) \tag{2}$$

where $I(x,y)$ is a pixel value at coordinates $(x,y)$ in the gap image, where $I(x,y) = 1$ for a black pixel and $I(x,y) = 0$ for a white pixel, $G$ is the gap horizontally split into three regions $R_1$, $R_2$, and $R_3$, with $R_2$ being the middle region. $N_{R_2}$ is the

number of pixels in region $R_2$, $S_{R_2}$ is the sum of black pixels in region $R_2$, and $T$ is the predefined threshold (5%).

If the density $D_{R_2}$ is less than $T$, then:

$$P = M = \left( \frac{x_1 + x_2}{2}, \frac{y_1 + y_2}{2} \right) \tag{3}$$

where $(x_1, y_1)$ and $(x_2, y_2)$ are the coordinates defining the horizontal bounds of the gap $G$. The condition for placing a segment point at the midpoint is:

$$\frac{S_{R_2}}{N_{R_2}} < T \Rightarrow P = \left( \frac{x_1 + x_2}{2}, \frac{y_1 + y_2}{2} \right). \tag{4}$$

In cases where the text density within a gap exceeds the mentioned threshold, which mostly arises with handwritten text rather than printed text due to ornate texts, the method employs histogram-based analysis to identify potential segment points. Vertical and horizontal histograms are generated from the gap image, where valleys represent potential locations for segment points, as they indicate regions with relatively low text density. The deepest valley in both the vertical and horizontal histograms is selected as the segment point. Figure 10 shows the generated histograms for the determination of the highly beneficial segment point position.

The two histograms are generated by summing the intensity values along the vertical and horizontal axes of the gap region to obtain two 1D arrays. The indices of the deepest valleys in the histograms are then identified to form the coordinates for segment point. The x-coordinate of segment point is calculated by adding the index of the deepest valley in the vertical histogram to the starting x-coordinate, whereas, the y-coordinate is determined by adding the index of the deepest valley in the horizontal histogram to the starting y-coordinate, as illustrated in the following formulas.

If $D_{R_2}$ is greater than $T$, then:

$$H_v(x) = \sum_{y=1}^{h} I(x, y) \quad \text{for} \quad x = 1, 2, \ldots, w, \tag{5}$$

where $H_v(x)$ is the vertical histogram, $x$ is the index representing the column of the gap region, ranging from 1 to $w$, the width of the gap image, and $y$ is the index representing the row of the gap region, ranging from 1 to $h$, the height of the gap region.

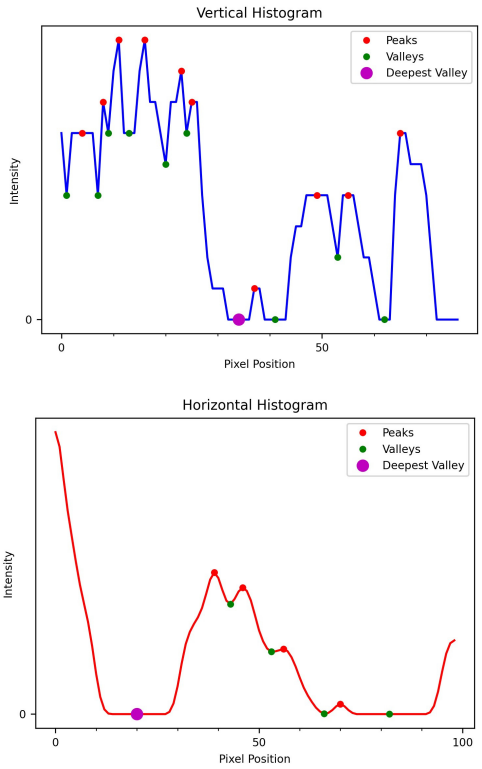



**Fig. 10** Process of generating vertical and horizontal histograms and identifying the deepest valleys to localize the segment points effectively.

$$H_h(y) = \sum_{x=1}^{w} I(x, y) \quad \text{for} \quad y = 1, 2, \ldots, h, \tag{6}$$

where $H_h(y)$ is the horizontal histogram, $y$ is the index representing the row of the gap region, ranging from 1 to $h$, the height of the gap image, and $x$ is the index representing the column of the gap region, ranging from 1 to $w$, the width of the gap region.

$$x_v = \arg\min_x H_v(x), \tag{7}$$

$$y_v = \arg\min_y H_h(y), \tag{8}$$

where $x_v$ and $y_v$ are the indices of the deepest valley in the vertical and horizontal histograms.

$$x_p = x_s + x_v, \tag{9}$$

$$y_p = y_s + y_v, \tag{10}$$

And the final equation:

$$(x_p, y_p) = (x_s + x_v, y_s + y_v). \tag{11}$$

Figure 11 depicts the segment points identification process based on thresholding technique and histogram analysis.

#### 3.2.5 Segmentation path identification

The segment points corresponding to each line of text are connected from the left to right border of the image to form a continuous path that covers the entire gap between each two consecutive text lines. Finally, the text lines are processed for the extraction process, which represents the region between the paths within the image. Figure 12 depicts an example of the path generation process, and Table 5 lists all the thresholds and parameters utilized through text lines identification process.

**Table 5** Thresholds and parameters values used in the text lines identification process

| Threshold/Parameter | Value |
|---|---|
| Minimum height of text contours | 1.5% |
| Minimum width of text contours | 30 pixels |
| Exclusion factor | 50% |
| Proximity threshold for merging contours | 50% |
| Angle difference threshold | 90 degrees |
| Number of interpolated points per line | 30 |
| Vertical kernel size for gap histograms | 9 |
| Horizontal kernel size for gap histograms | 9 |
| Black pixel proportion threshold | 5% |

### 3.3 Structured Data Generation

As per our prior research [52], a unique deep learning approach is required to transcribe the French BCRB, considering the significant impediments it poses. This stage necessitates the development of a precise technique for correlating the segmented text line images with the manually transcribed text to construct a training dataset for a deep learning model.

