Multimedia Tools and Applications
https://doi.org/10.1007/s11042-025-20945-5

1250: HERITAGE PRESERVATION IN THE DIGITAL AGE: ADVANCES IN MACHINE LEARNING, MONOMODAL AND MULTIMODAL PROCESSING, AND HUMAN-MACHINE INTERACTION

Handwritten ink segmentation algorithms for hyperspectral images of historical documents

Marco Buzzelli[1] · **Francisco Moronta-Montero**[2] · **Ramón Fernández-Gualda**[2] · **Ana Belén López-Baldomero**[2] · **Juan Luis Nieves**[2] · **Eva M. Valero**[2]

Received: 9 February 2024 / Revised: 12 May 2025 / Accepted: 23 May 2025

Abstract
This study focuses on the segmentation of handwritten ink in historical documents using hyperspectral imaging in two spectral ranges (visible and near-infrared). Binarization is useful as a pre-processing step for material identification using the reflectance spectra. To showcase the challenges of using hyperspectral imaging, classical single-band (Howe and Sauvola) and deep learning-based algorithms (DeepLabv3, SAM, DINOv2) are compared. For algorithms that take a single image as input, a procedure is presented to select the optimal band for binarization. The deep learning-based semantic segmentation algorithm DeepLabv3 uses the full spectrum instead. A hyperspectral database encompassing 226 samples is introduced as a benchmark to compare the performance of the algorithms. The study also introduces a novel semi-automatic method for generating ground truths, which are needed for computing performance metrics. DeepLabv3 performs on par with the best traditional algorithm in both ranges, but overall, it offers more consistent and reliable results. DINOv2 demonstrates good semantic understanding in separating foreground and background but suffers from limited spatial resolution. Conversely, SAM excels at capturing fine details but lacks the ability to identify text regions. The binarization quality obtained with three-channel images is also assessed, generally resulting in lower average performance. Our findings contribute to the advancement of technologies for the analysis of text in documents of historical interest.

Keywords Image binarization · Historical documents · Hyperspectral imaging

1 Introduction

Historical documents hold immense cultural and scientific significance. Serving as a testament to our legacy, manuscripts constitute a valuable source for knowledge retrieval, making them crucial elements of cultural heritage worthy of preservation and study [1–4].

The main goal of this study is to develop and evaluate a methodology for the task of binarization in hyperspectral images (HSI) of handwritten documents. For this purpose, several algorithms will be tested to assess their performance, focusing in particular on the

Extended author information available on the last page of the article

Published online: 05 June 2025

challenges posed by the increased dimensionality and variability of spectral data. Besides, the intrinsic value of spectral information for document binarization will be highlighted by comparing the quality of binarization between spectral and conventional three-channel digital images. The following considerations will introduce in a general way both topics — hyperspectral imaging and binarization as a segmentation procedure to separate text from substrate. Then, the main contributions of the paper will be summarized.

Hyperspectral imaging as an analytical tool in the context of historical documents is a growing field that offers several advantages: it is non-destructive, portable, relatively fast and low-cost in comparison to other techniques, and has high spatial and spectral resolution [5, 6]. Access to spectral data enables the identification of materials in documents [7], providing crucial insights into the authenticity and age of the ink [8]. In addition, it can be used to recover degraded texts [9], and retrieve features not discernible by the human eye [1, 10]. In the case of illuminated manuscripts, the technique can be used to extract information about the distribution of pigments, in a similar way as it is done in drawings and paintings [11–13].

Binarization is a critical preprocessing step in which a multi-tone image is converted into a binary image. The substrate pixels (parchment or paper in the case of documents) are usually labelled in black and the foreground pixels (text and illuminations) are labelled in white [14]. This binary image can then be used, in the case of historical documents, for further processing, such as Optical Character Recognition (OCR), page layout analysis, image enhancement, or material classification [6]. Due to the state of conservation of some documents of historical interest, their binarization can be challenging [15]. Some of the problems that can occur are ink fading due to humidity and paper deterioration [16], bleed-through if both sides of the paper are written [17], and the presence of stains, smear, and creases, uneven illumination, varying font size and thin strokes [2, 14]. Despite numerous past attempts in this domain, the binarization of degraded documents using conventional images remains an open challenge [3]. In our work, both ink strokes and bleed-through text are treated as image foreground to ensure comprehensive analysis and accurate segmentation.

Thresholding is the most straightforward method for binarization, with two main approaches: global thresholding and local thresholding. Global thresholding defines a single threshold for the entire image, making it fast but less effective for documents with complex backgrounds [18]. In contrast, local thresholding adapts to image areas, providing a more flexible solution. A widely used local adaptive thresholding method was proposed by Sauvola et al. [19], based on the assumption that the local intensity distribution of text pixels is different from the local intensity distribution of background pixels. However, such algorithms do not work well with background noise [20], and a fixed window is not optimal if different font sizes and stroke widths are present [14]. Additional drawbacks of traditional image segmentation algorithms are limited accuracy and non-uniformity [21].

A different binarization method, not based on local thresholding, was proposed by Howe et al. [22]. It is based on random Markov field classifiers and has also been used in different binarization contests, being one of the best performing algorithms, and serving as a basis for other robust algorithms [23, 24].

In order to solve the problems of traditional algorithms, deep learning models have recently become more popular for pixel-wise image inference [25], and specifically for binarizing document images [14]. Although these algorithms may exhibit slower processing speeds, they are generally characterized by higher accuracy values [26]. In the field

of semantic segmentation, the DeepLab series [27] are classic models that have achieved impressive performance on a variety of datasets [28]. Recent foundation models [29, 30] in deep learning have also shown excellent performance in semantic understanding of natural images. However, the applicability of these models in separating text from background in historical documents has yet to be fully explored.

According to Ciortan et al. [1], using hyperspectral images can enhance ink separation from the substrate. They used a simple distance-based classification, without considering common binarization algorithms, and only performed tests in the VNIR (visible to near-infrared) range. Some previous studies have used multispectral images in the context of document binarization [5, 31–37] with promising results. However, none of them have explored the short-wavelength infrared (SWIR) range in historical documents, nor compared SWIR with visible range results. The SWIR range can be potentially useful if false color images are used to highlight the presence of inks that fade at these wavelengths, like those with metallo-gallate components.

The main contributions of this study are the following: first, a comparative analysis of the performance of two traditional algorithms – Sauvola and Howe – and three deep learning based models – DeepLab, SAM, and DINOv2 – in the context of historical document binarization with hyperspectral image data in both SWIR and visible ranges. Second, a new hyperspectral dataset containing 226 samples is used for benchmarking, extracted from the Hyperdoc project database [38]. This represents a relevant contribution to researchers interested in historical documents and mock-up samples made with historically documented recipes. Third, a new semi-automatic procedure to obtain the Ground Truths (GTs) images of the samples is proposed to facilitate performance assessment for binarization tasks. The analysis of results leads to new insights into the usefulness of spectral information in different spectral ranges for binarization, and the limitations of current segmentation algorithms for this task in challenging cases.

2 Methods

In this section we present the methods underlying our research.

The approach includes three main steps, corresponding to the subsections:

1. Hyperspectral data collection: hyperspectral fragments are captured across VNIR and SWIR spectral ranges using Resonon Ltd. cameras, and annotated. They are categorized into a training set, and two test sets with different levels of difficulty.
2. Segmentation methods: five binarization approaches are implemented. The Howe and Sauvola methods depend on a small number of tunable parameters. DeepLab is a trainable deep-learning model for semantic segmentation, SAM is a general-purpose model for prompt-based segmentation, and DINOv2 is a foundation model for semantic feature extraction.
3. Selection of hyperspectral image channel: to optimize the input for traditional algorithms, the best-performing single channel from the hyperspectral data is selected using a Signal-to-Noise Ratio (SNR) ber-band metric.

Figure 1 illustrates the different steps followed in the methodology of this study, as developed in the following subsections.

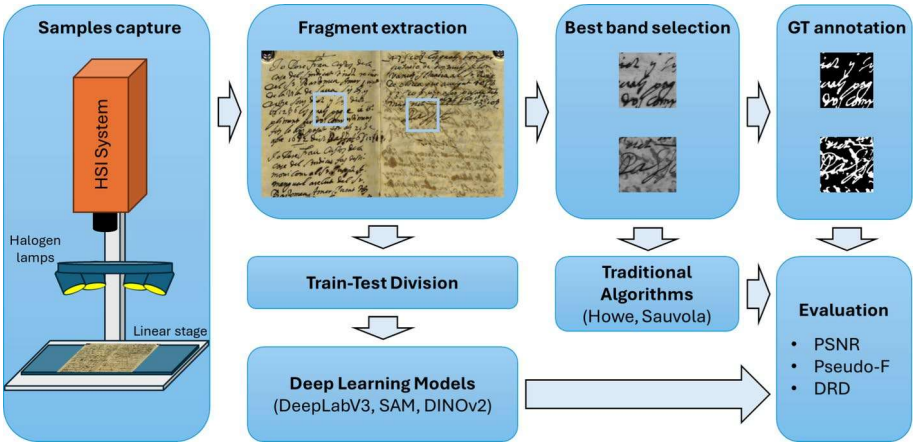

Fig. 1 Steps followed in the methodology for this study on handwritten ink segmentation algorithms for hyperspectral images of historical Documents

2.1 Hyperspectral data collection

2.1.1 Image fragments characteristics

The samples used in this study are fragments of spectral images of documents containing different types of inks of historical interest on different substrates. There are a total of 111 fragments in the VNIR range and 115 fragments in the SWIR range (see Section 2.1.2 for details of both ranges). The fragments have variable sizes of around 3cm^2, and digital dimensions ranging from 45×30 to 150×150 pixels. The fragments were extracted from several types of documents, all containing only ink and substrate.

An overview of the primary characteristics of these samples is provided in Table 1.

