Color and Hyperspectral Image Segmentation for Historical Documents

Irina Ciortan*, Hilda Deborah*†, Sony George*, and Jon Y. Hardeberg*,

*The Norwegian Colour and Visual Computing Laboratory,
Department of Computer Science and Media Technology, Gjøvik University College
P.O.Box 191, N-2802 Gjøvik, Norway

†Laboratory XLIM-SIC UMR CNRS 7252, University of Poitiers
Bât. SP2MI, Téléport 2, 11 Bd Marie et Pierre Curie, BP 30179
86962 Futuroscope Chasseneuil Cedex, France

***Abstract*—Several historical documents from the collection of the National Library of Oslo were acquired using a hyperspectral scanner. While each of the documents has its specific characteristics, requiring different image preprocessing steps, the common goal for all documents is to increase their legibility. The aim of this study is to show the advantage of hyperspectral imaging compared to traditional color imaging, in particular for the task of ink separation using distance-based classification method.**

***Index Terms*—hyperspectral imaging, cultural heritage, historical documents, classification, segmentation, distance function**

I. Introduction

Historical documents represent the legacy of our past and predecessors. Manuscripts act as testimony of the culture, a recording of historical happenings, they can comprise a collection of first drafts or sketches that were later turned into great works of art or they might hold numerous conversation between world important personalities and thus, they represent over all a valuable source of potential new discoveries. As the Latin proverb goes, "verba volant, scripta manent", spoken words fly away, while written words remain carved in history and stay as proof for further generations and investigations. It might be that many of the answers to secrets and unsolved dilemmas of the past lie in the tomes of manuscripts and papers stored in libraries and archives or simply lost in private or anonymous collections. This only, to begin with, is a strong motivation by itself to regard with increased care the preservation of manuscripts as cultural heritage items and to start studying and analyzing the information content of the historical documents in a conservation-friendly way.

An important sub-field of study in document investigation is ink examination, as it provides relevant clues on the authenticity, backdating and age of the manuscript [1]. As well as the substrate of the document, the ink is prone to various degradation processes, such as ink fading or invisibility, ink-corrosion, ink-bleeding, ink-overlapping and ink-mixing. According to the level of intrusion, there are two main methods of ink analysis [1]: destructive and non-destructive. The first category refers to chemical analysis, that implies contact, sample withdrawal and chemical experiments on the document. The second category includes spectroscopy and digital imaging. In this article, we focus on the hyperspectral imaging techniques as a state-of-the-art, useful and non-destructive tool for ink analysis in handwritten documents [2].

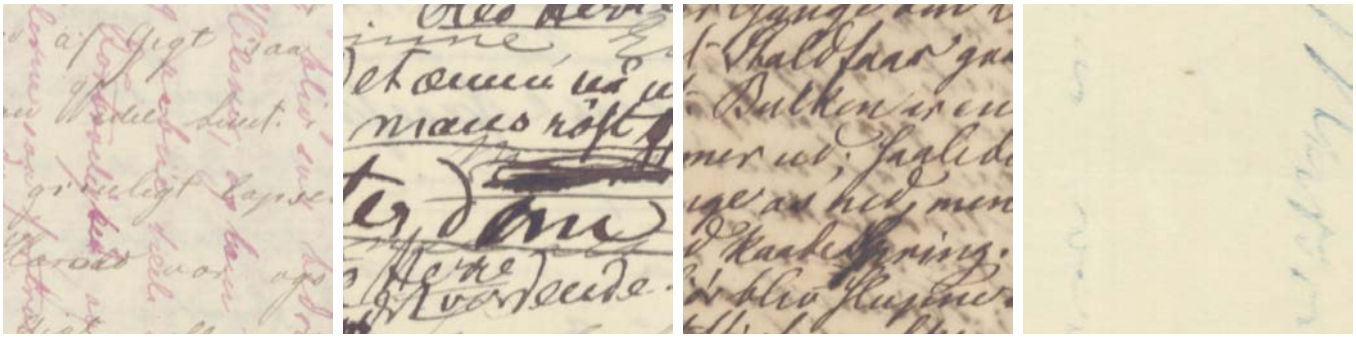
Spectral imaging has been used in various applications, including remote-sensing [3], [4], cultural heritage e.g. pigment mapping, material identification, discoveries of underwritings [5], medical skin analysis [6], food quality [7], etc. Depending on the number of bands from the electromagnetic spectrum that can be acquired, the spectral imaging systems are multispectral or hyperspectral [8]. The hyperspectral imaging(HSI) systems output a high spectral resolution along the spatial domain, so that every pixel in the resulting image is a continuous spectra. Meanwhile, multispectral images comprises less spectral bands and estimation methods are needed for spectrum reconstruction at pixel level. As detailed in the following section, spectral imaging is an emerging instrument for document and ink analysis.

