A Methodological Framework for Digital Enhancement and Text Recognition in Palimpsests

Katia Rasheva-Yordanova[0000-0001-5859-9114], Georgi P. Dimitrov[0000-0002-4785-0702], Inna Dimitrova[0000-0002-8480-7965]

University of Library Studies and Information Technologies, Sofia, Bulgaria
k.rasheva@unibit.bg, g.dimitrov@unibit.bg, i.dimitrova@unibit.bg

Abstract. This paper presents a methodology for enhancing palimpsest readability using image processing. Applying CLAHE, noise reduction, and adaptive thresholding enables the extraction of hidden text layers. The results show effective separation of scriptio inferior and superior, supporting further analysis through digital reconstruction and OCR.

Keywords: Palimpsest, Image Processing, Optical Character Recognition (OCR), Contrast Enhancement, Digital Preservation.

1 Introduction

A palimpsest is a multilayered artifact. It is a manuscript whose original text has been partially or completely erased by scraping or washing, after which the surface has been reused to record new content. It most often owes its appearance to the high cost of parchment and the difficulty of finding it during the Middle Ages, which necessitated the use of available parchments repeatedly (Perino et al., 2023; Teasdale et al., 2017). Palimpsests symbolize the physical and conceptual multi-layerness of cultural artifacts, as well as their interrelationships with historical context, social structures, and the multiple interpretations that can arise over time (Dimitrov, 2019; Namicheva-Todorovska et al., 2021). They are the subject of extensive research in modern scholarship, especially in the context of heritage and the preservation of cultural identity. They are a valuable source of information, as they often contain undisclosed data about historical texts that have been erased and rewritten.

The extraction and interpretation of palimpsests are often associated with various methodological and technical challenges (e.g. the specifics of the ink used, the parchment and the age of the artifact, the availability of technical capabilities for their study, etc.), often giving rise to additional limitations. The goal is to conduct the study with minimal risk to the integrity of the artifact.

Digital Presentation and Preservation of Cultural and Scientific Heritage, Vol. 15, 2025.
Sofia, Bulgaria: Institute of Mathematics and Informatics – BAS. ISSN: 1314-4006, eISSN: 2535-0366

Traditional methods of manuscript restoration are often perceived as risky for artifacts and palimpsests. Chemical treatments designed to reveal the original text can further damage manuscripts, limiting the effectiveness of conventional conservation techniques. Furthermore, the human eye is unable to capture all the details, especially in heavily faded texts, necessitating the use of modern analysis and visualization technologies (Rachman, 2017; Şahin, et al., 2017).

Due to these limitations, in recent years multispectral imaging has emerged as one of the leading methods for studying ancient manuscripts. This technique allows the discovery of hidden layers of text by capturing information at different wavelengths that remain invisible to the human eye (Rasheva-Yordanova et al., 2024).

New methods for digital text processing and recognition offer promising solutions that significantly facilitate the process of palimpsest restoration and analysis. This paper presents a structured methodological approach for digital text enhancement and recognition in palimpsests. The proposed approach integrates image preprocessing techniques, such as contrast enhancement (CLAHE), noise reduction, and adaptive thresholding, together with machine learning-based optical character recognition (OCR), to extract and restore the erased text layers. By systematically evaluating different processing methods, this study aims to establish an optimized process sequence for palimpsest analysis. The results demonstrate efficiency in separating scriptio inferior and scriptio superior, which allows for further analysis of the ancient manuscripts through digital reconstruction and OCR.

The article is organized into 3 relatively independent sections: Section one presents the general challenges in the study of palimpsests. Section two describes the methodological approach for digital processing and reading of palimpsests, including the stages of capture, pre-processing, separation of text layers, application of OCR and text reconstruction. Section three presents the results of the application of the methods and emphasizes the analysis of the effectiveness of the different techniques.

2 Challenges and Modern Approaches in the Study of Palimpsests

The study of palimpsests is a challenge due to multiple factors – the physical condition of the manuscripts, the multi-layered nature of the texts, the characteristics of the inks used, and the limitations of traditional restoration methods. They often suffer from damage caused by time, storage conditions, and the process of reusing the parchment itself. The ink can fade and the parchment can become worn or warped, making analysis significantly more difficult (Rachman, 2017; Şahin et al., 2017).