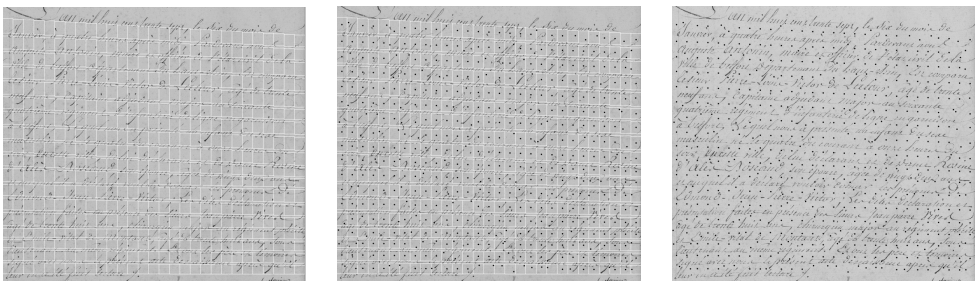
#### 3.3.1 XML File Generator

A novel tool has been developed to generate XML files. This tool establishes a link between each segments of the images and their corresponding transcriptions at both the paragraph and text line levels by means of added tags. These XML files have been designed to include several essential properties such as:

- Image Information: This section of the XML file provides details related to the image, including its register name, type (single or double page), image height, and image width.
- Reading Order: This part of the file identifies the sequential reading order of components within the declaration and assigns a unique index number to each, specifying their respective type (title number, title name, paragraph, margin).
- Declaration Number and Name: It includes information regarding the declaration's number and name, with their coordinates within the image and the corresponding transcribed text.
- Paragraphs and Margins: This part provides information regarding the primary paragraphs and marginal annotations within the declaration, including their coordinates within the image and the corresponding transcribed text. Additionally, it provides details about the text lines within these paragraphs and margins, including a unique identification number, their respective coordinates within the image, and the corresponding transcribed text.

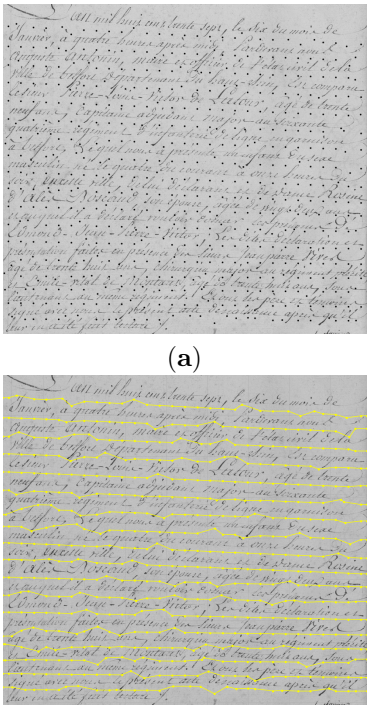
## 4 Results

The proposed method has been applied to samples of the BCRB images described in Section 3.1, which as already stated possess challenges such as text skewness, degradation, and text overlapping. Additionally, we evaluated the performance of our method on other datasets used in the International Conference on Document Analysis and Recognition (ICDAR), such as the IAM historical document database (Saint-Gall) and the READ Bozen dataset. The ground truth of these samples has been annotated manually using VGG image annotator (VIA) version 2.0.12 [58]. Figure 13 displays examples of the annotated datasets utilized during the evaluation process.



(**a**) (**b**) (**c**)

**Fig. 11** The process of segment points identification within the gaps based on thresholding technique and histogram analysis. (a) Gaps identification. (b) Segment points placement. (c) Final placement.



(**a**) (**b**)

**Fig. 12** The process of identifying the paths that represent the borders of the text lines within the image. (a) Segment points placement. (b) Paths identification.

An ablation study has been conducted to understand the performance of the proposed method and determine the values of the parameters and thresholds for accurate segmentation processes, highlighting the accuracy obtained at each stage. Additionally, an accuracy comparison with state-of-the-art methods has been carried out to validate the effectiveness of the proposed approach.

Finally, XML files have been generated to structure the significant components presented by the BCRB at both the paragraph and text line levels. These files will play a crucial role in facilitating the data labeling process during the training stage of the text recognition models.

## 4.1 Evaluation metric

In this study, several evaluation metrics have been employed to assess the performance of the proposed method. These include the evaluation metrics presented in [59] and used in the ICDAR for image segmentation contests during the years 2001, 2003, 2005, 2007, and 2013, respectively.

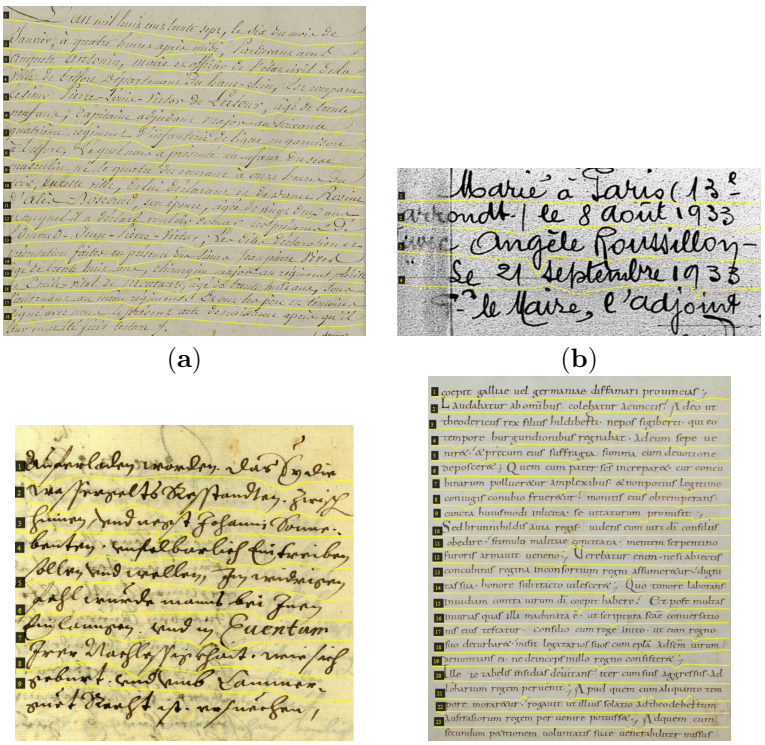
The Intersection over Union ($IoU$) measures the overlap between the predicted text line image regions and ground truth image regions. $IoU$ is computed using the following formula:

$$IoU = \frac{|\text{predicted mask} \cap \text{ground truth mask}|}{|\text{predicted mask} \cup \text{ground truth mask}|}, \tag{12}$$

where $|\text{predicted mask} \cap \text{ground truth mask}|$ is the number of pixels in the intersection of the predicted image and the ground truth image, and $|\text{predicted mask} \cup \text{ground truth mask}|$ is the number of pixels in the union of the predicted image and the ground truth image.

