Journal of Instrumentation

PAPER • **OPEN ACCESS**

Assessing readability of the text in ancient paper fragments by a photometric statistical analysis

To cite this article: Martina Franchi *et al* 2024 *JINST* **19** C05022

Published by IOP Publishing for Sissa Medialab

Received: *December 31, 2023*
Revised: *March 21, 2024*
Accepted: *April 12, 2024*
Published: *May 13, 2024*

International Workshop on Imaging
Varenna (Lake Como), Italy
26–29 September 2023

Assessing readability of the text in ancient paper fragments by a photometric statistical analysis

Martina Franchi[a,b,*] **Stefania Colonnese**[c] **Alessia Cedola**,[b] **Lia Barelli**[d] **and Simona Morretta**[e]

[a] *SBAI Department, Sapienza University of Rome, P.le Aldo Moro 5, 00185 Rome, Italy*
[b] *Institute of Nanotechnology — CNR, Rome, Italy*
[c] *DIET Department, Sapienza University of Rome, Via Eudossiana 18, 00184 Rome, Italy*
[d] *DSDRA Department, Sapienza University of Rome, P.zza Borghese 9, 00186 Rome, Italy*
[e] *Soprintendenza Speciale Archeologia Belle Arti e Paesaggio di Roma (SSABAP), P.zza dei Cinquecento 67, 00185 Rome, Italy*

E-mail: martina.franchi@uniroma1.it

Abstract: Ancient documents are important historical sources that are often found in a fragmented condition due to their conservation status. In this study, we examined fragments of paper found in 1996 during excavation of the Santi Quattro Coronati complex, in Rome. The archaeological site where the fragments were found is situated on the first floor of the tower within the complex. This location was used as a disposal pit approximately between the 15th and 16th centuries. The fragments exhibit text discoloration, hindering automatic recognition and human readability. To reveal the faded text, the fragments have been digitalized, converted into a perceptually uniform color space and the contrast has been enhanced. The photometric characteristics of the input and enhanced images have been statistically characterized, and the contrast enhancement assessed by a state-of-the-art metric. The statistical analysis of the text colour coordinates was carried out to develop supervised and unsupervised image segmentation, isolating the text.

The results of the method show that it effectively identifies text regions within images, improving readability, even for faded text. It can be integrated into deep learning-based character recognition systems, facilitating the automatic analysis of historical handwritten documents.

Keywords: Analysis and statistical methods; Image filtering; Image processing; Data analysis

*Corresponding author.

https://doi.org/10.1088/1748-0221/19/05/C05022

Contents

1 Introduction 1

2 Materials and methods 1

3 Results 2

4 Conclusion 4

1 Introduction

This work investigates paper fragments unearthed in 1996 during the excavation of the tower of the Santi Quattro Coronati complex in Rome [1]. Previous studies of materials collected during the excavation led to significant results in the reconstruction of Renaissance diets in Latium [2]. The fragments were found in a staircase, used as a disposal pit (garbage dump) between the 15th and the 16th centuries [2]. These fragments contain crucial information on the postal correspondence of the Roman Catholic cardinals residing in the complex and beyond, including interactions with the newly discovered continent, America. Due to their poor state of preservation, the manuscript documents have been reduced to fragments, affected by discolouration of the text, compromising their readability and the possibility of carrying out historical analyses.

This study addresses the challenge of enhancing the readability of these manuscripts through non-invasive techniques to allow automatic character recognition. Preprocessing through contrast enhancement and color segmentation is crucial to enhance legibility and ensure reliable recognition of faded text characters. We processed the fragment images acquired by a digital camera, focusing on contrast enhancement to improve readability and on statistically grounded color space segmentation to isolate the text. We evaluated the enhancement qualitatively and quantitatively.

To study manuscripts affected by ink loss or brown spotting, multispectral imaging techniques such UV fluorescence and VIS-NIR imaging can be employed [3, 4]. Text enhancement and segmentation may leverage deep learning [5, 6]. We present a method for handling images from digital libraries, which typically consist of RGB images captured by portable digital cameras or flatbed scanners [7]. In our work, we leverage state-of-the-art, ideally unsupervised, image enhancement and segmentation methods suitable for application in a training free environment. The results led to an improvement in text readability, laying the groundwork for future enhancements in the digitization of written content.

2 Materials and methods

Materials. The samples consist of paper fragments with different sizes — ranging from 2 cm to 10 cm — shapes and yellowish color tones. They have holes and brown spots on the surface and, for some fragments, text is present on both sides (recto and verso) of the leaves. For this study, we selected fragments with more lines of written text. The text characters are brown in colour and are approximately 0.5 centimeters in height, see figure 1(a). The images were acquired by a Canon EOS 6D camera with a 24–105 mm lens at a range of 80 to 85 mm focal length. The aperture was set to f/13, and various exposure times ranging from 1 to 5 seconds were used. The ISO was set to 50. The camera was fixed on a tripod and placed at a distance of 50 cm from the sample.