Table 1 Primary characteristics of the samples, including fragment count, ink types, substrates, time period, and source locations

Category	VNIR fragments	SWIR fragments	Inks	Substrates	Period	Source
Synthetic	50	55	Iron gall, sepia, lampblack, madder lake red dye	Somerset® paper, watercolor paper, modern parchment	Contemporary (recipes from 13 th-17 th centuries)	Synthesized
Alhambra	34	32	Chinese ink, red ink	Translucent paper	Early 20 th century	Alhambra Museum Archive
Alamas	11	12	Iron gall, sepia/ carbon black)	Parchment (authentic), cotton/ linseed paper (forgery)	1461 (authentic), 1487 (forgery)	Archive of the Royal Chancellery of Granada
Selva	16	16	Pure/mixed iron gall	Linseed or hemp fiber paper	1682-1683	Historical Archive of the Town of Selva
Total	111	115	-	-	-	-

Modern synthetic samples contain inks and mixtures all bound with Arabic gum and elaborated according to traditional recipes from the 13 th to 17 th centuries [39]. Some of them were dyed with madder lake red (*rubia tinctoria*), a practice common in the Andalusian region during the Arab domination [40], or with diluted lamp black ink. The substrates used for the synthetic samples are Somerset® paper with different treatments (gelatin, wheat starch, or gum Arabic), watercolor paper, and modern parchment.

"Alhambra" samples were extracted from spectral images of a collection of nineteen hand-made architectural plans or transfers depicting sculptures present in some of the Alhambra buildings. The type of ink is likely to be Chinese ink (carbon-based), but this is a hypothesis based on materials currently used for the same purpose in the period of the documents; the samples that we have used for this study have not been analyzed to determine the ink composition yet. They are preserved in the Alhambra Museum Archive [41].

"Alamas" samples were extracted from two instances of a set of official documents of the Islamic period called Alamas. One of the documents is of authentic Islamic origin, while the other has been proven to be a forgery commissioned by a Christian noble for the purpose of asserting his noble origins. Both documents are preserved at the Archive of the Royal Chancellery of Granada, and were studied for preservation purposes during 2022 [42].

Finally, "Selva" samples were extracted from a small notebook documenting several commercial transactions found in the Historical Archive of the Town of Selva (Mallorca Island, Spain) [43]. The inks used in this document were determined by X-Ray Fluorescence (XRF) techniques through an internal investigation.

In Fig. 2 we show two instances of pages of documents used in the study, with the fragments extracted from them highlighted in yellow. In the Selva page, the deterioration due to ink transfer from the back of the page can be observed.

The fragments were divided into three subsets for each spectral range: Train, Easy Test and Hard Test. Most of the samples approximately correspond across both ranges, but in some cases it was not possible to find exactly the same area of the documents in both ranges. For this reason, there are a different number of samples in some subsets in both ranges.

In Table 2, the subdivision of the samples among the subsets is presented, along with information about the source documents, number of fragments, number of pixels, inks, and substrates. As can be observed, the Easy Test subsets are formed by samples extracted from the same document set as the Train subset, while the Hard Test contains entirely different documents that are naturally aged, none of them dated after the 17 th century and showing clear deterioration due to time and preservation conditions (see Fig. 2 below). For some of the binarization algorithms used (see Subsections 2.2.1 and 2.2.2), the division between training and test subsets is not relevant because they are not learning-based. But it is very relevant for the deep-learning-based approach (see Subsection 2.2.3). In this case, the bias introduced by training data sources may have a significant impact on the measured test performance. As such, it is particularly useful to present results on two test scenarios: one that simulates a use case more similar to the training knowledge base (Easy Test), and one that might be more representative of novel real-world applications (Hard Test).

2.1.2 Instrumentation specifications

We used two cameras from Resonon Ltd. coupled to a linear stage to capture the full cubes from where the mini-cube samples were extracted afterwards. The first image capture device

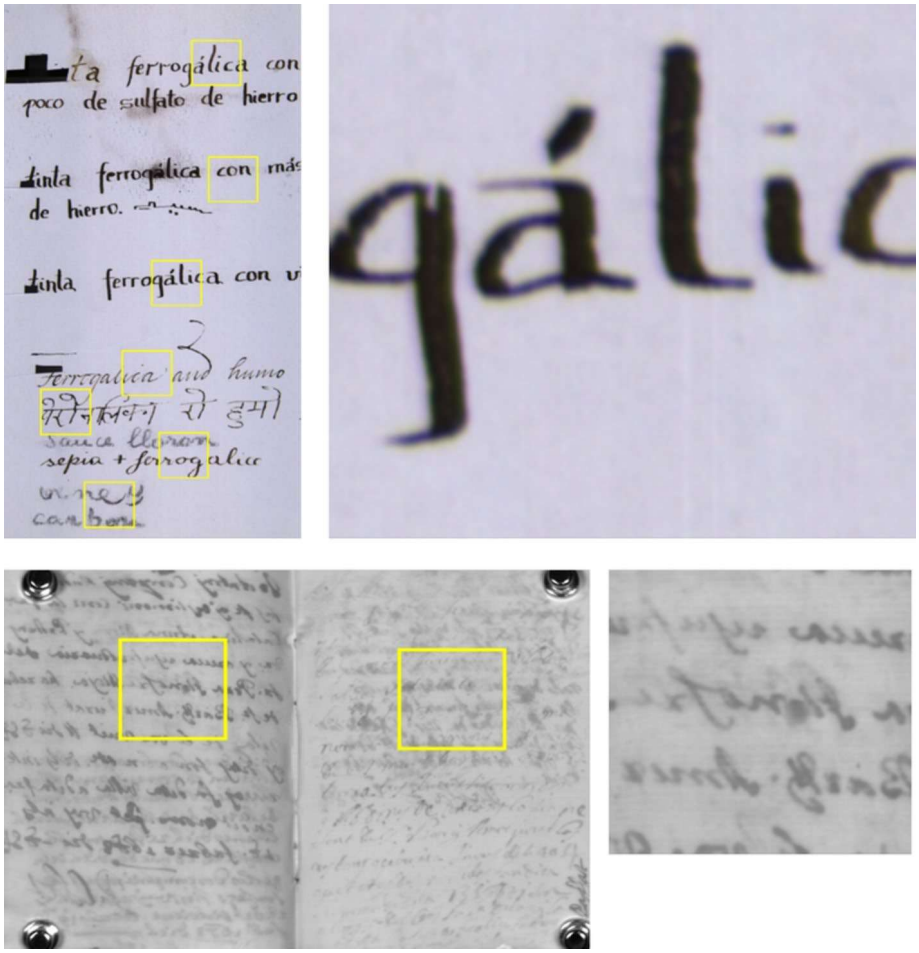

Fig. 2 False color image of a document within the synthetic sample set in the VNIR range (above) and from a page of the Selva manuscript in the SWIR range (below). The extracted fragment areas are marked in yellow. On the right side, one of the fragments extracted from each of the images with size 150×150 pixels is shown. The false color bands are (645, 565 and 440) nm for VNIR and (1015, 1140 and 1220) nm for the SWIR range

(Pika L) covers the spectral range from 380 to 1080 nm (VNIR range) with 900 pixels per line and a field of view (FOV) of 13.5 cm at the working distance (60 cm approximately), resulting in a spatial resolution of 0.15 mm/pixel. [44]. The second (Pika IR+) covers the range from 888 to 1732 nm (SWIR range) with 640 pixels per line and a FOV of 14.5 cm at the working distance of 40 cm approximately, resulting in a spatial resolution of 0.227 mm/pixel [45]. We cropped the extremes of the range, obtaining 121 bands in VNIR from 400 to 1000 nm and 161 bands in SWIR from 900 to 1700 nm. The sampling interval was 5 nm for both ranges. In all cases, both dark and flat field correction with a white reference surface were applied. The light source was a set of four halogen lamps oriented to avoid specular reflection from the samples.

Table 2 Summary of subset features including names, documents used to build them (and number of fragments from each document), total number of fragments, total number of pixels, and materials

Sample subset	Document of origin	Fragments	Pixels	Inks	Substrates
Train VNIR	Synthetic (43), Alhambra (26)	69	798550	Iron gall, Sepia, Lamp black, Mixed	Somerset Watercolor, Parchment, Translucent paper
Easy Test VNIR	Synthetic (7), Alhambra (8)	15	86700	Iron gall, Sepia, Lamp black, Mixed	Somerset, Translucent paper
Hard Test VNIR	Alama (11), Selva (16)	27	582500	Iron gall or Mixed iron gall	Parchment, Linseed+cotton paper, Linseed or hemp paper
Train SWIR	Synthetic (42), Alhambra (26)	68	689875	Iron gall, Sepia, Lamp black, Mixed	Somerset, Watercolor, Parchment, Translucent paper
Easy Test SWIR	Synthetic (13), Alhambra (6)	19	77500	Iron gall, Sepia, Lamp black, Mixed	Somerset, Translucent paper
Hard Test SWIR	Alama (12), Selva (16)	28	571000	Iron gall or Mixed iron gall	Parchment, Linseed+cotton paper, Linseed or hemp paper

2.1.3 Ground truth (GT) data

For each fragment, a GT binary image was generated using a semi-automatic procedure (see steps in Fig. 3). The procedure involved first selecting the band with the highest contrast between ink and background (Fig. 3 (b)), as described in subsection 2.3. Then, foreground skeleton was extracted using the *bwskel* function in Matlab R2023a, which applies the medial surface axis thinning algorithm [46] (see Fig. 3 (d)). The skeleton was then forced to increase its width until the intensity of surrounding pixels matched the average of the borders of a Canny edge detector. Figure 3 (c) shows the detected borders, while Fig. 3 (e) presents the result after skeletal growth. This is a variation on the method proposed in [47], in which the skeleton was manually corrected and then forced to grow until it met those borders.