Moreover, we utilized the Detection Rate (DR) metric to calculate the ratio of correctly detected text lines to the total number of ground truth text lines, with a defined acceptance threshold of 0.9 to count the correct text line detections. DR is computed using the following formula:

$$DR = \frac{O2O}{N_{\text{gt}}}, \tag{13}$$



(a) (b) (c) (d)

**Fig. 13** Samples of the ground truth annotations generated by VGG image annotator tool. (a) Primary paragraph from Belfort civil registers of birth. (b) Marginal annotation from Belfort civil registers of birth. (c) Paragraph from READ Bozen dataset. (d) Page from Saint Gall dataset.

where one-to-one matching ($O2O$) is the number of correctly matched text lines between the predicted and ground truth text lines, and $N_{\text{gt}}$ is the total number of ground truth text lines.

Additionally, the Recognition Accuracy (RA) metric is utilized to calculate the ratio of correctly detected text lines to the total number of predicted text lines. RA is computed using the following formula:

$$RA = \frac{O2O}{N_{\text{pred}}}, \tag{14}$$

where $O2O$ is the number of correctly matched text lines between the predicted and ground truth text lines, and $N_{\text{pred}}$ is the total number of predicted text lines.

Finally, F-Measure (FM) is utilized to calculate the harmonic mean of the Detection Rate (DR) and the Recognition Accuracy (RA), providing an overall metric for performance evaluation. It is computed using the following formula:

$$FM = 2 \times \frac{DR \times RA}{DR + RA}. \tag{15}$$

Employing these metrics provides a robust and comprehensive evaluation of various performance aspects of the proposed method. Detection Rate (DR) and Recognition Accuracy (RA) aim to measure the capability to detect and recognize text lines, providing insights into the practical performance of the method. F-Measure (FM), as a classical F1-score, offers a balanced view of the method's performance by considering both DR and RA.

## 4.2 Text line segmentation

The proposed text line segmentation method has been applied to the BCRB on both the primary paragraph and the marginal annotations parts. The images have been selected from different registers and span different periods. Additionally, it contains most of the challenges that could evaluate the effectiveness of our method such as hybrid

text, text skewness, text overlapping, and different handwritten styles.

The method uses a polygon shape border in extracting the text line images from the dataset image which results from the process of connecting the segment points. Thus, it was crucial to identify one image size for both the predicted and ground truth text line images. This process involves calculating the max width and height of the images, and padding the small one to achieve similar size. The padding process involves calculating the dominant background color within the image to fill the padding space. This approach ensures maintaining the characteristics of the handwritten text in both the detected and ground truth ones. Additionally, the evaluation process also involves applying Gaussian blur and Otso thresholding to eliminate background noise and clear the handwritten text for the evaluation metrics. Figure 14 depicts examples of the segmentation process on both the primary paragraph and the marginal annotations of the dataset in comparison with the ground truth images.

We have also evaluated our method on other datasets such as the IAM historical document database (Saint-Gall) and READ Bozen dataset, which were parts of datasets used in the International Conference on Document Analysis and Recognition (ICDAR) and International Conference on Frontiers in Handwriting Recognition (ICFHR).

The Saint Gall dataset presented in [60] comprises 60 pages of a handwritten historical manuscript written in the Latin language during the 9th century. The dataset is available online in format (JPEG, 300dpi), binarized, normalized, and transcribed at the text line level. Furthermore, the READ Bozen dataset [61] is part of the European Union's Horizon 2020 project that represents the minutes of council meetings held between 1470 and 1805 (about 30,000 pages), and that contains early modern German handwriting. The available dataset comprises 400 pages with annotations at text line levels. Figure 14 depicts examples of the segmentation results on the datasets in comparison with the ground truth images, while Table 6 shows the performance accuracy of the proposed method.

Despite the effectiveness of our proposed method for text line extraction, some limitations were identified. Firstly, the method struggles to detect the cores of text lines in cases of significant degradation, such as severe ink smearing or faded text, which can obscure key features required for accurate segmentation. Secondly, in instances where text images exhibit extreme skewness or rotated text, notably within marginal annotations, the method may fail to segment text lines correctly. Figure 15 depicts examples of these limitations. Finally, the method has not been tested on documents with vastly different formats, such as those with vertical writing systems.

**Table 6** Accuracy results reported on Belfort civil registers of birth, Saint Gall, and READ Bozen datasets.

| Component | Accuracy | | | |
|---|---|---|---|---|
| | IoU | DR | RA | FM |
| Primary paragraphs | 97.5% | 99% | 98% | 98.50% |
| Marginal annotations | 93.1% | 96% | 94% | 94.79% |
| Saint Gall | 98.3% | 99% | 99% | 98.91% |
| READ Bozen | 92.8% | 97% | 94% | 95.41% |

### 4.2.1 Ablation study

An ablation study is carried out to systematically test variations and identify the optimal configuration for the segmentation process. This study involves testing the performance of the method with different thresholds and parameter values, such as the kernel sizes of the Gaussian blur filter and the morphological operations. Additionally, it examines the number of gap regions (windows) between each two consecutive text lines. Furthermore, it considers the vertical and horizontal kernel sizes for smoothing the histograms to localize the segment points. Finally, the study assesses the black pixel proportion threshold used to detect the presence of text in the middle region of the gaps.

While many of the parameters are dynamically computed based on the input image, certain parameters still require manual tuning for specific cases. This manual tuning, particularly for kernel sizes, plays a critical role in achieving optimal performance across different types of historical documents.