Further complications arise from erasure processes, such as scraping or washing, which were not always effective and over time can further complicate the recovery of the original text (Kasso et al., 2022). Furthermore, the presence of different text layers, which may contain similar handwriting or colors, makes the reading of palimpsests an extremely complex task (Kasso et al., 2022). The interaction between the "scriptio inferior" (the erased text) and the "scriptio superior" (the new writing) often leads to confusion in interpretation, further complicated by the inks used - iron, organic or soot

compounds (Rachman, 2017). The oxidation process, the aging of the material and the chemical interactions between the different inks further complicate the differentiation of the texts (Kasso et al., 2022).

These challenges are stimulating the development of new methods for studying hidden texts. Innovative approaches are being developed to improve the visibility of different text layers and preserve the integrity of original documents, overcoming the limitations of traditional conservation techniques (Şahin et al., 2017).

Over the past decade, palimpsests have become the subject of increased scholarly interest, especially among researchers using digital and non-invasive methods of analysis (Easton et al., 2018; Starynska et al., 2021). The increasing focus on digitization and new technologies in the humanities emphasizes the need for a multidisciplinary approach to their study (Kasso et al., 2022; Şahin et al., 2017). The main methods for analyzing palimpsests are presented in Table 1.

Table 1. Non-invasive methods for studying palimpsests.

Method	Description	Expected Result
Multispectral and hyperspectral imaging (Brenner & Miklas, 2019; Brenner et al., 2020; Fischer & Kakoulli, 2006)	Infrared (IR), ultraviolet (UV), and X-ray spectroscopy reveal hidden ink layers and parchment structure.	UV imaging enhances contrast between old and new text.
Image processing and artificial intelligence (AI) (Rasheva-Yordanova et al., 2024)	Techniques such as Contrast-Limited Adaptive Histogram Equalization (CLAHE) improve the contrast of hidden texts.	Machine learning and computer vision assist in the automatic extraction and classification of text layers. Additional processing in Adobe Illustrator allows for manual enhancement of the results.
Noise reduction and image cleaning algorithms (Rasheva-Yordanova et al., 2024)	Noise suppression using cv2.fastNlMeansDenoising eliminates artifacts from the images without losing important details.	
Digital reconstructions (Grusková & Gregorio, 2023)	Once the text is extracted, algorithms can be used to digitally reconstruct lost or fragmented parts of the document.	

The above non-invasive methods demonstrate significant potential for revealing and restoring hidden texts in palimpsests, as the combination of multispectral visualization, artificial intelligence, and image processing algorithms leads to more precise analysis and digital reconstruction of these valuable manuscripts.

3 Defining the Main Stages in Digitization and Reading of Palimpsests

The information from the previous section allows us to define six key stages in the digital processing of palimpsests: (1) manuscript preparation, (2) image capture, (3) pre-processing, (4) text recognition, (5) reading and reconstruction, and (6) digital archiving.

Fig. 1. Stages of digitization of palimpsests.

Each of these stages, along with its specifics, will be presented in the following table.

Table 2. Key stages in the digital processing of palimpsests.

Stage	Completed tasks
1. **Manuscript Preparation** (Physical Cleaning and Condition Analysis)	The palimpsest's condition is assessed, non-invasively cleaned, and the ink and parchment types are identified to select suitable visualization methods.
2.**Image capturing** (Using photographic techniques for digitization)	The master copy is captured in RGB. IR and UV imaging reveal hidden ink layers, while multispectral techniques distinguish text layers by spectral properties. 3D scanning is used if needed.
3.**Image Preprocessing (**Contrast Enhancement and Noise Removal)	First, the UV image is converted to grayscale to simplify processing. Then, cv2.fastNlMeansDenoising removes artifacts and isolates text contours. Finally, CLAHE enhances local contrast to reveal faint text.
4. **Recognition of text layers** (different text separation)	To separate scriptio superior and inferior, adaptive thresholding is applied: Gaussian for the upper text and median for the lower. Contrast normalization, FFT filtering, and morphological operations then enhance the readability of the palimpsest by suppressing the upper layer.
5. **Text Decoding and Reconstruction** (Application of OCR and Machine Learning)	An OCR algorithm, supported by machine learning, is used to reconstruct fragmented texts. Results undergo manual verification by philologists. Initial analysis links the Old Bulgarian layer to a 12th-century Menaion and the Greek text to 11th-century liturgical fragments, aiding studies of medieval manuscript reuse.
6. **Digital archiving and storage** (Saving data for future research)	Original and processed images are archived with metadata on digitization methods and techniques, ensuring public access via digital libraries and repositories.