Once the automatic GT is obtained, it is manually reviewed and may be corrected using three different approaches. First, the intensity threshold at which growth stops can be modified to include or exclude lighter strokes based on human criteria. Secondly, several functions included in Matlab R2023a version can be applied to process the original image. Functions like *imadjust* and *locallapfilt*, based on [48], help to increase contrast. For noise reduction, a flat-field correction (*imflatfield*) or median filtering (*medfilt2*) can be used. This image processing helps skeleton identification and therefore improves the quality of the automatic GT generation. Finally, once we have the best possible GT with the algorithm, GIMP software [49] can be used for the final check (see Fig. 3 (f) for the final result). We used partially transparent layers to superimpose the GT and the reference band images, and either expand or erode the automatic GT in the portions of the image that required it. The main limitations of this approach are that in this checking stage, the distinction between ink and background is made based on visual criteria, a matter which is addressed in the experimental section with tests based on mathematical morphology. Also, the GT for the reference

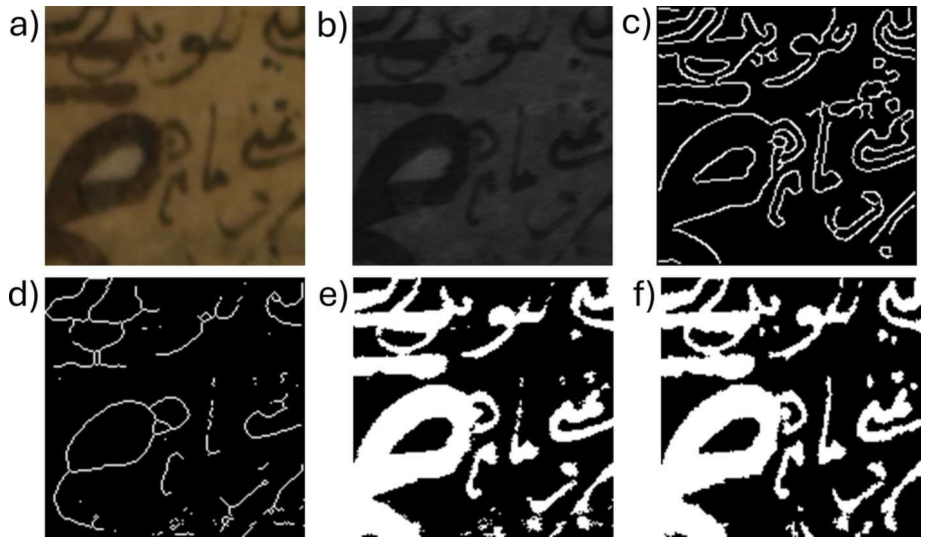

Fig. 3 Steps for GT creation for a fragment of the Alamas set: (**a**) VNIR false color image (bands 645, 565, and 440 nm), (**b**) single-band image with highest contrast (420 nm), (**c**) foreground boundaries detected using the Canny edge detector, (**d**) foreground skeleton, (**e**) skeletal growth result, and (**f**) final GT after manual correction in GIMP

band is not necessarily valid for all bands, since iron gall and sepia inks tend to fade in the SWIR range [7].

Nevertheless, we are convinced that this is the best strategy compatible with time constraints for providing the GT to our data. Fully automatic methods are available for building GT images, but they can be described ultimately as binarization algorithms. So instead of testing our collection of algorithms against another particular algorithm, we preferred to test it against the performance of a human observer in the binarization task.

In Fig. 4 two instances of false color, reference bands, and GT images are shown, one in the VNIR and one in the SWIR range. The very small details in the ink traces in the image below are not exactly transferred into the GT because they are very difficult to reproduce manually. This is an example of systematic errors induced by the GT used for evaluation. We quantify in Section 3.2.5 the impact of these errors, by testing the sensitivity of segmentation metrics to small imperfections in the annotation, and we show that they do not substantially affect the performance of the binarization methods in a significant way.

2.2 Segmentation methods and models

2.2.1 Sauvola method

The Sauvola method [19] is a local adaptive thresholding algorithm that is similar to the method proposed in [50]. However, the Sauvola method uses a different formula to calculate the local mean and standard deviation of pixel intensities. It is also less sensitive to noise than Bradley's Local Image Thresholding [51]. The threshold for this method is calculated as

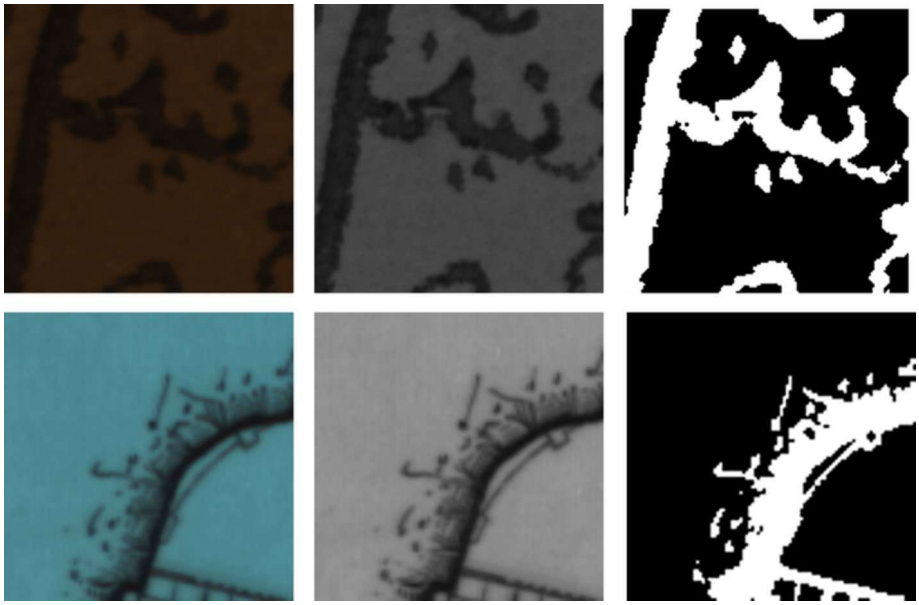

Fig. 4 (Left column) False color image of two of the fragment images in the Hard test VNIR (above, bands 645, 565, and 440 nm) and Train SWIR (below, bands 1000, 1200, and 1600 nm). (Center column) Reference band image used for building the GT (660 nm above and 1120 nm below). (Right column) GT image corresponding to the fragments

$$T(x,y) = m(x,y) \times \left[1 + k \times \left(\frac{s(x,y)}{R} - 1\right)\right], \tag{1}$$

where *T(x,y)* represents the local threshold for pixel *(x,y)*; *m(x,y)* and *s(x,y)* are the average and standard deviation of pixel intensities in the local neighborhood, k is an empirically chosen parameter, and R is the dynamic range of pixel values. The dynamic range factor is a measure of the relative intensity of a pixel compared to the maximum possible intensity. A higher value of R indicates that the pixel is brighter than most of the pixels in the local window and helps to preserve objects that are well-contrasted with the background. Empirically determined parameters, k and R, are critical for its performance. In this study, the value of k is set to 0.4, and R is the maximum standard deviation of the window used at the evaluated pixel. The local window size is set to 1/3 of the image width by 1/3 of the image height. To handle border pixels effectively, padding is applied by replicating the values of the border pixels.

2.2.2 Howe method

The binarization method by Howe et al. [22] is not based on local thresholding, but on a classifier that labels the pixels based on minimizing a random-field Markov energy function. This function uses the Laplacian of the image for separating the two classes, enhancing the uniformity of the background areas. It also incorporates edge detection into the energy function, aiming to align the classes boundaries with the detected edges. Since it is a parametric

method, it also proposes a way to automatically find the best performing parameters. The energy function for a given binarization B is computed according to this equation:

$$\begin{aligned}\mathcal{E}_I(B) &= \sum_{x=0}^{p}\sum_{y=0}^{q}\left[L_{xy}^{0}\left(1-B_{xy}\right)+L_{xy}^{1}B_{xy}\right]\\ &+\sum_{x=0}^{p-1}\sum_{y=0}^{q}C_{xy}^{h}\left(B_{xy}\neq B_{x+1,y}\right)\\ &+\sum_{x=0}^{p}\sum_{y=0}^{q-1}C_{xy}^{v}\left(B_{xy}\neq B_{x,y+1}\right),\end{aligned} \tag{2}$$

where $L_{xy}^{\{0,1\}}$ are penalties for mismatch between the class and the appearance of the pixel (either text or background), while C_{xy}^{h} and C_{xy}^{v} are weights that try to prevent fast changes of classes among adjacent vertical or horizontal pixels (irregularities). C_{xy}^{h} and C_{xy}^{v} depend on a parameter that has an image-dependent optimal value. Besides, the Canny edge detector also needs an additional parameter. These two parameters can be set automatically. The original Howe's implementation [52] for automatic parameterization aims to suppress bleed-through pixels and selecting only foreground text, while in our case the aim is to detect all pixels containing ink. So we have modified the Canny edge threshold parameter specifically for those images that have some bleed-through.

2.2.3 DeepLab model

We selected the DeepLabv3 [27] model as the first neural architecture for our deep-learning-based approach, based on its well-documented performance in semantic segmentation tasks. DeepLab has proven its ability to capture fine-grained details and accurately delineate object boundaries, leveraging the dilated convolution technique [53] to integrate multi-scale contextual information. These features make it potentially effective in handling complex and intricate patterns in historical documents. Nonetheless, we opted to increase the level of detail by feeding an upscaled version of the spectral image to the model, and subsequently downscaling the segmentation output. The annotated fragments used in training and validation, ranging from 30 pixels to 150 pixels per side, are upscaled to a fixed 512×512 resolution. This process enables the model to adapt to various spatial scales, allowing it to learn to capture both fine-grained and larger structural information present in the historical documents.