The study demonstrated improvement in the core text line identification process when tuning the kernel sizes of the Gaussian blur and morphological operations, while the number of the gabs, the kernel size of the histograms, and the

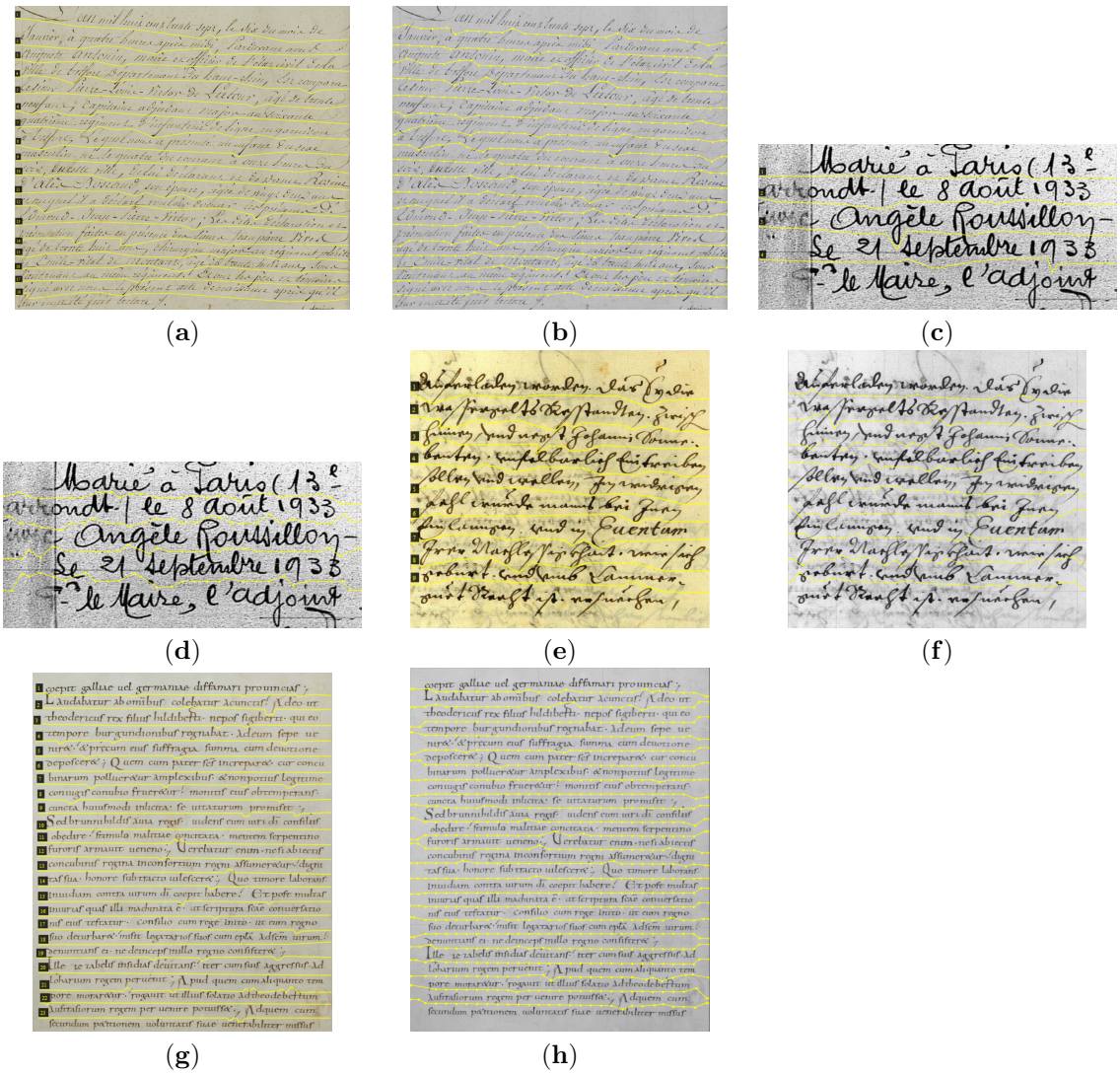


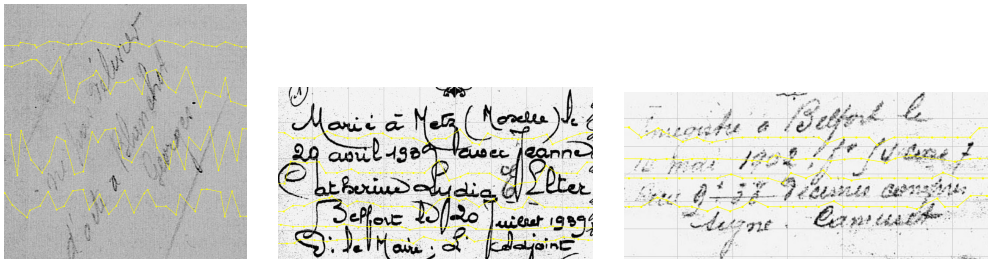
**Fig. 14** Examples of the segmentation results. (a) Ground truth of a primary paragraph in Belfort civil registers of birth. (b) Segmentation result of the primary paragraph in Belfort civil registers of birth. (c) Ground truth of a marginal annotation in Belfort civil registers of birth. (d) Segmentation result of the marginal annotation in Belfort civil registers of birth. (e) Ground truth of a paragraph in the READ Bozen dataset. (f) Segmentation result of the paragraph in the READ Bozen dataset. (g) Ground truth of a page in the Saint Gall dataset. (h) Segmentation result of the page in the Saint Gall dataset.

text threshold show improvements in the segment point localization process, which facilitates the detection process of the text lines borders.

## 4.3 Accuracy comparison with stat of the art methods

We conducted an accuracy comparison with state-of-the-art approaches that utilize different methodologies for the text line segmentation process, including both traditional image processing techniques and modern artificial intelligence approaches. Table 8 shows the performance accuracy of these methods when applied to the Saint Gall and READ Bozen datasets.

Several of the methods compared utilize deep learning techniques, such as Convolutional Neural Networks (CNNs) and Fully Convolutional Networks (FCNs) for text line segmentation [41, 64, 65]. These AI-based approaches have demonstrated effectiveness in handling challenging handwritten text. Despite relying on traditional image



(a) (b) (c)

**Fig. 15** Examples of segmentation challenges encountered in marginal annotations: (a) Extreme skewed text. (b) Heavily overlapped text. (c) Degraded text.