In this paper, we show how we conducted a case study for historical manuscripts belonging to the collection of the National Library of Oslo with hyperspectral image acquisition. Then, we propose to improve the legibility of the handwritten documents, with a distance-based ink separation approach and we compare the hyperspectral results with the ones obtained by conventional RGB color methods.

II. State Of The Art

Spectral imaging is a promising approach to address specific degradation aspects present in historical handwritten documents and successful results have been previously obtained [9], [10], [11], [1], [12], [13]. Processing multispectral images, the faded writing from historical manuscripts was recovered by detecting the invisible ink with a constrained energy maximization filter [11]. Hidden text was recovered with Principal Component Analysis and Independent Component Analysis, operations allowed by hyperspectral images [13].

978-1-5090-0048-7/15/$31.00 ©2015 IEEE



(a) M1 (b) M2 (c) M4 (d) M5

Fig. 1. Historical documents dataset where each document has its particular problem. Nevertheless one common processing goal is set for this particular dataset, i.e. ink segmentation. The color images shown above were generated using the CMF method, see Sec. III-C.

In [14], eighteen aqueous stabilization treatments proposed for iron-gall documents were evaluated with hyperspectral imaging after the documents were exposed to accelerating aging conditions (heat, humidity and light). HSI was able to quantify spectral changes of inks caused by either the aqueous treatments or the aging factors.

A quantitative hyperspectral imaging system based on tunable light sources was designed for acquisition of historical documents at the National Archive of Netherlands [12]. The resulted hyperspectral information allowed for numeric assessment of paper yellowing, monitoring of ink aging, identification of biological and physical damages, such as ink corrosion, ink discrimination and legibility enhancement. The hyperspectral images offered additional content and the same accuracy of results would have been otherwise difficult to establish based on conventional visual inspection or conventional color-photography. The advantage of hyperspectral imaging over traditional 3-channel RGB color scans of handwritten documents is emphasized within an ink-mismatch detection, where two different inks with really close colors, but different spectral signatures appear identical in the visible range to the human eye, a phenomenon known as metamerism [1]. For this reason, access to the invisible range of the electromagnetic spectrum, where ink dissimilarity is more recognizable, is necessary. By means of hyperspectral imaging, this is possible, as information from the invisible electromagnetic bands of the spectra can be obtained thanks to a more complete sampling than that provided by the RGB subset. The strength of the HSI method in ink discrimination in comparison with other analytical methods existing in the literature is reinforced in [15], where different gel inks were studied.

The information from the invisible range can be gathered from traditional spectroscopic measurements as well [16]. However, when dealing with documents and inks, the traditional spectrometric instruments (e.g: spectrophotometer), have the measurement port wider than the ink line (3-8mm diameter in comparison to 1mm), so that it outputs the spectra of the ink combined with the spectra of the surrounding paper [14]. By contrast, hyperspectral imaging has the advantage of capturing spectral information for every pixel along the spatial dimension [10], so it is possible to acquire the spectra of the thin ink without major influence from the surrounding substrate [14].

With our study, we will further on prove the superiority of hyperspectral imaging over conventional color method by focusing on the particular aspect of ink-separation.

III. Image Dataset

A. Hyperspectral Acquisition

This study was conducted by Colour and Visual Computing Laboratory, Gjøvik University College, Norway in collaboration with National Library of Norway. Manuscripts selected for this study are representative cases of more general legibility issues that ancient manuscripts are undergoing. This includes fading of characters, overwritten text, mixing of writing from different authors, mixing of inks, bleed-through effect etc. In addition to the acquisition of hyperspectral images, the case study focused on improving the legibility and visualization of the manuscripts.