The original DeepLab architecture is designed to work with RGB images, as opposed to hyperspectral images. Therefore, a primary modification entails replacing the first convolutional layer of DeepLab, originally designed to process 3-channel inputs, with a new convolutional layer that can accept N channels, corresponding to the number of spectral bands in our hyperspectral data. This adaptation allows us to seamlessly integrate hyperspectral information into the model while preserving the subsequent layers and their learned weights, a critical aspect of our adaptation process since we aim to leverage transfer learning [54].

Deep learning models are capable of learning some degree of normalization during training. Despite this, applying appropriate normalization techniques remains beneficial, as it can expedite convergence and enhance the model's overall robustness. Taking inspiration from the inherent normalization in the Sauvola and Howe methods, we implemented an efficient neural local normalization. Specifically, we construct a convolutional filter of size $f \times f$ with dilation factor l, using uniform weights that sum up to 1, and zero bias. This

filter is then convolved with the input image I, effectively producing a normalization matrix J having the same resolution. Finally, a local-dilated normalized image is obtained by per-pixel division between I and J, and provided as input to DeepLab. This process enables the definition of an arbitrarily-wide receptive field for the computation of local normalization statistics, avoiding computationally-expensive dense statistics.

Preliminary experiments were conducted to accurately select a proper configuration of the DeepLab model, in terms of backbone, normalization, and data cleanup. Three backbone architectures have been considered for feature extraction, namely: ResNet101 [55], ResNet50 [55], and MobileNetV3 [56], with ResNet101 yielding the best performance. For the local normalization, a filter of size 5×5 with dilation parameter 8 has been found to strike the best balance between computational resources and segmentation quality. Bands at the extrema of the acquired spectrum are found to be of lower quality, as corroborated by the metrics described in Section 2.3; for this reason, we trim a number of trailing channels from both the beginning and end of the spectrum, with 4 channels from each end producing the best results.

2.2.4 Segment anything model

The Segment Anything Model (SAM) [30] is a general-purpose segmentation model developed with a prompt-based paradigm, allowing for both interactive and automatic segmentation. Its architecture consists of an image encoder, a prompt encoder, and a lightweight mask decoder, enabling it to generalize across a wide variety of segmentation tasks without task-specific fine-tuning.

Since SAM was originally designed for natural RGB images, we assess its performance using both single-band images (selected as the most informative spectral band) and three-channel false-color composites.

A key advantage of SAM is its ability to process various types of input prompts, including sparse user-defined points and bounding boxes. However, SAM does not inherently assign semantic meaning to its dense segmentation results, meaning its automatic mode will indiscriminately segment different regions of the document. To address this limitation, we explore SAM in its "Language SAM" configuration [57], which enables segmentation through a multimodal approach. In this setup, the input image is processed with Grounding DINO [58], an open-world object detector tasked with identifying text elements within the document. This allows for a more targeted segmentation, guiding SAM to focus on ink regions rather than extraneous features.

2.2.5 DINOv2 features

DINOv2 features [29] have demonstrated remarkable capability in distinguishing semantic content in open-world scenarios. As a self-supervised vision transformer model, DINOv2 learns rich feature representations without the need for labeled data, making it particularly valuable for applications across a variety of domains. By leveraging learned feature embeddings, in fact, we anticipate it possible to highlight structural patterns within the document, including ink traces, while mitigating the influence of background textures and degradations.

A key limitation of DINOv2 features, however, is their relatively low spatial resolution. Since the model's embeddings are derived from hierarchical transformer layers, the

resulting feature maps tend to be significantly downsampled compared to the input image. This can be problematic for ink segmentation, where fine details are important to detect. To address this, supersampling techniques are often employed to enhance spatial resolution while preserving the integrity of the learned features. One of the most recent and effective methods for feature supersampling is FeatUp [59], which reconstructs high-resolution representations by learning an upsampling function tailored to feature embeddings: unlike naive interpolation-based approaches, FeatUp instead adapts to the structure of the feature space. By applying FeatUp to DINOv2 features, we aim to recover finer spatial details in the segmentation maps while maintaining the robustness of the underlying semantic representations.

2.3 Selection of hyperspectral image channel

Before employing traditional segmentation algorithms, it is essential to determine the optimal image channel for binarization, as these algorithms require a single-channel input. With a hyperspectral image at our disposal, there exists a multitude of channels to choose from within a broad spectral range. To identify the channel that best supports the task of binarization, an image quality (IQ) metric is employed. Several image quality metrics, such as Gradient Magnitude [60], Sharpness Index [61], Entropy [62], and Signal-to-Noise Ratio (SNR) [63] were considered. Preliminary tests were carried out with these metrics, visually evaluating their results and noting that some exhibited limitations, either focusing on specific image characteristics like sharpness, or on general qualities such as the naturalness of the image.

These limitations, observed in both VNIR and SWIR ranges, led us to ultimately select the SNR-based metric. The SNR-based formula [63] is derived as follows:

$$SNR(\lambda) = 10 \times \log_{10}\left(\frac{{I_{ave}}^2}{\sigma^2}\right), \tag{3}$$

where I_{ave} represents the mean intensity of pixels in the image, σ is the standard deviation of pixel intensities, and λ denotes the channel index. In the context of this metric, I_{ave} can be interpreted as the signal, while σ can be regarded as the noise in the SNR formula. The channel with the lowest SNR, as determined by this formula, is chosen for subsequent image segmentation with the traditional approaches, since it corresponds to the channel with the highest standard deviation, and therefore, the greatest contrast. Two examples with channels selected using this approach can be found in the central column of Fig. 4.

3 Experiments and results

3.1 Binarization quality evaluation metrics

To comprehensively assess the quality of the binarization results, we employ evaluation metrics that compare the algorithmic predictions to the semi-automatically created GT. The evaluation process involves binarized images, where the two primary classes are foreground

(indicating the presence of ink, labeled in white) and background (no ink presence, labeled in black). The following three metrics are used for this evaluation:

3.1.1 Peak signal-to-noise ratio (PSNR)

Peak Signal-to-Noise Ratio is a widely used metric for evaluating the quality of segmented or binarized images [64]. It quantifies the level of distortion between the ground truth and the segmented image. PSNR is calculated as:

$$PSNR = 10 \times \log_{10}\left(\frac{{I_{max}}^2}{MSE}\right), \tag{4}$$

where I_{max} is the maximum possible pixel value (255 in our case), and *MSE* is the Mean Squared Error between corresponding pixels in the ground truth and segmented images. A higher PSNR value indicates a lower level of distortion, implying a more accurate binarization. PSNR values range from 0 to ∞, but an acceptable PSNR value in the context of segmentation is usually around 20 [65].

3.1.2 Pseudo-F measure

Given the subjectivity involved in creating ground truths, particularly around the edges, it was necessary to implement a weighted measure that takes into account the edges of the GT in order to enhance document-oriented evaluation outcomes. The Pseudo-F Measure offers an alternative approach to assessing performance, similar to the traditional F-Measure [66]. However, it employs the pseudo function with respect to recall and precision instead of their direct functions [67]. Pseudo-Recall (*pREC*) and pseudo-Precision (*pPR*) rely on a weighted penalization of pixels surrounding the borders of the ground truth characters, considering both the local stroke width and the distance from the contour of the ground truth. The Pseudo-F Measure is calculated as:

$$Pseudo{-}F = \frac{2 \times pPR \times pREC}{pPR + pREC} \tag{5}$$

In Pseudo-Recall (*pREC*), the weights assigned to the foreground of the ground truth are adjusted based on the local stroke width. Conversely, in pseudo-Precision (*pPR*), the weights are confined within a region that extends to the background of the ground truth, considering the stroke width of the nearest ground truth component. A higher Pseudo F Measure indicates a better balance between precision and recall. Pseudo-F Measure values range from 0 to 100.

3.1.3 Distance reciprocal distortion (DRD)

The Distance Reciprocal Distortion Metric (DRD) is a measure employed to assess the visual distortion in binary document images [68]. This metric considers the distortion for each flipped pixel and the number of non-uniform (not exclusively black or white pixels) 8×8 blocks in the ground truth image. A “flipped pixel” refers to a change in the binary

value of a pixel in an image (from black to white or vice-versa) representing a distortion or alteration, and its visibility is influenced by factors such as the proximity of pixels and the observer's focus. Even a subtle change in a single pixel can be noticeable, particularly when it occurs within the viewer's field of vision, emphasizing the importance of pixel relationships in visual perception. The DRD calculation is expressed as follows:

$$DRD = \frac{\sum_k DRD_k}{NUBN}, \tag{6}$$

where *NUBN* is the number of non-uniform 8×8 blocks in the GT image and DRD_k is defined as the weighted sum of the pixels in the 5×5 block of the GT that differ from the centered k^{th} flipped pixel at *(x,y)* in the binarization result image *B*:

$$DRD_k = \sum_{i=-2}^{2} \sum_{j=-2}^{2} |GT_k(x+i, y+j) - B_k(x,y)| \times W_{NM}(i+i_C, j+j_C) \tag{7}$$

W_{NM} is a 5×5 normalized weighted matrix defined in [68], and (i_C, j_C) are the coordinates of its central value, which in our case is equal to (3,3). The DRD metric offers a comprehensive evaluation of visual distortion in binary document images, considering both individual flipped pixels and non-uniform blocks in the ground truth. A lower DRD score indicates more effective binarization.

3.2 Results

In Table 3 the three quality metrics results are shown on both spectral ranges, for each of the two subsets using either single band images or three-channel images built as shown in Fig. 4. The results shown for DeepLabv3 correspond to either full spectral information or three-channel images.

In addition to the three tested algorithms, a practical bound on the metrics is also provided based on mathematical morphology: given the inherent inaccuracies that are present in ground truth annotations, it is useful to include a reference to gauge the sensitivity of the involved metrics to minor variations in the binarization maps. For this reason, we performed a stress test of the metrics by generating an artificial prediction via the application of mathematical morphology to the ground truth.