**Table 7** Configuration variations and performance accuracy

| Parameter/ Threshold | Configuration | | | | | | Accuracy | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | GB-KS | Morph-KS | E-KS | Gaps no. | Hist-KS | BPx | IoU | DR | RA | FM |
| **GB-KS** | (51, 1) | (2, 250) | (4, 4) | 30 | (7, 7) | 0.05 | 97.12% | 11% | 22% | 14.67% |
| | (101, 1) | (2, 250) | (4, 4) | 30 | (7, 7) | 0.05 | 94.72% | 57% | 65% | 60.74% |
| | (151, 1) | (2, 250) | (4, 4) | 30 | (7, 7) | 0.05 | 96.21% | 93% | 95% | 94.10% |
| | (201, 1) | (2, 250) | (4, 4) | 30 | (7, 7) | 0.05 | 97.16% | 94% | 97% | 95.41% |
| | (251, 1) | (2, 250) | (4, 4) | 30 | (7, 7) | 0.05 | 97.42% | 99% | 98% | 98.50% |
| **Morph-KS** | (251, 1) | (2, 50) | (4, 4) | 30 | (7, 7) | 0.05 | 89.47% | 95% | 90% | 92.44% |
| | (251, 1) | (2, 100) | (4, 4) | 30 | (7, 7) | 0.05 | 89.73% | 95% | 90% | 92.44% |
| | (251, 1) | (2, 150) | (4, 4) | 30 | (7, 7) | 0.05 | 90.34% | 95% | 90% | 92.44% |
| | (251, 1) | (2, 200) | (4, 4) | 30 | (7, 7) | 0.05 | 94.24% | 95% | 90% | 92.44% |
| | (251, 1) | (2, 250) | (4, 4) | 30 | (7, 7) | 0.05 | 97.42% | 99% | 98% | 98.50% |
| **E-KS** | (251, 1) | (2, 250) | (1, 1) | 30 | (7, 7) | 0.05 | 96.97% | 97% | 97% | 97.00% |
| | (251, 1) | (2, 250) | (2, 2) | 30 | (7, 7) | 0.05 | 97.10% | 97% | 97% | 97.00% |
| | (251, 1) | (2, 250) | (3, 3) | 30 | (7, 7) | 0.05 | 97.23% | 97% | 97% | 97.00% |
| | (251, 1) | (2, 250) | (4, 4) | 30 | (7, 7) | 0.05 | 97.24% | 99% | 97% | 98.00% |
| | (251, 1) | (2, 250) | (5, 5) | 30 | (7, 7) | 0.05 | 97.42% | 99% | 98% | 98.50% |
| **Gaps no.** | (251, 1) | (2, 250) | (4, 4) | 10 | (7, 7) | 0.05 | 95.64% | 99% | 98% | 98.50% |
| | (251, 1) | (2, 250) | (4, 4) | 30 | (7, 7) | 0.05 | 97.42% | 99% | 98% | 98.50% |
| | (251, 1) | (2, 250) | (4, 4) | 50 | (7, 7) | 0.05 | 96.96% | 99% | 98% | 98.50% |
| | (251, 1) | (2, 250) | (4, 4) | 70 | (7, 7) | 0.05 | 96.68% | 99% | 98% | 98.50% |
| | (251, 1) | (2, 250) | (4, 4) | 90 | (7, 7) | 0.05 | 96.63% | 99% | 98% | 98.50% |
| **Hist-KS** | (251, 1) | (2, 250) | (4, 4) | 30 | (3, 3) | 0.05 | 97.09% | 99% | 98% | 98.50% |
| | (251, 1) | (2, 250) | (4, 4) | 30 | (5, 5) | 0.05 | 97.36% | 99% | 98% | 98.50% |
| | (251, 1) | (2, 250) | (4, 4) | 30 | (7, 7) | 0.05 | 97.42% | 99% | 98% | 98.50% |
| | (251, 1) | (2, 250) | (4, 4) | 30 | (9, 9) | 0.05 | 97.50% | 99% | 98% | 98.50% |
| | (251, 1) | (2, 250) | (4, 4) | 30 | (11, 11) | 0.05 | 97.17% | 99% | 98% | 98.50% |
| **BPx** | (251, 1) | (2, 250) | (4, 4) | 30 | (9, 9) | 0.03 | 97.37% | 99% | 98% | 98.50% |
| | (251, 1) | (2, 250) | (4, 4) | 30 | (9, 9) | 0.04 | 97.33% | 99% | 98% | 98.50% |
| | (251, 1) | (2, 250) | (4, 4) | 30 | (9, 9) | 0.05 | 97.50% | 99% | 98% | 98.50% |
| | (251, 1) | (2, 250) | (4, 4) | 30 | (9, 9) | 0.06 | 97.50% | 99% | 98% | 98.50% |
| | (251, 1) | (2, 250) | (4, 4) | 30 | (9, 9) | 0.07 | 96.84% | 99% | 98% | 98.50% |

***GB-KS:** Gaussian blur kernel size, **Morph-KS:** Morphological operation kernel size, **E-KS:** Erosion kernel size, **Gaps no.:** Number of gaps, **Hist-KS:** Histogram kernel size, **BPx:** Black pixel proportion.

**Table 8** Accuracy comparison of state-of-the-art text line segmentation methods on Saint Gall and READ Bozen datasets.

| Dataset | Reference | FM (%) |
|---|---|---|
| Saint-Gall | [41] | 98.76 |
| | [62] | 93.99 |
| | [63] | 97.2 |
| | Ours | 98.5 |
| READ Bozen | [64] | 97.5 |
| | [65] | 99.6 |
| | [66] | 94.2 |
| | Ours | 95.41 |

processing techniques, our method delivers competitive outcomes. Overall, the proposed method has shown highly effective performance in the text line segmentation process, achieving competitive results on both datasets in comparison with state-of-the-art methods developed using both traditional image processing and artificial intelligence techniques.

## 4.4 XML File Generation

We have developed an automated verification tool to validate our manual transcriptions. This tool formats and corrects the usage of tags within the transcriptions, ensuring precise alignment between the image components and the corresponding transcriptions when generating the XML files, providing a flexible and scalable structure that facilitates future processing.

These XML files play an important role in Optical Character Recognition (OCR) systems, where the structured data can improve the

accuracy of both printed and handwritten text recognition. Furthermore, the XML representation greatly aids deep learning model training by allowing for metadata extraction and the indexing of important data, supporting tasks such as writer identification, keyword spotting, and historical document retrieval. An example of one of these XML files is depicted in Figure 16.