These key stages form a comprehensive process of digital processing of palimpsests, each of which plays an essential role in the restoration, reading, and long-term preservation of historical texts.

4 Implementation of Digital Processing of Palimpsests

In this section of the article, the main results of the study and digitization of palimpsest layers in the Dragotin Apostolus (codex NBKM 880), a manuscript from the late 12th - early 13th century, kept at the National Library of the Republic of Bulgaria, will be presented. It contains text in Cyrillic and Greek, hidden under the Cyrillic upper layer. A fragment of the Greek text is identified as 11th century Octoechos, and the Cyrillic palimpsest folios (48 in number) contain an Old Bulgarian 12th c. menaion (Christova-Shomova, 2018). The codex is partially studied and digitalized in 2011 within the framework of the project "The Enigma of the Sinaitic Glagolitic Tradition", funded by the Austrian Science Fund. Multispectral images were captured of the 48 Cyrillic palimpsest folios and the text is read and published by Christova-Shomova (Christova-Shomova, 2018). Further studies of the Cyrillic and Greek palimpsest layers were not performed. In comparison to the Cyrillic palimpsest folios, the majority of the Greek scriptio inferior have poor legibility which require elaboration of multispectral capturing and image processing. The current study under the project "Interdisciplinary methods and tools for the study of manuscripts", funded by the Bulgarian Science Fund (BSF) aims at improving the capacity of multispectral capturing to provide better quality images using machine learning and AI modelling. The algorithm will be implemented for the digitalization and study of palimpsest manuscripts from Bulgarian repositories.

The results of the RGB capture of the main digital copy of the manuscript, as well as the work on revealing the hidden layers of ink by applying infrared (IR) and ultraviolet (UV) photography, can be seen in Fig. 2.

Fig. 2. Fragment of the Dragotin Menaion palimpsest, captured by multispectral imaging.

To distinguish the different text layers, multispectral and hyperspectral imaging was applied, which allows the detection of specific spectral characteristics of the ink and

parchment. This method identifies differences in the optical properties of materials invisible to the human eye, which aids the extraction of erased text elements.

A combination of contrast enhancement, noise reduction and binarization techniques was used for post-processing the images, aiming to extract the lower text layer (scriptio inferior). Contrast Limited Adaptive Histogram Equalization (CLAHE) was initially applied to highlight textual variations and local contrast differences. Non-Local Means Denoising was then used to reduce noise while preserving fine textual details. This multi-layered processing improves the clarity of the palimpsest and facilitates subsequent content analysis.

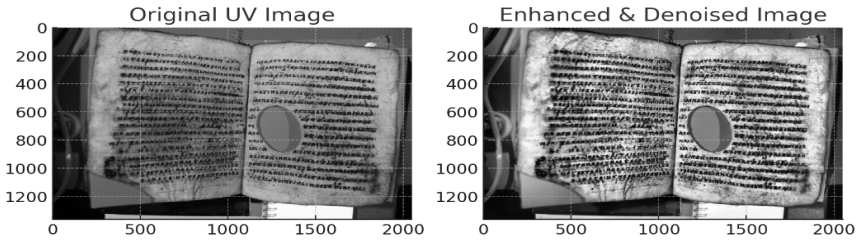

Fig. 3. Improving the legibility of a palimpsest through UV imaging and digital processing.

Adaptive Thresholding, which combines Gaussian and mean adaptive binarization, is then applied to effectively separate the different ink layers. This approach allows for local adaptation of segmentation thresholds, taking into account the illumination and textural characteristics of the manuscript.