3.2.1 Comparison between test sets

Overall, with the exception of the Sauvola algorithm in the VNIR range for the DRD metric, the binarization results for the *Easy Test* subset are better than for the *Hard Test* subset. Since neither Sauvola nor Howe algorithms rely on a training set, this means that at least some of the images in the *Hard Test* subset are intrinsically harder to binarize. This is also supported by the fact that the standard deviation is higher for the *Hard Test* subset in both spectral ranges across the three metrics. The *Hard Test* subset is then in principle an adequate choice for the purpose of challenging the trainable model, and this is supported by the

Table 3 Binarization quality metrics for the three algorithms tested in the two spectral ranges (VNIR and SWIR) and for each of the test sets considered: Easy Test and Hard Test with single band for Sauvola, Howe, SAM, and DINOv2, and full spectral information for DeepLab

Method	Range	Easy Test	Hard Test	All Tests	Three-channel
PSNR ↑					
Sauvola	VNIR	13.3 ± 2.4	13.4 ± 2.6	13.4 ± 2.5	13.2 ± 2.5
	SWIR	13.2 ± 4.7	9.7 ± 3.6	11.3 ± 4.1	9.6 ± 4.2
Howe	VNIR	12.6 ± 2.3	10.8 ± 3.6	11.4 ± 3.2	11.6 ± 2.8
	SWIR	14.4 ± 2.8	9.9 ± 3.1	11.9 ± 3.7	10.6 ± 4.0
DeepLabv3	VNIR	15.7 ± 3.2	11.9 ± 3.6	13.3 ± 3.6	11.5 ± 3.4
	SWIR	16.3 ± 4.0	9.6 ± 2.9	12.3 ± 4.7	12.6 ± 4.8
DINOv2	VNIR	12.6 ± 2.9	9.4 ± 2.1	10.6 ± 2.8	9.9 ± 2.9
	SWIR	14.2 ± 3.3	8.5 ± 2.4	10.8 ± 3.9	10.7 ± 4.0
SAM	VNIR	9.3 ± 6.9	1.8 ± 1.3	4.4 ± 5.5	3.8 ± 4.7
	SWIR	10.6 ± 5.1	2.2 ± 3.4	5.6 ± 5.8	6.1 ± 6.0
(Morphology bound)	VNIR	22.7 ± 6.3	20.4 ± 3.2	21.3 ± 4.7	21.3 ± 4.7
	SWIR	17.9 ± 4.1	22.0 ± 5.5	19.5 ± 5.1	19.5 ± 5.1
Pseudo F-Measure (%) ↑					
Sauvola	VNIR	96.8 ± 2.4	96.4 ± 2.9	96.6 ± 2.7	96.4 ± 2.9
	SWIR	94.6 ± 6.5	91.7 ± 4.8	93.0 ± 5.7	90.2 ± 7.7
Howe	VNIR	96.5 ± 2.7	92.6 ± 6.6	94.0 ± 5.4	93.9 ± 6.7
	SWIR	96.3 ± 6.2	92.1 ± 4.8	93.9 ± 5.8	91.7 ± 7.4
DeepLabv3	VNIR	97.1 ± 3.2	94.2 ± 3.6	95.2 ± 3.7	91.8 ± 5.8
	SWIR	96.1 ± 6.4	91.7 ± 4.3	93.5 ± 5.6	93.2 ± 6.0
DINOv2	VNIR	93.6 ± 6.5	89.9 ± 6.0	91.3 ± 6.4	91.9 ± 5.0
	SWIR	95.4 ± 6.5	89.8 ± 4.5	92.1 ± 6.0	90.4 ± 7.2
SAM	VNIR	87.7 ± 28.0	27.7 ± 25.7	62.4 ± 40.2	61.8 ± 37.9
	SWIR	82.7 ± 21.5	38.5 ± 38.0	66.9 ± 35.2	76.3 ± 29.4
(Morphology bound)	VNIR	97.7 ± 2.8	98.4 ± 0.9	98.2 ± 1.8	98.2 ± 1.8
	SWIR	97.3 ± 1.9	97.7 ± 4.5	97.5 ± 3.2	97.5 ± 3.2
DRD ↓					
Sauvola	VNIR	5.8 ± 4.4	4.3 ± 2.6	4.8 ± 3.4	5.3 ± 4.0
	SWIR	6.0 ± 6.3	9.6 ± 4.5	8.0 ± 5.6	13.1 ± 8.6
Howe	VNIR	5.8 ± 3.3	10.4 ± 7.8	8.8 ± 6.9	9.1 ± 9.2
	SWIR	3.2 ± 2.3	9.3 ± 4.6	6.6 ± 4.8	10.4 ± 7.1
DeepLabv3	VNIR	3.4 ± 2.7	6.4 ± 3.9	5.3 ± 3.8	10.8 ± 13.4
	SWIR	2.9 ± 4.7	10.0 ± 4.5	7.2± 5.8	7.3 ± 6.5
DINOv2	VNIR	7.6 ± 7.6	11.7 ± 5.5	10.2 ± 6.6	11.4 ± 6.0
	SWIR	3.8 ± 3.4	13.1 ± 4.0	9.4 ± 5.9	10.5 ± 8.6
SAM	VNIR	57.1 ± 83.5	91.7 ± 43.0	79.4 ± 61.9	82.2 ± 61.4
	SWIR	18.4 ± 28.3	88.8 ± 58.8	60.3 ± 59.7	57.3 ± 60.3
(Morphology bound)	VNIR	1.0 ± 1.2	0.7 ± 0.4	0.8 ± 0.8	0.8 ± 0.8
	SWIR	1.3 ± 0.7	0.9 ± 1.5	1.1 ± 1.1	1.1 ± 1.1

"All Tests" is the weighted average of Easy and Hard test. The "Three-channel" set is composed of false color images

results obtained by DeepLabv3, which are considerably worsened for the *Hard Test* subset in both ranges and the three quality metrics shown in Table 3.

3.2.2 Comparison between VNIR and SWIR results

This comparison offers different results according to the algorithm. For *Sauvola*, the VNIR range has better binarization quality. While for the *Howe* algorithm, the SWIR range results are approximately equal or markedly better. One possible explanation is that Howe is less prone to introduce artifacts when an image has relatively little contrast between text and background, and the background is not uniform. This situation frequently occurs for the *Hard Test* subset. It is precisely for this subset that the differences between spectral ranges are more marked in general. For DeepLabv3, the SWIR range results tend to be worse, save for the PSNR and DRD metrics and the *Easy Test* subset. DINOv2 and SAM present an inverted behavior, with SWIR generally offering better results than VNIR. However, the relevance of this effect should be reconsidered in light of the overall worse performance, as discussed in Section 3.2.3.

As shown in previous studies [7, 69], pure iron gall ink becomes transparent in the SWIR range, particularly in spectral bands above approximately 1500 nm. For the single-band based algorithm, this is not a problem because the best channel selected according to the procedure explained in Subsection 2.3 is never above 1500 nm. But it is a potential problem for DeepLabv3 because the different bands of the input spectra have conflicting information. In fact, if we look at the *All tests* column in Table 3 there is a noticeable effect of the spectral range in the DeepLabv3 results for all three quality metrics.

An alternative way to visualize the comparison between VNIR and SWIR results is to use the examples shown in Fig. 5 for the VNIR and in Fig. 6 for the SWIR range. In these two figures, the first, third, fourth, and fifth rows correspond to similar fragments from both ranges. By looking at these rows, it can generally be observed that the quality of the binarization is worse in the SWIR range, with the differences being more noticeable for the first and fourth rows. The two fragments in these rows belong to the *Hard Test* subset.

3.2.3 Comparison between methods

For the VNIR range, the *Sauvola* method is slightly better than *Howe* (and markedly better according to the PSNR metric and DRD metric for the *Hard Test* subset). The results tend to be much more similar between the two single-band-based algorithms in this range, although in principle the DIBCO competition results [23, 24] indicate that for the datasets used in those competitions, *Howe* is able to outperform *Sauvola*. DeepLab outperforms Sauvola for the *Easy Test* subset, but not for the *Hard Test* subset. On average, the performance of Sauvola and DeepLab is similar, although Sauvola is the best in all three metrics.

For the SWIR range, the situation when comparing the two single-band approaches is reversed. The explanation for this result is similar to the one offered in the previous subsection: *Howe* is less prone to introduce artifacts when there are spurious blotches, highlights, or illumination-induced inhomogeneities in the page, especially when the contrast between background and foreground is less marked. DeepLab outperforms both single-band approaches in this spectral range for the PSNR metric, and performs better than *Sauvola* according to the three metrics. This means that DeepLab can offer more consistency and