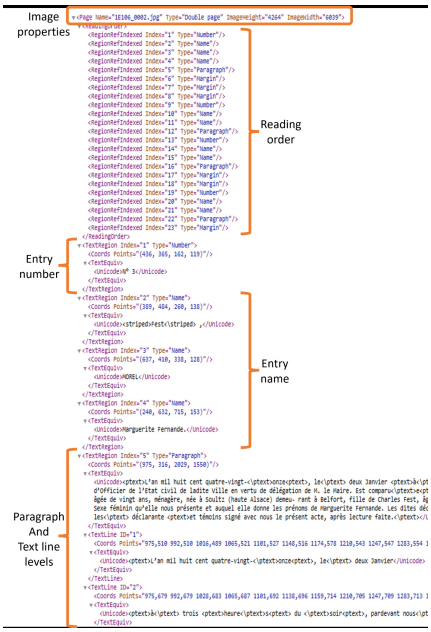



**Fig. 16** Example of the structured data file (.xml) for Belfort civil registers of birth

# 5 Conclusion and Future Work

This paper presents an unsupervised approach for text line extraction in historical documents, concentrating on the BCRB. The proposed method exhibits a robust segmentation process based on different image processing techniques to address the challenges posed by these historical registers. The results demonstrate high accuracy in segmenting text lines, with competitive performance compared to state-of-the-art methods.

The method requires no extensive preprocessing and there is no need for text line image skew correction. Additionally, many parameters values are computed dynamically based on the input image, eliminating the demand for a training dataset. Thus, it is appropriate for a wide range of historical documents with varying text styles and formats, including those written in French and other languages. Moving forward, we will focus on developing adaptive techniques that can automatically learn the optimal parameters for each document. This enhancement would increase the method's generalizability to a wider range of historical documents with varying levels of degradation.

The development of the structured data generation tool ensures that transcriptions are correctly aligned with their corresponding image components for further research, such as integrating with OCR systems to evaluate end-to-end transcription accuracy and streamline the entire digitization process. Additionally, it enables the development of approaches for segmenting the dataset into word and character levels to enhance the preservation and accessibility of historical documents.

# Authorship contribution statement

**Wissam AlKendi**: Conceptualization, Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis. **Franck Gechter**: Conceptualization, Review & editing, Supervision. **Laurent Heyberger**: Conceptualization, Writing – original draft, Review & editing, Supervision. **Christophe Guyeux**: Conceptualization, Review & editing, Supervision.

# Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

# Data availability

The datasets (Belfort civil registers of births) analyzed during the current study are available online as mentioned, and from the corresponding author on reasonable request.

## Acknowledgements

This work has been supported by the EIPHI Graduate School (contract "ANR-17-EURE-0002") and by the Bourgogne-Franche-Comté Region.

## References

[1] Rosenzweig, R.: Clio Wired: The Future of the Past in the Digital Age. Columbia University Press, New York (2011)

[2] Terras, M., Nyhan, J., Vanhoutte, E.: Digital Humanities in Practice. Facet Publishing, London (2013)

[3] Chen, K., Seuret, M., Liwicki, M., Hennebert, J., Ingold, R.: Page segmentation of historical document images with convolutional autoencoders. In: Proceedings of the 2015 13th International Conference on Document Analysis and Recognition (ICDAR), pp. 1011–1015. IEEE, Tunis, Tunisia (2015)

[4] Mechi, O., Mehri, M., Ingold, R., Ben Amara, N.E.: Text line segmentation in historical document images using an adaptive u-net architecture. In: Proceedings of the 2019 International Conference on Document Analysis and Recognition (ICDAR), pp. 369–374. IEEE, Sydney, Australia (2019)

[5] Likforman-Sulem, L., Zahour, A., Taconet, B.: Text line segmentation of historical documents: A survey. International Journal of Document Analysis and Recognition (IJDAR) **9**, 123–138 (2007)

[6] Mehri, M., Sellami, A., Tabbone, S.: Historical document image segmentation combining deep learning and gabor features. In: Proceedings of the 16th International Conference on Document Analysis and Recognition (ICDAR), pp. 395–410. IEEE, San José, California, USA (2023)

[7] Jemni, S.K., Ammar, S., Souibgui, M.A., Kessentini, Y., Cheddad, A.: St-keys: Self-supervised transformer for keyword spotting in historical handwritten documents. arXiv preprint arXiv:2303.03127 (2023)

[8] Christlein, V., Diem, M., Kleber, F., Mühlberger, G., Schwägerl-Melchior, V., Gelder, E., Maier, A.: Automatic writer identification in historical documents: A case study. Zeitschrift für digitale Geisteswissenschaften (2016)

[9] He, S., Schomaker, L.: Deep adaptive learning for writer identification based on single handwritten word images. Pattern Recognition **88**, 64–74 (2019)

[10] Nguyen, Q.D., Phan, N.M., Krömer, P., Le, D.A.: An efficient unsupervised approach for ocr error correction of vietnamese ocr text. IEEE Access **11**, 58406–58421 (2023)

[11] Bennani-Smires, K., Musat, C., Hossmann, A., Baeriswyl, M., Jaggi, M.: Simple unsupervised keyphrase extraction using sentence embeddings. arXiv preprint arXiv:1801.04470 (2018). Available at https://arxiv.org/abs/1801.04470

[12] Simistira, F., Papavassiliou, V., Stafylakis, T., Katsouros, V.: Enhancing handwritten word segmentation by employing local spatial features. In: Proceedings of the 2011 International Conference on Document Analysis and Recognition (ICDAR), pp. 1314–1318. IEEE, Beijing, China (2011)

[13] Ryu, J., Koo, H., Cho, N.I.: Word segmentation method for handwritten documents based on structured learning. IEEE Signal Processing Letters **22**(7), 838–842 (2015)

[14] Sharma, M.K., Dhaka, V.S.: Segmentation of handwritten words using structured support vector machine. Pattern Analysis and Applications **23**(3), 1355–1367 (2020)

[15] Sahoo, P.K., Soltani, S., Wong, A.K.C.: A survey of thresholding techniques. Computer Vision, Graphics, and Image Processing **41**(2), 233–260 (1988)

[16] Ali, A.A., Suresha, M.: Efficient algorithms for text lines and words segmentation for recognition of arabic handwritten script. In: Emerging Research in Computing, Information, Communication and Applications:

ERCICA 2018, Volume 1, pp. 387–401. Springer, Cham, Switzerland (2019)

[17] Sun, Y., Butler, T.S., Shafarenko, A., Adams, R., Loomes, M., Davey, N.: Word segmentation of handwritten text using supervised classification techniques. Applied Soft Computing **7**(1), 71–88 (2007)

[18] Delsalle, P.: Histoires de Familles: Les Registres Paroissiaux et D'état Civil, du Moyen Âge à Nos Jours: Démographie et Généalogie (Family History: Parish and Civil Status Registers, from the Middle Ages to the Present Day: Demography and Genealogy). Presses universitaires de Franche-Comté, Besançon (2009)

[19] Gourdon, V.: L'histoire sociale de la famille en france à l'époque moderne et au xixe siècle : traditions historiographiques et renouvellements thématiques (the social history of the family in france in the modern era and the 19th century: Historiographical traditions and thematic renewals). In: García González, F., Guzzi-Heeb, S. (eds.) Historia de la Familia, Historia Social. Experiencias de Investigación en España Y en Europa (siglos XVI-XIX) (History of the Family, Social History. Research Experiences in Spain and Europe (16th-19th Centuries)), pp. 167–193. Ediciones TRea/Ediciones de la Universidad de Castilla-La Mancha, Gijón (2023)

[20] Laslett, P., Oosterveen, K., Smith, R.M.: Bastardy and Its Comparative History: Studies in the History of Illegitimacy and Marital Nonconformism in Britain, France, Germany, Sweden, North America, Jamaica, and Japan. Edward Arnold, London (1980)

[21] Heyberger, L.: L'industrialisation de belfort : une conséquence positive du siège de 1870-1871 ? approche par l'histoire anthropométrique (the industrialization of belfort: A positive consequence of the siege of 1870-1871? an approach through anthropometric history). In: Belot, R. (ed.) 1870 de la Guerre à la Paix (1870 from War to Peace), pp. 207–217. Hermann, Strasbourg-Belfort, Paris (2013)

[22] Memon, J., Sami, M., Khan, R.A., Uddin, M.: Handwritten optical character recognition (ocr): A comprehensive systematic literature review (slr). IEEE Access **8**, 142642–142668 (2020)

[23] Li, M., Lv, T., Chen, J., Cui, L., Lu, Y., Florencio, D., Zhang, C., Li, Z., Wei, F.: Trocr: Transformer-based optical character recognition with pre-trained models. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, pp. 13094–13102 (2023)

[24] Lefta, A.S., Daway, H.G., Jouda, J.: Red blood cells detecting depending on binary conversion at multi threshold values. Al-Mustansiriyah Journal of Science **33**(1), 69–76 (2022)

[25] Hussain, S.A.K., Al-Nayyef, H., Al Kindy, B., Qassir, S.A.: Human earprint detection based on ant colony algorithm. International Journal of Intelligent Systems and Applications in Engineering **11**(2), 513–517 (2023)

[26] Al-Khalidi, F.Q., Alkindy, B., Abbas, T.: Extract the breast cancer in mammogram images. Technology **10**(2), 96–105 (2019)

[27] Leedham, G., Chen, Y., Takru, K., Tan, J.H.N., Mian, L.H.: Comparison of some thresholding algorithms for text/background segmentation in difficult document images. In: Proceedings of the 7th International Conference on Document Analysis and Recognition (ICDAR), pp. 859–864. IEEE, Edinburgh, UK (2003)

[28] Susan, S., Rachna Devi, K.M.: Text area segmentation from document images by novel adaptive thresholding and template matching using texture cues. Pattern Analysis and Applications **23**(2), 869–881 (2020)

[29] Naseir, A.: A comparison study of image edge segmentation methods using prewitt, sobel and laplacian of gaussian for medical images. Journal of Education for Pure Science - University of Thi-Qar **12**(2), 96–106 (2022)

[30] Qiu, X., Chen, Z., Adnan, S., He, H.: Improved mr image denoising via low-rank

approximation and laplacian-of-gaussian edge detector. IET Image Processing **14**(12), 2791–2798 (2020)

[31] Haralick, R.M.: Digital step edges from zero crossing of second directional derivatives. In: Readings in Computer Vision, pp. 216–226. Elsevier, San Francisco, CA, USA (1987)

[32] Cho, H., Sung, M., Jun, B.: Canny text detector: Fast and robust scene text localization algorithm. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3566–3573. IEEE, Las Vegas, NV, USA (2016)

[33] Le, V.P., Nayef, N., Visani, M., Ogier, J.M., De Tran, C.: Text and non-text segmentation based on connected component features. In: Proceedings of the 13th International Conference on Document Analysis and Recognition (ICDAR), pp. 1096–1100. IEEE, Tunis, Tunisia (2015)

[34] Wang, Y., Wang, W., Li, Z., Han, Y., Wang, X.: Research on text line segmentation of historical tibetan documents based on the connected component analysis. In: Pattern Recognition and Computer Vision: First Chinese Conference, PRCV 2018, Guangzhou, China, November 23-26, 2018, Proceedings, Part III, pp. 74–87. Springer, Cham, Switzerland (2018)

[35] Ghosal, A., Nandy, A., Das, A.K., Goswami, S., Panday, M.: A short review on different clustering techniques and their applications. In: Emerging Technology in Modelling and Graphics: Proceedings of IEM Graph 2018, pp. 69–83. Springer, Cham, Switzerland (2020)

[36] Rong, Y.: Staged text clustering algorithm based on k-means and hierarchical agglomeration clustering. In: Proceedings of the IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA), pp. 124–127. IEEE, Dalian, China (2020)

[37] Kumar, A., Ingle, Y.S., Pande, A., Dhule, P.: Canopy clustering: A review on pre-clustering approach to k-means clustering. International Journal of Innovative and Advanced Computer Science (IJIACS) **3**(5), 22–29 (2014)

[38] Murtagh, F., Contreras, P.: Algorithms for hierarchical clustering: An overview. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **2**(1), 86–97 (2012)

[39] Murtagh, F., Contreras, P.: Algorithms for hierarchical clustering: An overview, ii. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **7**(6), 1219 (2017)