In the final stage, contrast normalization and frequency component filtering were performed, aimed at suppressing the upper text (scriptio superior) and improving the legibility of the erased palimpsest (scriptio inferior). These methods reduce the visual dominance of the later inscription, optimizing the extraction of hidden textual information.

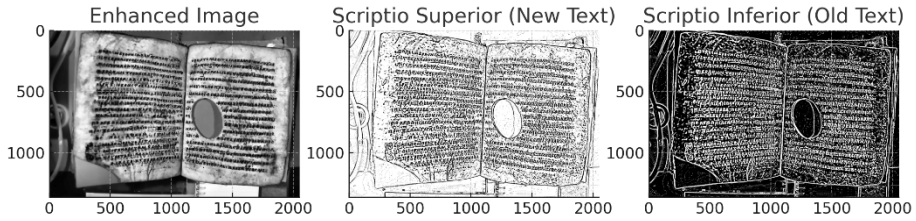

Fig. 4. Separation of text layers of a palimpsest through digital processing.

In the next stage, a Tesseract-based OCR algorithm, trained on 12th-century Old Bulgarian and Greek samples, was applied to the processed images (Fig. 5). The pipeline included contrast normalization and noise filtering to enhance legibility. The OCR module achieved an average recognition accuracy of 76.4% for Old Bulgarian and 84.1% for Greek texts, based on a manually verified test set.

```
custom_config = "--psm 6 --oem 3 -l bul+equ+osd"  # 'bul' for Bulgarian, 'equ' for equation recognition, 'osd' for orientation
recognized_text = pytesseract.image_to_string(processed, config=custom_config)
print("Recognized Old Slavonic Text:")
print(recognized_text)
with open("recognized_old_slavonic_text.txt", "w", encoding="utf-8") as f:
    f.write(recognized_text)
def correct_text(text):
    dictionary = {"#": "□", "@": "□", "0": "□", "9": "□", "□": "□", "□": "□"}
    for old_char, new_char in dictionary.items():
        text = text.replace(old_char, new_char)
    return text
corrected_text = correct_text(recognized_text)
with open("corrected_old_slavonic_text.txt", "w", encoding="utf-8") as f:
    f.write(corrected_text)
print("Corrected Text:")
print(corrected_text)
```

Fig. 5. Fragment of the OCR processing module for Old Bulgarian palimpsests.

The algorithm uses Tesseract OCR, trained on Cyrillic and Greek texts, and includes the ability to pre-process images to improve recognition. The module was configured with Tesseract 5.0 using LSTM-based recognition, and the training data included manually annotated samples from 12th-century Old Bulgarian manuscripts and Greek liturgical texts. The preprocessing pipeline combined contrast normalization and median filtering to enhance input quality before recognition. For model evaluation, accuracy metrics were collected on a test set of 500 manually verified lines, yielding an average character recognition rate of 76.4% for Old Bulgarian and 84.1% for Greek text layers.

Despite these measures, the OCR results still contain significant random errors and unrelated symbols, illustrating the challenges associated with palimpsest processing. (Fig. 6).

Fig. 6. Recognized Old Bulgarian characters after applying the OCR algorithm.

We acknowledge that the presence of errors may be due to several factors: (1) Noise in the image – despite pre-processing, there may be traces of the above text or random occurrences from the aging of the material; (2) Suboptimal Tesseract settings for Old Bulgarian – additional training of the OCR model with specific Old Bulgarian manuscripts may be necessary; (3) Deformation of the letters in the palimpsest – ink fading and the structure of the material may make recognition difficult.

Based on the experimental results, the advantages and disadvantages of the applied methods were analyzed, with the main goal being to identify an optimal strategy for

recovering the hidden texts. Through a comparative analysis of the different palimpsest processing techniques, their effectiveness in improving the readability and extracting the lower text layer (scriptio inferior) is assessed.

The results obtained allow for a conclusion on the most appropriate combination of methods, which provides an opportunity for future improvement of the process through machine learning and advanced text reconstruction algorithms. A readability index was used to compare methods. The CLAHE + FFT approach achieved 47.3 points, significantly higher than UV imaging alone (21.7), confirming the benefit of combining techniques.