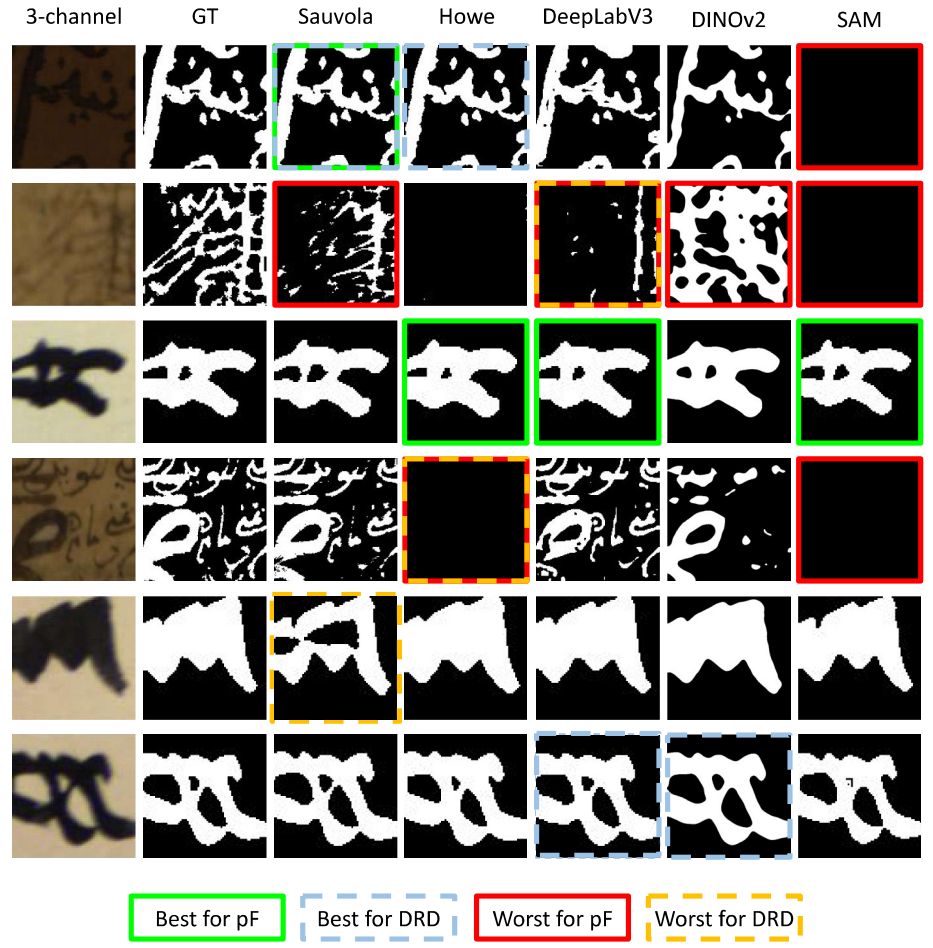

Fig. 5 Binarization examples in the VNIR range. The best and worst fragments according to either Pseudo-F or DRD metrics are highlighted

reliability in the binarizations under relatively poor image quality conditions, and despite the inherent problems of the SWIR range for iron gall inks.

For "text"-prompt SAM the quality of the segmentation is extremely high for images where the text is prominent in the image, and semantically identified as such. However, in scenarios where denser writing is present, or where it is not easily identified as such, language-SAM fails to offer graceful degradation. This significantly impacts the overall performance and its usability for our task. For DINOv2+FeatUp, semantic identification of ink against substrate is significantly improved compared to SAM. However, the model struggles with providing sufficient detail in the segmentation, despite the FeatUp-based augmentation.

In Fig. 5, it can be observed how *Sauvola* is able to deal better than the other two algorithms in general with dark substrate and faint traces conditions (rows 1 and 2). However, very thick traces pose a clear problem (see row 5). In this fragment, the thickness of the trace

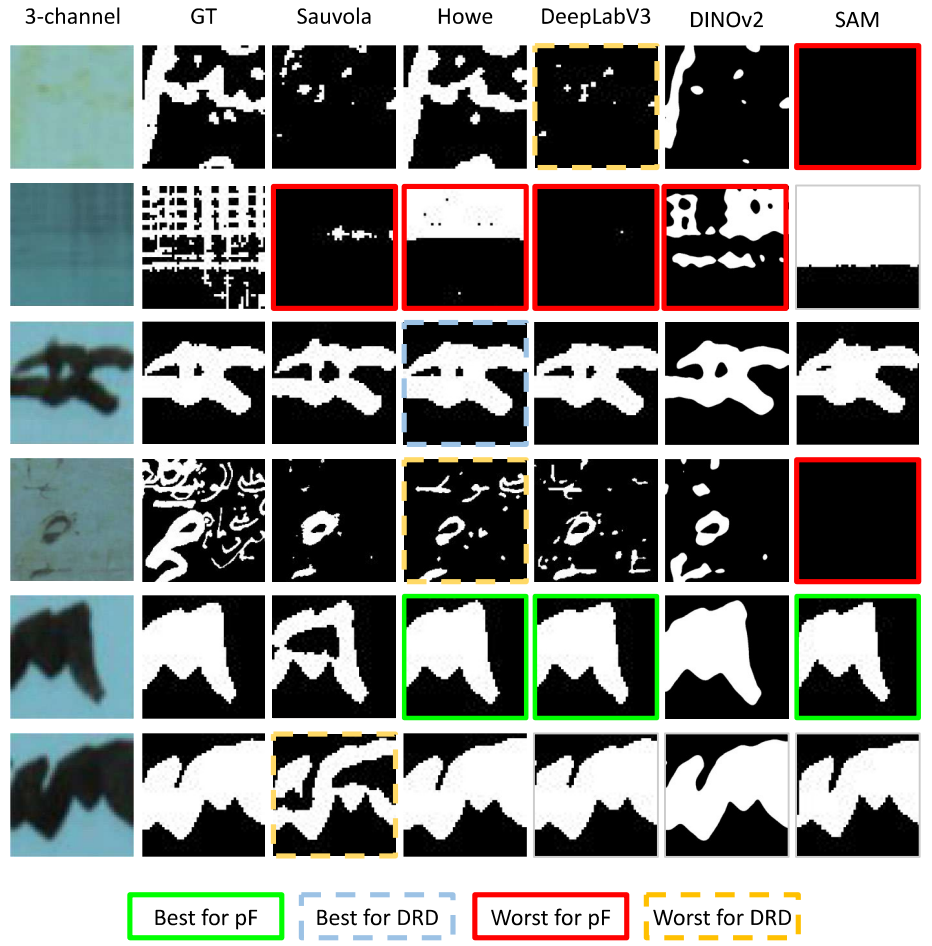

Fig. 6 Binarization examples in the SWIR range. The best and worst fragments according to either Pseudo-F or DRD metrics are highlighted

is higher or of the same order as the window size (one third of the image). For central pixels in the trace, the standard deviation is very low, and this results in a local threshold value that is similar to the one found for the background pixels. Lowering the value of the k parameter in (1) from 0.4 to 0.1 can improve this situation, although it does not solve it completely, and this change worsens the average results of *Sauvola*, so in the end we chose to keep the k value unchanged. Lowering the value of k results in a higher local threshold for images that have less contrast with the background and relatively thin traces, which is more common in the fragments belonging to the test sets. *Howe*, on the other hand, is almost not able to find any foreground pixels for the fragments in rows 2 and 4. This is related to the image dependency of the optimal parameters related to Canny edge detection. We used the values recommended in [22] (algorithm version 3), but these parameters are clearly not optimal for our particular set of images. Again, changing these parameters will improve results for some instances, but may cause an overall decrease in performance. Another reason that explains

the lower quality of *Howe*'s binarization in the VNIR range is the fact that this algorithm is designed to mark bleed-through pixels as background. But in our GT images, the bleed-through pixels are marked as foreground, because we are interested in them for material identification in the document.

As commented before, *DeepLab* offers a slightly worse overall performance than *Sauvola* in the VNIR range, but on the other hand it is less affected by the image dependency of optimal parameter settings, which in the end makes its results more reliable and general.

Looking at the examples shown in Fig. 6, the opposite trend from the one found in the VNIR range is observed for *Howe*, in the sense that it tends to find more foreground pixels than the other algorithms (see rows 1 and 2, for instance). For the third row, the central hole in the character is also narrower with *Howe*, and too wide for *Sauvola*. This is due to the effect of the k parameter setting commented above. Again, the global performance of *DeepLab* is more consistent, even if it is not able to find acceptable results for the fragments in the first two rows.

3.2.4 Three-channel vs single band images

This comparison is done to see in which cases it is worth performing the best band search. In Table 3, the average results for the two test subsets obtained using an RGB or three-band image transformed into grayscale by a standard transform ($Im_{grs} = 0.2989 \times R + 0.5870 \times G + 0.1140 \times B$) as input for the binarization are presented in the column labelled **three-channel**. The initial three-channel image was obtained as a pseudo-color image using the bands mentioned in Fig. 4 caption. Since there is no standard for the selection of these bands in the SWIR range, our choice of 1600, 1200, and 1000 nm is based on findings from previous experiments [70].

The results suggest that the difference between the optimal band and the three-channel image is negligible in the VNIR range for the classical algorithms, with a trend to a higher standard deviation in all metrics when the three-channel images are used as input. For *DeepLab*, the results are consistently worse for the three-channel images. For SAM and DINOv2, single channel images appear to be consistently better than three-channel images, as they better represent the natural-looking inputs that the underlying foundation models have been exposed to.

In the SWIR range, the three-channel image offers clearly worse results than the best band. This can very likely be explained by the fact that one of the bands used to form the pseudo-color image in the SWIR range is 1600 nm, which is over the fading limit for iron gall inks. Then, using several bands in the SWIR range can result in a comprehensive loss of information for certain materials, but not for others, as can be appreciated in Fig. 6, which shows examples of SWIR fragments that contain iron gall inks in rows 1 and 4. Moreover, the higher penetration depth of SWIR radiation tends to make the SWIR images slightly more blurred, and this blurriness is also dependent on wavelength, which makes it more noticeable for the three-channel than for the single-channel images.

3.2.5 Morphology bound

Among the adopted metrics for evaluation, PSNR is theoretically unbounded (upper limit is infinity), Pseudo F-Measure is upper-bounded at 100%, and DRD is lower-bounded at 0.

In practice, it is potentially interesting to evaluate the sensitivity of such metrics to minor inaccuracies that are unavoidably present in the ground truth annotations, due for example to ambiguous pixels lying on the threshold between ink and substrate. To this extent, binary opening and binary closing with a square 3×3 structural element have been applied to the ground truth, providing three fake binarizations, whose results are averaged and reported in Table 3 as "(Morphology bound)".

PSNR, despite being potentially infinite, in practice settles around 20 units, which remains distant from all reported configurations, with the exception of DeepLab on SWIR Easy Test coming closer to its bound. Pseudo F-Measure is, on the other hand, close to the actual predictions, bringing the practical upper bound from 100% down to roughly 97%. Finally, DRD's lower bound is only raised by 1% on average, so the general observations on the algorithms' performance remain unchanged.