The hyperspectral data for this research was obtained using the HySpex line scanning hyperspectral camera VNIR-1600 manufactured by Norsk Elektro Optikk.. This camera captures images in the visible and near infrared spectral range 400 to 1000 nm with 160 spectral bands. The VNIR-1600 has a spectral sampling of 3.7 nm and captures 1600 spatial pixels across the field of view (10cm in the present case). The camera and the light source was fixed and manuscript was placed in a platform, which was allowed to move in a linear translator. Acquisition distance was 21 cm from the camera. Software synchronizes the speed of the translator with the camera frame rate. Light from a Xenon source guided via fiber optic cable was used for illuminating the manuscript.An equalization filter was used in front of the focusing lens to reduce the radiation in the bands in which the sensor is the most sensitive, allowing for correct exposure at the wavelength edges. This arrangement improves the noise levels in the blue region as well as the NIR (near infrared) region. Spectralon reference was also acquired along with the images and this has been used to normalize the images.

B. Hyperspectral Data

Several subsets of the acquired hyperspectral images are shown in Fig. 1. Representative subsets were used instead of the full images in order to reduce the computational requirements for processing purposes. Each of these four images are of spatial dimension 500×500 pixels and 160 spectral channels. In the following, the characteristics of each images will be briefly discussed. Also, the 4 documents will be referred further on with M1, M2, M4 and M5, according to the correspondence defined in Fig. 1.

Manuscript M1 represents a "Letter from Nathalia Munch (1812-1900) to Peter Andreas Munch (1810-1863)", dated 27.09.1833. There are two different inks used in document M1, i.e. purple and black ink. Even though the ink separation task in this case might seem trivial because the two inks are perceptually different, through image processing techniques, the benefit is that the two inks can be visualized separately, discarding the overlap.

Document M2 is a draft of "Bjørn Bjørnson "Over Ævne II" fair copy of Karoline Bjørnson with Bjørnson corrections". Thus, there are two different types of writing for each one of the two authors, supposedly with different types of black ink. What is known about the characteristics of the two different handwriting is that one handwriting generally has larger shape and the letters are more rounded and the other handwriting is more slanted and of smaller width.

Manuscript M4 is a "Letter from Gustav Kielland to unknown", dated 14.06.1859. The document's characteristic is the bleed-through effect of the ink from the verso side of the page appearing visible on the other side.The bleed-through is strongly visible in the front-side. In addition to this, there is an overlap between the bleed-through and the ink on the front-side.

Document M5 is a "Letter from Roald Amundsen to Douglas Mawson", dated 20.07.1914 and it contains text in process of fading, written on the substrate, where a watermark that has little to no visibility for the naked eye is present.

C. Color Image Simulation

Given a hyperspectral image, we can obtain a three-channel color image through several ways. To simulate how the same acquired objects look like according to human visual system, a color transformation using the spectral power distribution of an illuminant (typically D65) and color matching function should be applied to the hyperspectral image. This color transformation shall further be referred to as CMF method.

Another way of simulating a color image is by directly using the grayscale images obtained at the peak sensitivity of the hyperspectral scanner as the RGB channels [17]. Since these band images represent the peak sensitivities of the hyperspectral scanner, they will provide better contrasts. Furthermore, by directly using three band images as the RGB channels, variations existing in the bands are preserved. If a color transformation incorporating either human or camera sensitivity function was used instead, these variations would have been lost, because the color transformation essentially works as a weighted average function. This color simulation method based on peak sensitivities will be further referred to as FIXED method.