Potential improvements include adaptive image processing models, automated artifact correction, and integration of specialized OCR systems trained on manuscripts in the respective language (Old Bulgarian, Greek etc.).

The following table presents the effectiveness of each method, as well as the degree of improvement in the readability of the lower layer of the manuscript.

Table 3. Influence of different methods on the legibility of the palimpsest.

Method	Description	Readability Before/After
RGB Imaging	Standard RGB image	Illegible → No improvement
UV Imaging	UV light for inks	Partially visible → Enhanced ink
CLAHE	Contrast enhancement	Faint text → Clearer outline
Noise Reduction	Noise reduction	Noisy text → Cleaner details
Binarization	Separation of text layers	Mixed layers → Separated texts
FFT Filtering	Isolation of the old text	Ink blending → Clearer old text
OCR	Automatic text recognition	Unrecognizable text → Partially recognized

From this, the advantages and disadvantages of each of the methods described above can be deduced (Table 4).

Table 4. Advantages and disadvantages of the applied methods.

Method	Advantages	Disadvantages
RGB Imaging	Preserves the original	Lack of contrast
UV Imaging	Highlights invisible inks	Does not always reveal details
CLAHE	Increases contrast	May lose some information
Noise Reduction	Preserves details	Requires tuning
Binarization	Separates layers	May introduce artifacts
FFT Filtering	Isolates underlying text	Requires powerful processing
OCR	Automates text recognition	Errors with specific characters

It can be concluded that the combination of CLAHE, Non-Local Means Denoising and Adaptive Thresholding gives the best results for improving the readability of the palimpsest. The additional use of FFT filtration helps to isolate the old text, while OCR algorithms can automate the reading, although with some challenges in the case of Old Bulgarian manuscripts.

5 Conclusions

The article presents a comprehensive approach for digital processing and reading of palimpsests, combining methods for capturing, image preprocessing, recognition of text layers and optical character recognition (OCR). The application of techniques such as contrast histogram equalization (CLAHE), noise suppression (Non-Local Means Denoising) and adaptive binarization significantly improves the readability of the hidden text, allowing the separation of scriptio inferior from scriptio superior. The results show that the use of FFT filtration helps to isolate the lower text layer, and the application of OCR algorithms provides the opportunity for automated reading of the content, although with challenges in the case of Old Bulgarian manuscripts.

The analysis of the results highlights the need to combine several digital techniques to achieve optimal results. The best results were obtained by an integrated approach involving UV imaging, contrast enhancement and text segmentation algorithms. However, the challenges in recognizing Old Bulgarian texts indicate the need for additional training of OCR models, as well as the use of Deep Learning for fragmented text reconstruction.

Future research is planned to make increased use of multispectral analyses and automated machine learning to improve the reading and reconstruction of palimpsests. The application of more precise models for handwriting analysis and the development of specific tools for palimpsest processing would contribute to an even more detailed and efficient study of historical documents.

Acknowledgements.

The current study is performed within the framework of the project "Interdisciplinary methods and tools for the study of manuscript monuments", КП-06-Н60/9 (2021), funded by the Bulgarian Science Fund (BSF).

References

Brenner, S., & Miklas, H. (2019). Multispectral imaging of degraded manuscripts in the Ivan Vazov National Library in Plovdiv. *Godišnik na Narodna biblioteka „Ivan Vazov" (2014–2018)* [Yearbook of the Ivan Vazov National Library (2014–2018)], 32–48.

Brenner, S., Sablatnig, R., Cappa, F., Vetter, W., Frühmann, B., Schreiner, M., & Miklas, H. (2020). Virtual conservation and restoration via multispectral imaging and

spectroscopy. In M. V. Korogodina (Ed.), *Issledovanie i restavraciâ rukopisej: Materialy konferencii - 2019* [Manuscript research and restoration: Conference materials - 2019, St. Petersburg, September 17–18, 2019] (pp. 15–26). SPb.: BAN. https://spbiiran.ru/issledovanie-i-restavracziya-rukopisej-sbornik-materialov-nauchnoj-konferenczii-2019-goda/

Christova-Shomova, I. (2018). *Dragotin minej. Bulgarski rukopis ot nachaloto na XII vek* [Dragotin Menaion. A Bulgarian manuscript from the beginning of the 12th century]. Univ. izd. "Sv. Kliment Ohridski."