3.2.6 Computational complexity comparison

The computational complexity of each binarization algorithm was evaluated based on the number of parameters, tuning method, and processing time on both VNIR and SWIR test sets. The results are reported in Table 4. Howe's method, while effective and depending on a low number of auto-tuned parameters, is the most time-consuming. The Sauvola method, with two empirically determined parameters, processes the images significantly faster. DeepLab, with 61,336,022 trainable parameters, demonstrates substantial computational efficiency on GPU, though it is relatively slower on CPU.

All experiments were conducted on a personal computer with the following hardware configuration for the traditional algorithms:

- Processor (CPU): Intel(R) Core(TM) i5-11400 F @ 2.60GHz (12 CPUs), 2.59 GHz
- Memory (RAM): 16 GB
- Storage: 512 GB NTFS SSD
- Operating System: Windows 10 Pro, v. 22H2, 64-bit

Table 4 Computational complexity comparison of binarization algorithms

Algorithm	Parameters	Tuning	VNIR (ms)	SWIR (ms)	Device
Howe	4/4	Auto-tuned	6691.428	6530.084	CPU
Sauvola	2/2	Empirical	370.804	351.227	CPU
DeepLabv3 (CPU)	61.3M/61.3M	Trainable	2005.100	2093.133	CPU
DeepLabv3 (GPU)	61.3M/61.3M	Trainable	97.761	101.289	GPU
SAM (CPU)	0/636.0M	None	3263.842	3183.056	CPU
SAM (GPU)	0/636.0M	None	1280.690	1286.715	GPU
DINOv2 (CPU)	770/22.2M	Trainable head	N/A	N/A	CPU
DINOv2 (GPU)	770/22.2M	Trainable head	248.709	245.710	GPU

Number of parameters is reported as trainable parameters over total ones

The DL-based method required a specialized hardware (GPU), run on a machine with the following configuration:

- Processor (CPU): Intel(R) Core(TM) i7-7700 CPU @ 3.60GHz (8 CPUs)
- Memory (RAM): 32 GB
- Graphics Card (GPU): NVIDIA Titan X, 12 GB
- Storage: 3 TB ext4 SSD
- Operating System: Ubuntu 22.04.3 LTS, 64-bit

4 Conclusions

This paper explores the use of hyperspectral imaging (HSI) for segmenting handwritten ink in historical documents, comparing traditional binarization methods (Sauvola and Howe) with the DeepLab deep learning model across VNIR and SWIR spectral ranges.

Results show that the "Easy Test" subset (with test samples from the same documents as training ones) yields better binarization outcomes than the "Hard Test" subset. Specifically, DeepLab demonstrates greater adaptability to complex samples, highlighting the value of the Hard Test subset as a benchmark for learning-based models. Between VNIR and SWIR ranges, Sauvola performs best in VNIR, while Howe is equally effective or better in SWIR. DeepLab's results, although generally reliable, are less consistent in SWIR due to conflicts in spectral bands, particularly where fading of iron gall inks above certain wavelengths impacts segmentation.

Comparing algorithms, Sauvola slightly outperforms DeepLab in VNIR, while DeepLab achieves superior results in SWIR, suggesting it offers broader adaptability in challenging imaging conditions. In comparing three-channel pseudo-color images to single optimal bands, results in VNIR are similar for traditional methods, but in SWIR, the three-channel approach significantly underperforms due to fading in the higher wavelengths, suggesting the need for a different band selection. For DeepLab, the transition to a three-channel input introduces a significant performance drop.

These findings highlight the potential of HSI for improving ink segmentation accuracy, especially where traditional imaging is insufficient. However, challenges remain, particularly in the SWIR range, where ink fading and data complexity affect performance.

Future research should focus on enhancing binarization methods by enabling Sauvola and Howe algorithms to process multiple bands simultaneously and produce results by consensus, potentially increasing segmentation reliability. For the DeepLab model, further developments could focus on enhancing the model's interpretability and performance.

Acknowledgements The authors have no additional acknowledgments to declare.

Author contributions **Marco Buzzelli**: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing - Original draft, Writing - Review and Editing, Visualization. **Francisco Moronta-Montero**: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing - Original draft, Writing - Review and Editing, Visualization. **Ramón Fernández-Gualda**: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing - Original draft, Writing - Review and Editing, Visualization. **Ana B. López-Baldomero**: Conceptualization, Methodology, Formal analysis, Investigation, Data curation, Writing - Original draft, Writing - Review and Editing, Visualization. **Juan Luis Nieves**: Resources, Conceptualization, Writing - Review and

Editing, Supervision. **Eva M. Valero**: Resources, Conceptualization, Methodology, Writing - Original draft, Writing - Review and Editing, Supervision, Project administration.

Funding This work was partially supported by grant PID2021-124446 NB-I00 funded by MICIU/AEI/10.13039/501100011033 and by ERDF, EU; grant PRE2022-101352 funded by MICIU/AEI/10.13039/501100011033 and "ESF+"; and grant FPU2020-05532 funded by Ministry of Universities (Spain). This work was partially supported by the MUR under the grant "Dipartimenti di Eccellenza 2023-2027" of the Department of Informatics, Systems and Communication of the University of Milano-Bicocca, Italy.

Data availability The data that support the findings of this study will be openly available following an embargo at the following URL: https://drive.google.com/drive/folders/1dqBEhxNRXuNMMWsXVW-uvBi2c8T9Z0BF?usp=sharing

Declarations

Competing interests The authors have no competing interests to declare that are relevant to the content of this article.

Availability of data and materials The data that support the findings of this study will be openly available following an embargo at the given URL.

References

1. Ciortan I, Deborah H, George S, Hardeberg JY (2015) Color and hyperspectral image segmentation for historical documents. In: 2015 Digital Heritage, IEEE, vol. 1, pp 199–206
2. Kavallieratou E, Stathis S (2006) Adaptive binarization of historical document images. In: 18th International conference on pattern recognition (ICPR'06), IEEE, vol. 3, pp 742–745
3. Sulaiman A, Omar K, Nasrudin MF (2019) Degraded historical document binarization: A review on issues, challenges, techniques, and future directions. J Imaging 5(4):48
4. Bartolini I, Moscato V, Pensa RG, Penta A, Picariello A, Sansone C, Sapino ML (2016) Recommending multimedia visiting paths in cultural heritage applications. Multimed Tools Appl 75:3813–3842
5. Salehani YE, Arabnejad E, Rahiche A, Bakhta A, Cheriet M (2020) Msdb-nmf: Multispectral document image binarization framework via non-negative matrix factorization approach. IEEE Trans Image Process 29:9099–9112
6. Magro N, Bonnici A, Cristina S (2021) Hyperspectral image segmentation for paint analysis. In: 2021 IEEE international conference on image processing (ICIP), IEEE, pp 1374–1378
7. López-Baldomero AB, Martínez-Domingo M, Valero EM, Fernández-Gualda R, López-Montes A, Blanc-García R, Espejo T (2023) Selection of optimal spectral metrics for classification of inks in historical documents using hyperspectral imaging data. In: Optics for Arts, Architecture, and Archaeology (O3A) IX, SPIE, vol. 12620, pp 99–111
8. Khan Z, Shafait F, Mian A (2013) Hyperspectral imaging for ink mismatch detection. In: 2013 12th International conference on document analysis and recognition, IEEE, pp 877–881
9. Shiel P, Rehbein M, Keating J et al (2009) The ghost in the manuscript: Hyperspectral text recovery and segmentation. Codicol Palaeography Digital Age, pp 159–174
10. Hedjam R, Cheriet M (2013) Historical document image restoration using multispectral imaging system. Pattern Recognit 46(8):2297–2312
11. Deborah H, George S, Hardeberg JY (2014) Pigment mapping of the scream (1893) based on hyperspectral imaging. In: Image and Signal Processing: 6th International Conference, ICISP 2014, Cherbourg, France, June 30–July 2, 2014. Proceedings 6, Springer, pp 247–256
12. Delaney JK, Dooley KA, Van Loon A, Vandivere A (2020) Mapping the pigment distribution of vermeer's girl with a pearl earring. Heritage Sci 8:1–16
13. Valero EM, Martínez-Domingo MA, López-Baldomero AB, López-Montes A, Abad-Muñoz D, Vílchez-Quero JL (2023) Unmixing and pigment identification using visible and short-wavelength infrared: Reflectance vs logarithm reflectance hyperspaces. J Cultural Heritage 64:290–300
14. Tensmeyer C, Martinez T (2020) Historical document image binarization: A review. SN Comput Sci 1(3):173