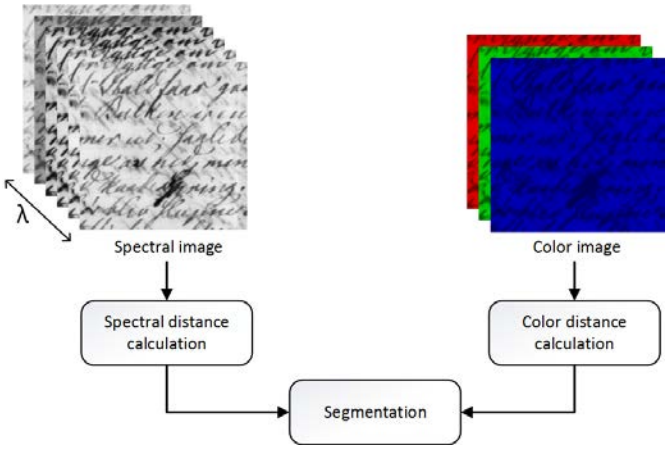


Fig. 2. Ink segmentation image processing steps

IV. Distance-Based Segmentation

Since the aim of this study is to show the advantage of hyperspectral imaging over that of traditional color imaging for the task of ink separation, both hyperspectral and simulated color images are used. The general algorithm flow is as shown in Fig. 2. Both hyperspectral and simulated color data of a historical document will be processed in its respective domain. After this distance calculation step, the ink separation task will be carried out using a simple binary classification.

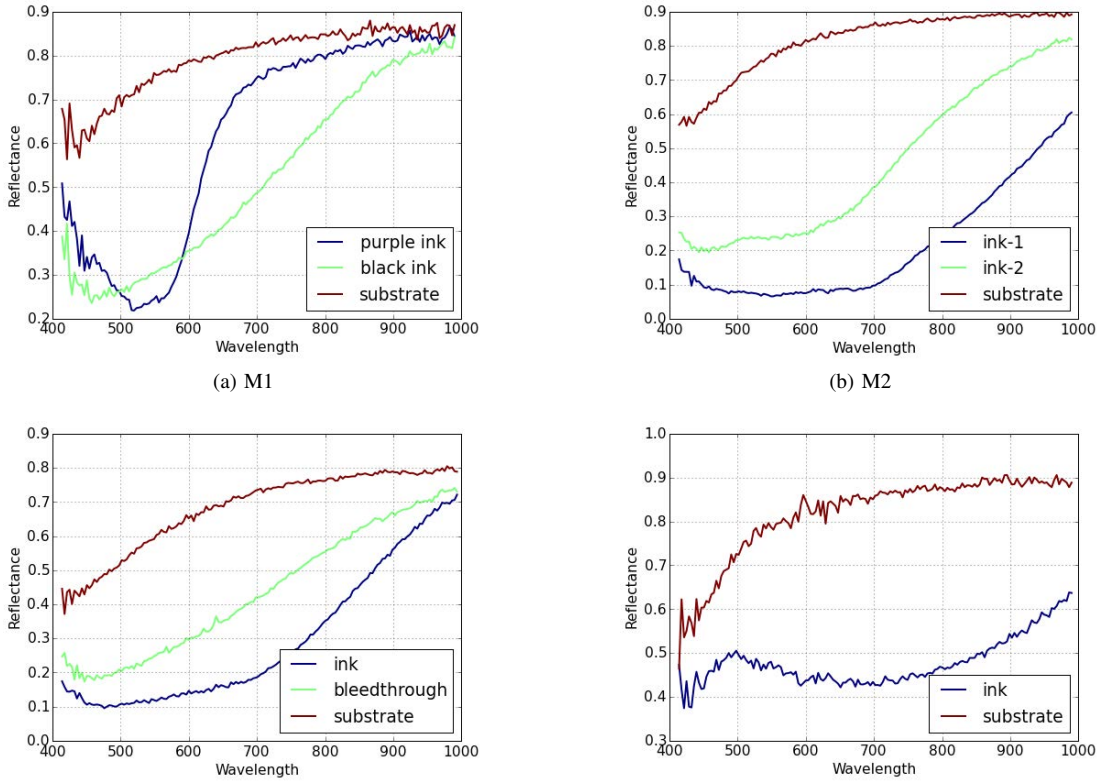
With distance-based classification, the class of a pixel is determined simply to which known class it is the most similar, or in other words, having the least distance value. In the following, the distance functions used for both color and spectral images are specified.