2024 JINST 19 C05022

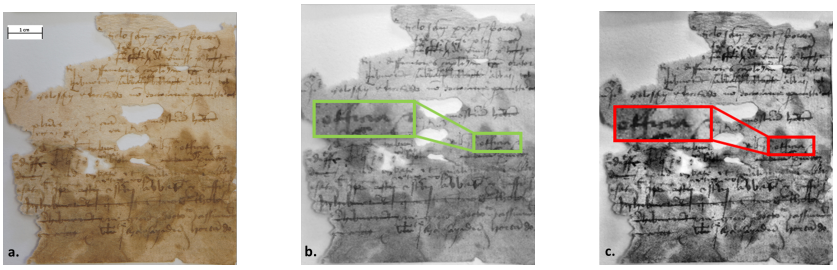

Figure 1. a) RGB fragment image, (b) corresponding image in L^* channel, (c) L^* channel image enhanced by CLAHE.

Methods. We processed the fragments' images by a three-tier algorithm, performing: i) color space conversion, ii) neighborhood-driven image enhancement, and iii) supervised and unsupervised segmentation of the enhanced image for text detection purposes.

We firstly convert the color representation from the original RGB space to the CIE $L^*a^*b^*$ space. CIE $L^*a^*b^*$ is a perceptually uniform color space, where the Euclidean distance between color points corresponds to their perceived difference [8]. After a preliminary assessment of different contrast enhancement techniques for text readability improvement, we applied Contrast Limited Adaptive Histogram Equalization (CLAHE) [9, 10] on the L^* component which represents the lightness, as illustrated by figure 1(b-c). The input image is divided into non-overlapping local neighborhoods of Window Size (WS). The CLAHE is then performed on the L^* channel, i.e. the histogram of each contextual area for L^* undergoes clipping at a Clip Limit (CL), and the remaining pixels are redistributed across the entire gray-level range [11]. The WS and CL parameters control the degree of contrast enhancement. Since CLAHE may emphasize undesired details in the background regions, the WS and CL parameters demand expert calibration to prevent artifacts. After visually analyzing the changes in the L^* distribution for the input and enhanced images of each fragment, we assessed the enhancement performance in terms of brightness preservation applying the Absolute Mean Brightness Error (AMBE) metric, formerly introduced on the luminance [10], directly to the perceptual lightness L^*. Given the values $\mathbf{L}^*_{\text{or}}$ and $\mathbf{L}^*_{\text{en}}$ of the input and the enhanced image, the metric is defined as the absolute difference $\left| E\{\mathbf{L}^*_{\text{or}}\} - E\{\mathbf{L}^*_{\text{en}}\} \right|$ where $E\{\cdot\}$ denotes the spatial average over the image.

As a third step, we segmented both the input and enhanced images to isolate the text in both the RGB and YCbCr color spaces. We performed supervised color segmentation, determining the typical pattern of the text areas in the color space and identifying the ranges of the RGB and YCbCr text coordinates.

Finally, we automated the segmentation by characterizing the text color in both RGB and YCbCr spaces. We selected 30 pixels from the text in each input fragment image, calculated their mean and standard deviation, and used this information to drive the color-based unsupervised segmentation.

3 Results

We transformed the images from RGB to CIELab space and applied CLAHE to the L^* channel. Then, we evaluated the performance of CLAHE by calculating the AMBE on a subset of 9 out of $N = 18$ images, showcasing all the features present in the set. AMBE values were $\{1.5, 0.4, 0.9, 2.9, 2.8, 4.7, 2.3, 1.4, 0.4\}$, relatively small w.r.t. the peak L^* values ($\approx 10^2$), indicating consistent preservation of lightness across images, with some variations due to individual content differences.

2024 JINST 19 C05022

Figure 2(a) shows the L* histograms of a sample before (left) and after contrast enhancement (right). After the enhancement the histogram shows an increase in the number of gray levels, suggesting improved contrast. The L* histograms bins corresponding to ink areas are highlighted by arrows. The ink-like zone was extended, improving clarity in text. Figure 2(b) shows a scatter plot of the L* channel before and after enhancement. We recognize the color stretching of the enhancement by the spread widths of input lightness value. For identity transformation, the scatter plot would reduce to a unitary slope straight line. Indeed, the smallest lightness are slightly scaled. Mid-range lightness values are spread over different enhanced values, depending on their neighborhood.