15. Baird HS (2004) Difficult and urgent open problems in document image analysis for libraries. In: First international workshop on document image analysis for libraries, 2004. Proceedings, IEEE, pp 25–32
16. Kavallieratou E, Antonopoulou H (2005) Cleaning and enhancing historical document images. In: Advanced Concepts for Intelligent Vision Systems: 7th International Conference, ACIVS 2005, Antwerp, Belgium, September 20-23, 2005. Proceedings 7, Springer, pp 681–688
17. Mello CA, Lins RD (2000) Image segmentation of historical documents. Visual2000, Mexico City, Mexico 30
18. Otsu N (1979) A threshold selection method from gray-level histograms. IEEE Trans Syst, Man, Cybernetics 9(1):62–66
19. Sauvola J, Pietikäinen M (2000) Adaptive document image binarization. Pattern Recognit 33(2):225–236
20. He J, Do Q, Downton AC, Kim J (2005) A comparison of binarization methods for historical archive documents. In: Eighth international conference on document analysis and recognition (ICDAR'05), IEEE, pp 538–542
21. Ran S, Lin L (2021) Painting element segmentation algorithm based on deep network. In: 2021 Photonics & electromagnetics research symposium (PIERS), IEEE, pp 2178–2182
22. Howe NR (2013) Document binarization with automatic parameter tuning. Int J Document Anal Recognit (ijdar) 16:247–258
23. Pratikakis I, Gatos B, Ntirogiannis K (2013) Icdar 2013 document image binarization contest (dibco 2013). In: 2013 12th International conference on document analysis and recognition, IEEE, pp 1471–1476
24. Ntirogiannis K, Gatos B, Pratikakis I (2014) Icfhr2014 competition on handwritten document image binarization (h-dibco 2014). In: 2014 14th International Conference on Frontiers in Handwriting Recognition, IEEE, pp 809–813
25. Bianco S, Buzzelli M, Schettini R (2019) A unifying representation for pixel-precise distance estimation. Multimed Tools Appl 78:13767–13786
26. Pratikakis I, Zagoris K, Barlas G, Gatos B (2017) Icdar2017 competition on document image binarization (dibco 2017). In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), IEEE, vol. 1, pp 1395–1403
27. Chen L-C, Zhu Y, Papandreou G, Schroff F, Adam H (2018) Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European conference on computer vision (ECCV), pp 801–818
28. Yu Y, Wang C, Fu Q, Kou R, Huang F, Yang B, Yang T, Gao M (2023) Techniques and challenges of image segmentation: A review. Electronics 12(5):1199
29. Oquab M, Darcet T, Moutakanni T, Vo H, Szafraniec M, Khalidov V, Fernandez P, Haziza D, Massa F, El-Nouby A et al (2023) Dinov2: Learning robust visual features without supervision. arXiv:2304.07193
30. Kirillov A, Mintun E, Ravi N, Mao H, Rolland C, Gustafson L, Xiao T, Whitehead S, Berg AC, Lo W-Y et al (2023) Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 4015–4026
31. Lettner M, Sablatnig R (2010) Higher order mrf for foreground-background separation in multi-spectral images of historical manuscripts. In: Proceedings of the 9th IAPR international workshop on document analysis systems, pp 317–324
32. Hedjam R, Cheriet M (2011) Novel data representation for text extraction from multispectral historical document images. In: 2011 International conference on document analysis and recognition, IEEE, pp 172–176
33. Mitianoudis N, Papamarkos N (2014) Multi-spectral document image binarization using image fusion and background subtraction techniques. In: 2014 IEEE International conference on image processing (ICIP), IEEE, pp 5172–5176
34. Diem M, Hollaus F, Sablatnig R (2016) Msio: Multispectral document image binarization. In: 2016 12th IAPR workshop on document analysis systems (DAS), IEEE, pp 84–89
35. Hedjam R, Nafchi HZ, Moghaddam RF, Kalacska M, Cheriet M (2015) Icdar 2015 contest on multispectral text extraction (ms-tex 2015). In: 2015 13th International conference on document analysis and recognition (ICDAR), IEEE, pp 1181–1185
36. Moghaddam RF, Cheriet M (2010) A multi-scale framework for adaptive binarization of degraded document images. Pattern Recognit 43(6):2186–2198
37. Hollaus F, Diem M, Sablatnig R (2015) Binarization of multispectral document images. In: Computer Analysis of Images and Patterns: 16th International Conference, CAIP 2015, Valletta, Malta, September 2-4, 2015, Proceedings, Part II 16, Springer, pp 109–120
38. Color Imaging Lab, University of Granada: (2025) Hyperdoc Database. https://colorimaginglab.ugr.es/pages/hyperdoc/database. [Online; accessed 17-Feb-2025]
39. Díaz Hidalgo RJ, Córdoba R, Nabais P, Silva V, Melo MJ, Pina F, Teixeira N, Freitas V (2018) New insights into iron-gall inks through the use of historically accurate reconstructions. Heritage Sci 6:1–15

40. Barkeshli M (2016) Historical persian recipes for paper dyes. Restaurator. Int J Preservation Library Archival Mater 37(1):49–89
41. Patronato de la Alhambra: (2019) Archivo - Patronato de la Alhambra y Generalife. https://www.alhambra-patronato.es/descubrir/investigacion/archivo. [Online; accessed 13-Nov-2024]
42. Real Chancillería de Granada: (2022) Alamas nazaríes. Los autógrafos de los sultanes (1454-1492). Consejería de Cultura y Patrimonio Histórico. Delegación Territorial en Granada. Archivo de la Real Chancillería de Granada. Granada 2022
43. Ministerio de Cultura - Gobierno de España: (2011) Censo-Guía de Archivos de España e Iberoamérica - ARCHIVO MUNICIPAL DE SELVA. http://censoarchivos.mcu.es/CensoGuia/archivodetail.htm?id=1511034
44. Resonon Inc.: Resonon PikaL. (2023) https://resonon.com/Pika-L. [Online; accessed 28-Nov-2023]
45. Resonon Inc.: (2023) Resonon PikaNIR. https://resonon.com/Pika-IR. [Online; accessed 28-Nov-2023]
46. Lee T-C, Kashyap RL, Chu C-N (1994) Building skeleton models via 3-d medial surface axis thinning algorithms. CVGIP: Graphical Models Image Process 56(6):462–478
47. Ntirogiannis K, Gatos B, Pratikakis I (2008) An objective evaluation methodology for document image binarization techniques. In: 2008 The Eighth IAPR international workshop on document analysis systems, IEEE, pp 217–224
48. Paris S, Hasinoff S, Kautz J (2011) Local laplacian filters: Edge-aware image processing with a laplacian pyramid. ACM Trans Graph 30(4)
49. GIMP.org: GIMP software. https://www.gimp.org/. [Online; accessed 28-Nov-2023] (2023)
50. Niblack W (1985) An Introduction to Digital Image Processing. Strandberg Publishing Company, DNK
51. Bradley D, Roth G (2007) Adaptive thresholding using the integral image. J Graphics Tools 12(2):13–21
52. Howe N (2023) Code implementation by Nicholas R. Howe. https://www.science.smith.edu/~nhowe/research/code/. [Online; accessed 01-Dec-2023]
53. Yu F, Koltun V (2015) Multi-scale context aggregation by dilated convolutions. arXiv:1511.07122
54. Pan SJ, Yang Q (2009) A survey on transfer learning. IEEE Trans Knowl Data Eng 22(10):1345–1359
55. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
56. Howard A, Sandler M, Chu G, Chen L-C, Chen B, Tan M, Wang W, Zhu Y, Pang R, Vasudevan V et al (2019) Searching for mobilenetv3. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 1314–1324
57. Medeiros L Language Segment-Anything. https://github.com/luca-medeiros/lang-segment-anything
58. Liu S, Zeng Z, Ren T, Li F, Zhang H, Yang J, Jiang Q, Li C, Yang J, Su H, Zhu J, Zhang L (2025) Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In: Leonardis A, Ricci E, Roth S, Russakovsky O, Sattler T, Varol G (eds) Computer Vision - ECCV 2024. Springer, Cham, pp 38–55
59. Fu S, Hamilton M, Brandt LE, Feldmann A, Zhang Z, Freeman WT (2024) Featup: A model-agnostic framework for features at any resolution. In: The Twelfth international conference on learning representations. https://openreview.net/forum?id=GkJiNn2QDF
60. Gonzalez RC, Woods RE (2018) Digital Image Processing, 4th edn., pp 184–201. Pearson,
61. Gonzalez RC, Woods RE (2018) Digital Image Processing, 4th edn. Pearson, Chap, p 10
62. Gonzalez RC, Woods RE, Eddins SL (2004) Digital Image Processing Using MATLAB. Pearson Education, Chap, p 11
63. Johnson DH (2006) Signal-to-noise ratio. Scholarpedia 1(12):2088. https://doi.org/10.4249/scholarpedia.2088. revision #126771
64. Fardo FA, Conforto VH, Oliveira FC, Rodrigues PS (2016) A formal evaluation of psnr as quality measurement parameter for image segmentation algorithms. arXiv:1605.07116
65. Sowmya B, Bhattacharya S (2005) Colour image segmentation using fuzzy clustering techniques. In: 2005 Annual IEEE India Conference-Indicon, IEEE, pp 41–45
66. Gatos B, Ntirogiannis K, Pratikakis I (2009) Icdar 2009 document image binarization contest (dibco 2009). In: 2009 10th International conference on document analysis and recognition, IEEE, pp 1375–1382
67. Ntirogiannis K, Gatos B, Pratikakis I (2012) Performance evaluation methodology for historical document image binarization. IEEE Trans Image Process 22(2):595–609
68. Lu H, Kot AC, Shi YQ (2004) Distance-reciprocal distortion measure for binary document images. IEEE Signal Process Lett 11(2):228–231
69. Sales F, Marconi C, Conte AM, Pulci O, Missori M (2024) Multispectral imaging and optical spectroscopy of two letters of st. The Collection of Chigi Palace in Ariccia, Italy, Advanced Technologies for Cultural Heritage Monitoring and Conservation, p 117

70. Moronta-Montero F, Gualda R, López-Baldomero AB, Buzzelli M, Martínez-Domingo M, Valero E (2024) Evaluation of binarization methods for hyperspectral samples of 16th and 17th century family trees. Archiving Conference 21:94–100

Authors and Affiliations

Marco Buzzelli[1] · Francisco Moronta-Montero[2] · Ramón Fernández-Gualda[2] · Ana Belén López-Baldomero[2] · Juan Luis Nieves[2] · Eva M. Valero[2]

✉ Marco Buzzelli
marco.buzzelli@unimib.it

Francisco Moronta-Montero
fmoronta@ugr.es

Ramón Fernández-Gualda
ramonz5@ugr.es

Ana Belén López-Baldomero
anabelenlb@ugr.es

Juan Luis Nieves
jnieves@ugr.es

Eva M. Valero
valerob@ugr.es

[1] Department of Informatics Systems and Communication, University of Milano-Bicocca, Viale Sarca, 336, 20126 Milan, Italy

[2] Department of Optics, Faculty of Sciences, University of Granada, Campus Fuentenueva sn, 18071 Granada, Spain