A. Spectral distance

The ink separation task can be regarded as a classification problem. However, depending on the characteristics of the ink and substrate within an image, we would have to use different distance functions [18]. If two inks are of different hue, in spectral domain it means that the two color signals will have different shape. In such case, a distance function that is based on the Spectral Correlation Mapper [19], see Eq. 2, will be employed. On the other hand, if two inks are of the similar hues but different lightness, the Euclidean distance function shown in Eq. 1 can be used. When both magnitude and shape difference are important, Euclidean distance of cumulative spectrum (ECS) shown in Eq. 3 will be used.

$$d_{Euc}(S_1,S_2) = \left(\sum_{\lambda} \left|s_{1,\lambda} - s_{2,\lambda}\right|^2\right)^{\frac{1}{2}} \tag{1}$$

$$R(S_1,S_2) = \frac{\sum_\lambda (s_{1,\lambda} - \overline{s_{1,\lambda}})(s_{2,\lambda} - \overline{s_{2,\lambda}})}{\sqrt{\sum_\lambda (s_{1,\lambda} - \overline{s_{1,\lambda}})^2}\sqrt{\sum_\lambda (s_{2,\lambda} - \overline{s_{2,\lambda}})^2}}$$



(a) M1 (b) M2

(c) M4 (d) M5

Fig. 3. Spectral reflectance signals of signatures found in the four historical documents.

$$d_{Cor}(S_1,S_2) = 1 - \frac{1+R(S_1,S_2)}{2} \tag{2}$$

$$d_{ECS}(S_1,S_2) = \left(\sum_{\lambda} \left| \int s_{1,\lambda} d\lambda - \int s_{2,\lambda} d\lambda \right|^2 \right)^{\frac{1}{2}} \tag{3}$$

B. Color distance

In the traditional colorimetric approach, the ink separation can be considered equivalent to color segmentation [1]. The simulated color images were transformed to the CIELAB color space and then, the CIELAB color difference was computed for the same pixel selections that were used in the spectral approach. More precisely, a patch of 2×2 pixels was used for taking the color or spectral value of every component: substrate, ink, bleed-through or watermark. The same procedure was applied for the color images transformed to HSV (Hue, Saturation, Value) space, using the Euclidean distance as the criteria for separation in the hue dimension (when there are different hues) or value dimension (when the hues are very similar, but lightness varies).

V. Result and Discussion

As mentioned previously, document M1 shown in Fig. I has two inks of different hue. The spectral reflectance signals of the two inks and the substrate are as shown in Fig. 3a. From this figure we can observe clearly that the two inks are indeed of different shape. Therefore, spectral correlation based distance (Eq. 2) is used for the segmentation of the two inks. On the other hand, in the color domain, the CIELAB difference is used on the simulated color image. The results are as shown in Fig. 4. While there is no ground truth data available, the obtained results for spectral approach are considered good based on subjective assessment, since we know that the two different inks are written in two different directions, i.e. horizontal and vertical. Comparing the spectral approach with the color approach, the two inks are better separated and less misclassification occurred in the former case.

The two different writings found on document M2 might be of identical inks as their spectral shapes are different mostly by magnitude, see Fig. 3b. Since the inks and the substrate have different spectral shape, correlation-based distance is used in order to extract the writings. Then, in order to differentiate the two inks, Euclidean distance is used. Two approaches are used in color domain, i.e. CIELAB difference and hue difference in HSV color space. The results are shown in Fig. 5. With magnitude/ lightness being the main difference of the two inks, in addition to the size and shape of the characters, it becomes a challenging task even for the spectral approach since spatial information needs to be incorporated. Without incorporating spatial information, one of the main issue is ink corrosion which happen at the edges of each characters. The advantage of spectral approach is that it is able to minimize the amount

of misclassified pixels caused by corrosion.

For document M4, we are dealing with the bleed-through phenomenon, where the ink written on the verso is overlapping with the writing on the front side of the paper [20], the spectral approach gives better results than the color approaches using CIELAB and HSV color spaces (see Fig. 6). ECS distance was used to compute spectral distances since in this case both magnitude and shape differences are of importance. See the relevant spectral reflectance signals in Fig. 3c. The results of color approaches are similar incorrectly classify as front-side ink some regions from the bleed-through that the spectral image discards, see regions inside the blue circles in Fig. 6.

There are two main interests in document M5, i.e. the fading ink and the almost-invisible watermark. Through ink segmentation using spectral and color approaches, we are able to recover the fading ink, see Fig. 7, with spectral approach giving better performance as expected. As for the watermark processing, the results are shown in Fig. 8. In the results of color processing, the watermark is still not so visible. On the contrary, with hyperspectral images, we are provided with images from the near-infrared (NIR) regions where the watermark is more visible. Single-band image processing is an effective method for contrast enhancement and watermark visualization [12], [21].