Dimitrov, V. (2019). And everybody leaves something so the next could ruin it. Along Via Diagonalis through Istria and Friuli. *Sledva: Journal for University Culture*, *39*, 67–81. https://doi.org/10.33919/sledva.19.39.9

Easton, R. L. Jr., Knox, K. T., Christens-Barry, W. A., & Boydston, K. (2018). Spectral imaging methods applied to the Syriac Galen Palimpsest. *Manuscript Studies: A Journal of the Schoenberg Institute for Manuscript Studies, 3*(1), 69–82.

Fischer, C., & Kakoulli, I. (2006). Multispectral and hyperspectral imaging technologies in conservation: Current research and potential applications. *Studies in Conservation, 51*(sup1), 3–16. https://doi.org/10.1179/sic.2006.51.supplement-1.3

Grusková, J., & Gregorio, G. D. (2023). Neue paläographische Einblicke in einige palimpsestierte Handschriften aus den griechischen Beständen der Österreichischen Nationalbibliothek [New palaeographic insights into palimpsested manuscripts from the Greek holdings of the Austrian National Library]. *Veröffentlichungen zur Byzanzforschung*, *2023*, 317–341. https://doi.org/10.1553/978oeaw91575s317

Grotans, A. A., Hendrix, J., & Kaczynski, B. M. (2009). Understanding medieval manuscripts: St. Gall's virtual library. *History Compass, 7*(3), 955–980. https://doi.org/10.1111/j.1478-0542.2009.00603.x

Kasso, T., Kytökari, M., Oinonen, M., Mizohata, K., Tahkokallio, J., & Heikkilä, T. (2022). A glance to the Fragmenta Membranea manuscript collection through FTIR and radiocarbon analyses. *Radiocarbon, 65*(1), 155–171. https://doi.org/10.1017/rdc.2022.81

Namicheva-Todorovska, E., & Namichev, P. (2021). Cultural sustainability and architectural heritage. *Palimpsest, 6*(11), 227–239. https://doi.org/10.46763/palim21116227n

Perino, M., Ginolfi, M., Felici, A. C., & Rosellini, M. (2023). A deep learning experiment for semantic segmentation of overlapping characters in palimpsests. In *Proceedings of the 2022 IMEKO TC4 International Conference on Metrology for Archaeology and Cultural Heritage*. https://doi.org/10.21014/10.21014/tc4-arc-2023.153

Rachman, Y. B. (2017). The use of traditional conservation methods in the preservation of ancient manuscripts: A case study from Indonesia. *Preservation, Digital Technology & Culture, 46*(3), 109–115. https://doi.org/10.1515/pdtc-2017-0006

Rasheva-Yordanova, K., Dimitrov, P. G., Tsvetkova, P. T., Bankovska, M., & Petrov, P. S. (2024). Improving palimpsest readability via image preprocessing: An investigation into adjustment techniques. *Digital Presentation and Preservation of Cultural and Scientific Heritage, 14*, 107–116. https://doi.org/10.55630/dipp.2024.14.9

Şahin, C. D., Coşkun, T., Arsan, Z. D., & Akkurt, G. G. (2017). Investigation of indoor microclimate of historic libraries for preventive conservation of manuscripts: Case study: Tire Necip Paşa Library, İzmir-Turkey. *Sustainable Cities and Society, 30*, 66–78. https://doi.org/10.1016/j.scs.2016.11.002

Starynska, A., Messinger, D., & Kong, Y. (2021). Revealing a history: Palimpsest text separation with generative networks. *International Journal on Document Analysis and Recognition, 24*(3), 181–195. https://doi.org/10.1007/s10032-021-00379-z

Teasdale, M. D., Fiddyment, S., Vnouček, J., Mattiangeli, V., Speller, C., Binois, A., & Collins, M. J. (2017). The York Gospels: A 1000-year biological palimpsest. *Royal Society Open Science, 4*(10), 170988. https://doi.org/10.1098/rsos.170988

Received: March 25, 2025
Reviewed: May 02, 2025
Finally Accepted: June 10, 2025