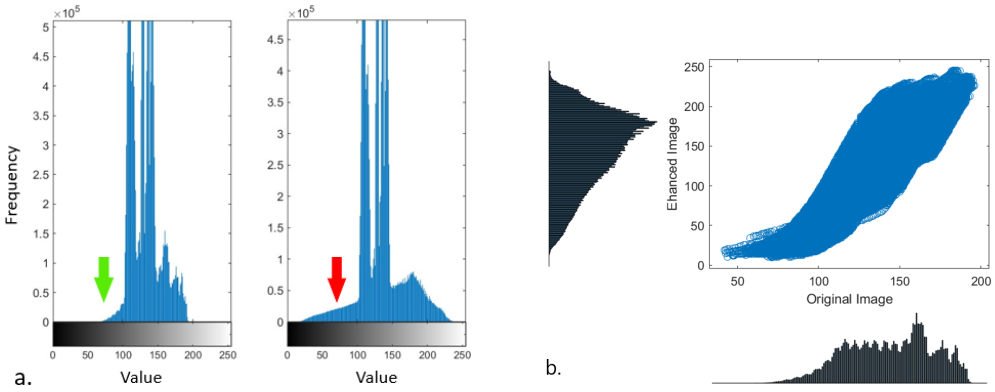

Figure 2. a) Histograms of the L* channel of the input (left) and enhanced (right) image (arrows identify the bins of the ink color). b) Scatter plot of the enhanced L* values versus the input ones.

After converting back the enhanced images from the CIE L*a*b* space to RGB, we applied color-based segmentation. We have evaluated the separability of the text from the background using only the color information, by supervised segmentation techniques. Specifically we calculated the histograms of the image color components and established the correspondence between the spatial text regions and the histogram bins; the operation was iterated across various color spaces.

We selected the RGB and YCbCr color spaces for conducting supervised segmentation. Figure 3 shows, for the input (left) and its segmented version (right), the histograms for individual color components and the color values of all pixels in a 3-D color space plot, as represented by the graphical interface provided by Matlab®. Real-time, supervised highlighting of the image regions linked to specific histogram components has shown that the text could be successfully isolated in the color space. Segmentation results in the RGB space revealed notable shifts in the 3D pixel distribution before and after the enhancement. Figure 4 displays the segmentation results for the RGB space, comparing the segmentation of the input image and of its enhanced version. By visual inspection we recognize that the enhancement improves text pixel detection, leading to a distinguished color space region for text representation.

After isolating text, we examined pixel colors. For each image, we chose $M = 30$ distinct text pixels, totaling $N = 540$ observations per color space. We computed the mean color coordinates and standard deviations $(\mu_R, \sigma_R) = (82, 35)$, $(\mu_G, \sigma_G) = (55, 33)$, $(\mu_B, \sigma_B) = (29, 25)$, and $(\mu_Y, \sigma_Y) = (51, 26)$, $(\mu_{Cb}, \sigma_{Cb}) = (-15, 5)$, $(\mu_{Cr}, \sigma_{Cr}) = (15, 2)$ in the RGB and YCbCr color spaces, respectively. Regions featuring faded, stained or dense ink show high ink color coordinates variability (e.g. $\sigma_Y \approx 50\%\mu_Y$), justifying local enhancement for easier automatic segmentation.

2024 JINST 19 C05022

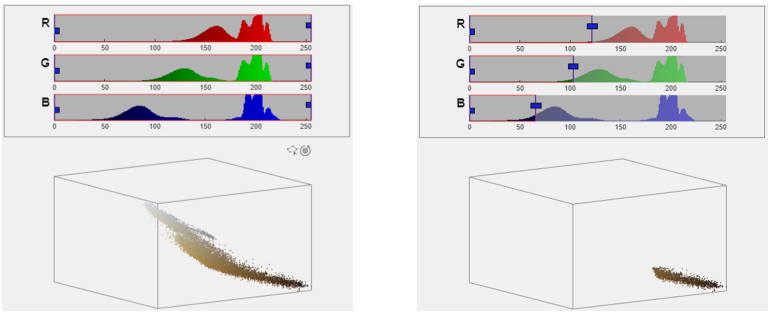

Figure 3. RGB histograms and 3-D color space for the input (left), and segmented image (right).

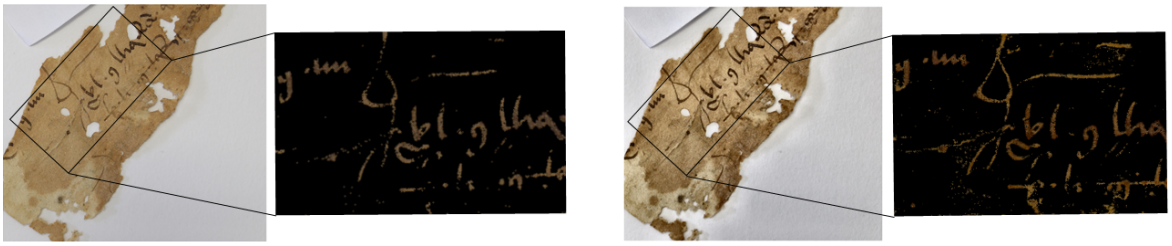

Figure 4. Supervised segmentation: (a) input image and (b) segmented image; (c) enhanced image and (d) segmented image.