VI. Conclusions

A case study was conducted for the historical manuscripts from the collection at the National Library of Oslo. Several handwritten documents of cultural importance were scanned with a hyperspectral camera. Then, ink segmentation results were compared between hyperspectral images and conventional color images, using distance-based criteria. Based on subjective visual assessment, the inks are more effectively separated based on their spectral signatures, than on their color coordinates. In addition, access to the invisible range of the spectra aids to visualize elements from the documents hardly visible otherwise, such as watermarks. Therefore, hyperspectral image processing stands as a strong instrument for improving the legibility of handwritten document analysis.

VII. Acknowledgements

The authors would like to offer their thanks to Norsk Elektro Optikk for their support in providing the hyperspectral camera for the acquisition campaign and to the National Library of Oslo for their hospitality and willingness to collaborate.

References

[1] Z. Khan, F. Shafait, and A. Mian, "Hyperspectral imaging for ink mismatch detection," in *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*, 2013, pp. 877–881.

[2] C. Balas, V. Papadakis, N. Papadakis, A. Papadakis, E. Vazgiouraki, and G. Themelis, "A novel hyper-spectral imaging apparatus for the non-destructive analysis of objects of artistic and historic value," *Journal of Cultural Heritage*, vol. 4, pp. 330–337, 2003.

[3] J. A. Richards and J. Richards, *Remote sensing digital image analysis*. Springer, 1999, vol. 3.

[4] G. A. Blackburn, "Hyperspectral remote sensing of plant pigments," *Journal of experimental botany*, vol. 58, no. 4, pp. 855–867, 2007.

[5] H. Deborah, S. George, and J. Hardeberg, "Pigment mapping of the scream (1893) based on hyperspectral imaging," in *Image and Signal Processing*, ser. Lecture Notes in Computer Science. Springer International Publishing, 2014, vol. 8509, pp. 247–256.

[6] A. Nunez, M. J. Mendenhall, and K. Gross, "Melanosome level estimation in human skin from hyperspectral imagery," in *Hyperspectral Image and Signal Processing: Evolution in Remote Sensing, 2009. WHISPERS '09. First Workshop on*, Aug 2009, pp. 1–4.

[7] A. A. Gowen, C. P. O'Donnell, P. J. Cullen, G. Downey, and J. M. Frias, "Hyperspectral imaging–an emerging process analytical tool for food quality and safety control," *Trends in Food Science & Technology*, vol. 18, no. 12, pp. 590–598, 2007.

[8] R. Shrestha and J. Y. Hardeberg, "Evaluation and comparison of multispectral imaging systems," *Color and Imaging Conference*, pp. 107–112, 2014.

[9] S. J. Kim, F. Deng, and M. S. Brown, "Visual enhancement of old documents with hyperspectral imaging," *Pattern Recognition*, vol. 44, no. 7, pp. 1461–1469, 2011.

[10] D. Goltz, M. Attas, E. Cloutis, G. Young, and P. Begin, "Visible (420–720 nm) hyperspectral imaging techniques to assess inks in historical documents," *Restaurator*, vol. 30, no. 3, pp. 199–221, 2009.

[11] R. Hedjam, M. Cheriet, and M. Kalacska, "Constrained energy maximization and self-referencing method for invisible ink detection from multispectral historical document images," in *Pattern Recognition (ICPR), 22nd International Conference on*. IEEE, 2014, pp. 3026–3031.

[12] R. Padoan, T. A. Steemers, M. Klein, B. Aalderink, and G. De Bruin, "Quantitative hyperspectral imaging of historical documents: technique and applications," *ART Proceedings*, 2008.

[13] P. Shiel, M. Rehbein, and J. Keating, "The ghost in the manuscript: Hyperspectral text recovery and segmentation," *Codicology and Palaeography in the Digital Age, M. Rehbein and PS und Torsten Schaßan, Eds. Norderstedt: Books on Demand*, pp. 159–174, 2009.