[40] Kaur, R.P., Kumar, M., Jindal, M.K.: Newspaper text recognition of gurumukhi script using random forest classifier. Multimedia Tools and Applications **79**(11), 7435–7448 (2020)

[41] Pastor-Pellicer, J., Afzal, M.Z., Liwicki, M., Castro-Bleda, M.J.: Complete system for text line extraction using convolutional neural networks and watershed transform. In: Proceedings of the 2016 12th IAPR Workshop on Document Analysis Systems (DAS), pp. 30–35. IEEE, Santorini, Greece (2016)

[42] Barakat, B., Droby, A., Kassis, M., El-Sana, J.: Text line segmentation for challenging handwritten document images using fully convolutional network. In: Proceedings of the 16th International Conference on Frontiers in Handwriting Recognition (ICFHR), pp. 374–379. IEEE, Niagara Falls, USA (2018)

[43] Siddiqua, S., Naveena, C., Manvi, S.K.: A combined edge and connected component based approach for kannada text detection in images. In: Proceedings of the International Conference on Recent Advances in Electronics and Communication Technology (ICRAECT), pp. 121–125. IEEE, Bangalore, India (2017)

[44] Sakhi, O.B.: Segmentation of heterogeneous document images: An approach based on machine learning, connected components analysis, and texture analysis. PhD thesis, Université Paris-Est (2012)

[45] Geetha, M.N., Samundeeswari, E.S.: Image text extraction and recognition using hybrid

approach of region based and connected component methods. International Journal of Engineering Research and Technology (IJERT) **3**(6) (2014)

[46] Kaur, R.P., Jindal, M.K., Kumar, M.: Text and graphics segmentation of newspapers printed in gurmukhi script: A hybrid approach. The Visual Computer **37**(7), 1637–1659 (2021)

[47] Papavassiliou, V., Stafylakis, T., Katsouros, V., Carayannis, G.: Handwritten document image segmentation into text lines and words. Pattern Recognition **43**(1), 369–377 (2010)

[48] Kundu, S., Paul, S., Bera, S.K., Abraham, A., Sarkar, R.: Text-line extraction from handwritten document images using gan. Expert Systems with Applications **140**, 112916 (2020)

[49] Gader, T.B.A., Echi, A.K.: Unconstrained handwritten arabic text-lines segmentation based on ar2u-net. In: Proceedings of the 17th International Conference on Frontiers in Handwriting Recognition (ICFHR), pp. 349–354. IEEE, Dortmund, Germany (2020)

[50] More, V., Kharat, M., Gumaste, S.: Segmentation of devanagari handwritten text using thresholding approach. International Journal of Scientific and Technology Research (IJSTR) **9**(3), 2277–8616 (2020)

[51] Kohli, M., Kumar, S.: Segmentation of handwritten words into characters. Multimedia Tools and Applications **80**, 22121–22133 (2021)

[52] AlKendi, W., Gechter, F., Heyberger, L., Guyeux, C.: Advancements and challenges in handwritten text recognition: A comprehensive survey. Journal of Imaging **10**(1), 18 (2024)

[53] Gedraite, E.S., Hadad, M.: Investigation on the effect of a gaussian blur in image filtering and segmentation. In: Proceedings of ELMAR-2011, pp. 393–396. IEEE, Zadar, Croatia (2011)

[54] Goh, T.Y., Basah, S.N., Yazid, H., Safar, M.J.A., Saad, F.S.A.: Performance analysis of image thresholding: Otsu technique. Measurement **114**, 298–307 (2018)

[55] Said, K.A.M., Jambek, A.B., Sulaiman, N.: A study of image processing using morphological opening and closing processes. International Journal of Control Theory and Applications **9**, 15–21 (2016)

[56] Said, K.A.M., Jambek, A.B.: Analysis of image processing using morphological erosion and dilation. In: Journal of Physics: Conference Series, vol. 2071, p. 012033. IOP Publishing, Bristol, UK (2021)

[57] Arbelaez, P., Maire, M., Fowlkes, C., Malik, J.: Contour detection and hierarchical image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **33**(5), 898–916 (2010)

[58] Dutta, A., Gupta, A., Zissermann, A.: VGG Image Annotator (VIA). Software available at http://www.robots.ox.ac.uk/~vgg/software/via/ (2016)

[59] Phillips, I.T., Chhabra, A.K.: Empirical performance evaluation of graphics recognition systems. IEEE Transactions on Pattern Analysis and Machine Intelligence **21**(9), 849–870 (1999)

[60] Fischer, A., Frinken, V., Fornés, A., Bunke, H.: Transcription alignment of latin manuscripts using hidden markov models. In: Proceedings of the 2011 Workshop on Historical Document Imaging and Processing, pp. 29–36 (2011)

[61] Sánchez, J.A., Romero, V., Toselli, A.H., Vidal, E.: READ Dataset Bozen. Dataset available at https://doi.org/10.5281/zenodo.218236 (2016). https://doi.org/10.5281/zenodo.218236

[62] Garz, A., Fischer, A., Bunke, H., Ingold, R.: A binarization-free clustering approach to segment curved text lines in historical manuscripts. In: Proceedings of the 12th

International Conference on Document Analysis and Recognition (ICDAR), pp. 1290–1294. IEEE, Washington, DC, USA (2013)

[63] Pastor-Pellicer, J., Garz, A., Ingold, R., Castro-Bleda, M.J.: Combining learned script points and combinatorial optimization for text line extraction. In: Proceedings of the 3rd International Workshop on Historical Document Imaging and Processing, pp. 71–78 (2015)

[64] Grüning, T., Leifert, G., Strauß, T., Michael, J., Labahn, R.: A two-stage method for text line detection in historical documents. International Journal on Document Analysis and Recognition (IJDAR) **22**(3), 285–302 (2019)

[65] Li, X.H., Yin, F., Xue, T., Liu, L., Ogier, J.M., Liu, C.L.: Instance aware document image segmentation using label pyramid networks and deep watershed transformation. In: Proceedings of the International Conference on Document Analysis and Recognition (ICDAR), pp. 514–519. IEEE, Sydney, Australia (2019)

[66] Kiessling, B.: A modular region and text line layout analysis system. In: Proceedings of the 17th International Conference on Frontiers in Handwriting Recognition (ICFHR), pp. 313–318. IEEE, Dortmund, Germany (2020)