These parameters enabled unsupervised text segmentation, facing color similarities in background, faded text, and stains. Figure 5 shows the results of unsupervised segmentation in the RGB color space; similar results in the YCbCr space are omitted for brevity. The proposed procedure made the automatic color segmentation process fast and applicable to fragments in varying states of preservation.

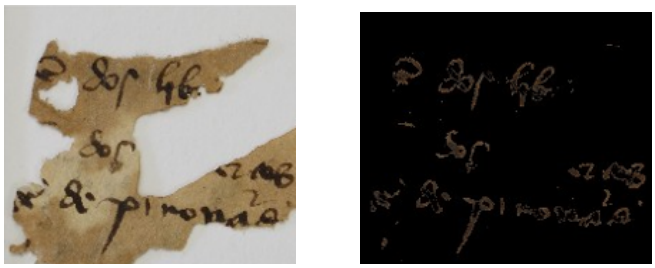

Figure 5. Unsupervised segmentation of RGB image: input image (left), and segmented image (right).

4 Conclusion

We present a method for enhancing the readability of digitized versions of stained and faded manuscript fragments from the Santi Quattro Coronati archaeological site in Rome. We firstly applied perceptually driven contrast enhancement to amplify faded text details, quantitatively assessing the enhancement performance across different text regions. Secondly, we developed both a supervised and unsupervised color-based segmentation task, isolating the written content and recovering faded characters. The

2024 JINST 19 C05022

combination of perceptually uniform color space enhancement and unsupervised segmentation, informed by a statistical analysis of text color coordinates, demonstrated the effectiveness of the approach. Future work will extend the technique to multispectral images and integrate it by deep learning driven by an objective text enhancement and detection performance metric.

Acknowledgments

We acknowledge SSABAP for granting access and permission to study the samples, and thank Dr. C. Moricca for her valuable support.

References

[1] L. Barelli, *The monumental complex of Santi Quattro Coronati in Rome*, translation by C. McDowall, Viella (2009).

[2] C. Moricca et al., *Early arrival of new world species enriching the biological assemblage of the Santi Quattro Coronati complex (Rome, Italy)*, *Interdiscip. Archaeol.* **9** (2018) 83.

[3] L. Pronti, M. Perino, M. Cursi, M. Santarelli, A. Felici and M. Bracciale, *Characterization and digital restauration of xiv-xv centuries written parchments by means of nondestructive techniques: three case studies*, *J. Spectro.* **2018** (2018) 2081548.

[4] C. Jones, C. Duffy, A. Gibson and M. Terras, *Understanding multispectral imaging of cultural heritage: Determining best practice in msi analysis of historical artefacts*, *J. Cult. Herit.* **45** (2020) 339.

[5] S. Khamekhem Jemni, M.A. Souibgui, Y. Kessentini and A. Fornés, *Enhance to read better: A Multi-Task Adversarial Network for Handwritten Document Image Enhancement*, *Pattern Recognit.* **123** (2022) 108370 [arXiv:2105.12710].

[6] R. Najam and S. Faizullah, *Analysis of recent deep learning techniques for arabic handwritten-text ocr and post-ocr correction*, *Appl. Sci.* **13** (2023) 7568.

[7] E.R. Leggett, *Digitization and digital archiving: A practical guide for librarians*, vol. 71, Rowman & Littlefield Publishers (2020).

[8] B. Preim, C. Botha, B. Preim and C. Botha, *An introduction to medical visualization in clinical practice*, in *Visual Computing for Medicine (Second Edition)*, Morgan Kaufmann (2014), p. 69–110 [DOI:10.1016/B978-0-12-415873-3.00003-1].

[9] K. Zuiderveld, *Contrast limited adaptive histogram equalization*, in *Graphics gems*, Academic Press (1994), p. 474–485.

[10] M.A. Qureshi, A. Beghdadi and M. Deriche, *Towards the design of a consistent image contrast enhancement evaluation measure*, *Signal Process. Image Commun.* **58** (2017) 212.

[11] S.H. Majeed and N.A.M. Isa, *Adaptive Entropy Index Histogram Equalization for Poor Contrast Images*, *IEEE Access* **9** (2021) 6402.

2024 JINST 19 C05022