[14] S. Tse, D. Goltz, S. Guild, V. Orlandini, M. Trojan-bedynski, and M. Richardson, "Effect of aqueous treatments on nineteenth-century iron-gall-ink documents: assessment using hyperspectral imaging," *Book Paper Group Ann*, vol. 28, p. 75, 2009.

[15] G. Reed, K. Savage, D. Edwards, and N. N. Daeid, "Hyperspectral imaging of gel pen inks: An emerging tool in document analysis," *Science & Justice*, vol. 54, no. 1, pp. 71–80, 2014.

[16] D. M. Goltz, B. Piniuta, E. Huebner, M. Attas, E. Cloutis, and J. Broomhead, "Spectroscopic approaches for studying faint text on a wooden tally from invincible (1758)," *International Journal of Conservation Science*, vol. 4, no. 1, 2013.

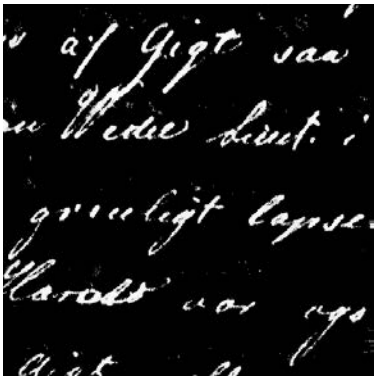
[17] Z. Khan, F. Shafait, and A. S. Mian, "Towards automated hyperspectral document image analysis," in *Automated Forensic Handwriting Analysis (AFHA), 2nd International Workshop on*, 2013, pp. 41–45.

[18] H. Deborah, N. Richard, and J. Y. Hardeberg, "A comprehensive evaluation on spectral distance functions and metrics for hyperspectral image processing," *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, 2015, to appear.

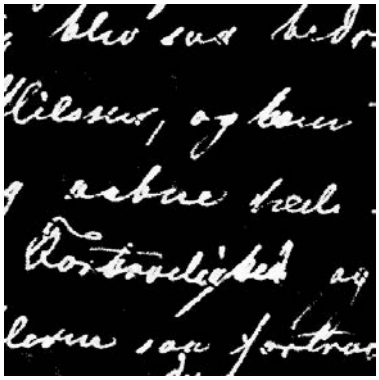
[19] O. de Carvalho Jr., R. Guimaraes, R. Gomes, A. de Carvalho, N. da Silva, and E. Martins, "Spectral multiple correlation mapper," in *Geoscience and Remote Sensing Symposium. IEEE International Conference on*, 2006, pp. 2773–2776.

[20] Y. Huang, M. S. Brown, and D. Xu, "A framework for reducing ink-bleed in old documents," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–7.

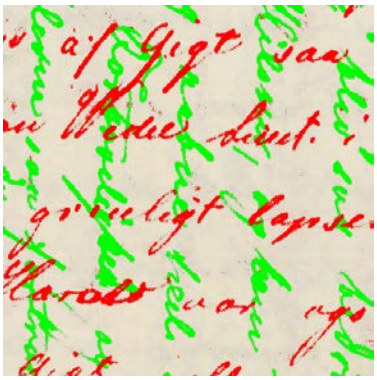
[21] S. J. Kim, S. Zhuo, F. Deng, C.-W. Fu, and M. S. Brown, "Interactive visualization of hyperspectral images of historical documents." *IEEE transactions on visualization and computer graphics*, vol. 16, no. 6, pp. 1441–1448, 2009.



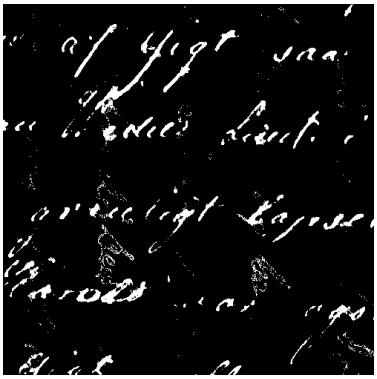
(a) Spectral: Black ink



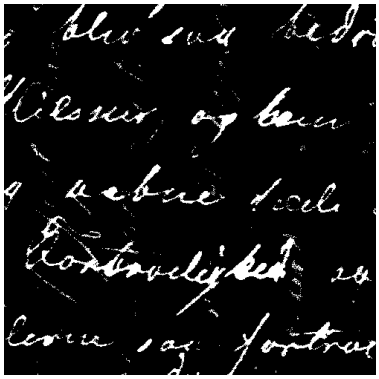
(b) Spectral: Purple ink (rotated)



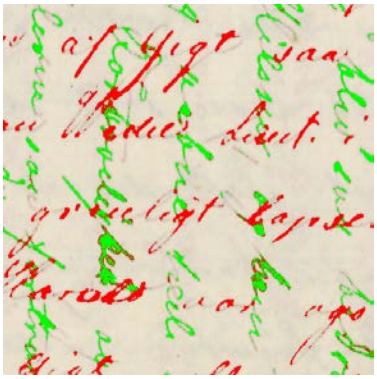
(c) Spectral: Segmented image



(d) CIELAB: Black ink

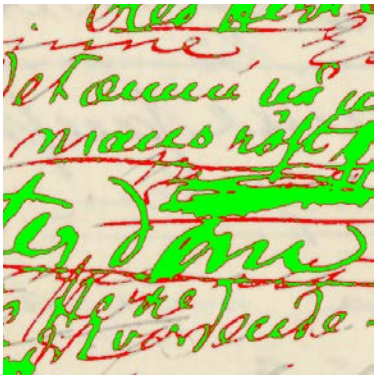


(e) CIELAB: Purple ink (rotated)

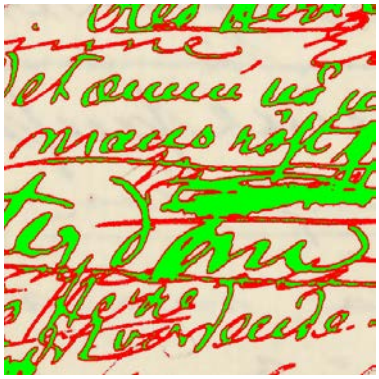


(f) CIELAB: Segmented image

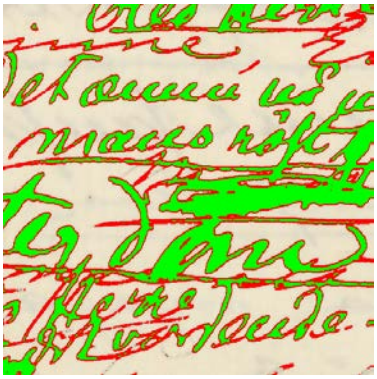
Fig. 4. Segmentation results of the two inks found in document M1, using spectral and color approaches.



(a) Spectral

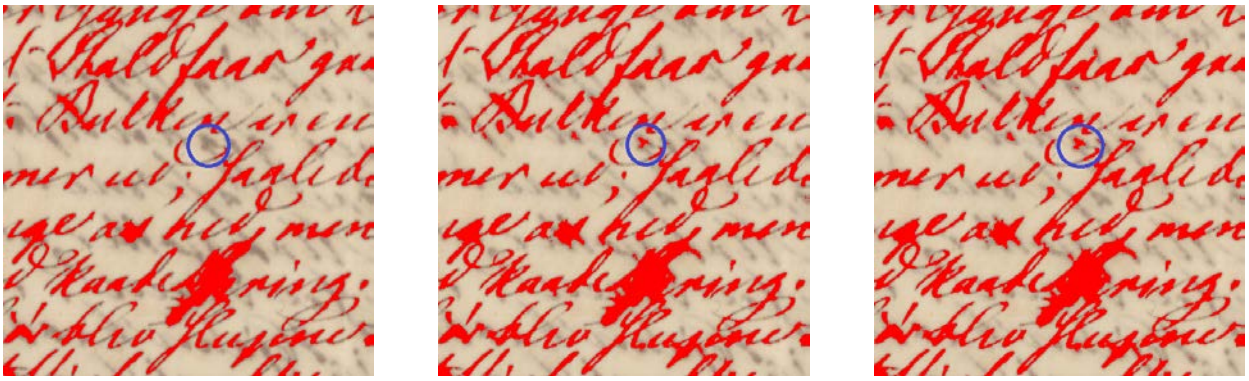


(b) CIELAB



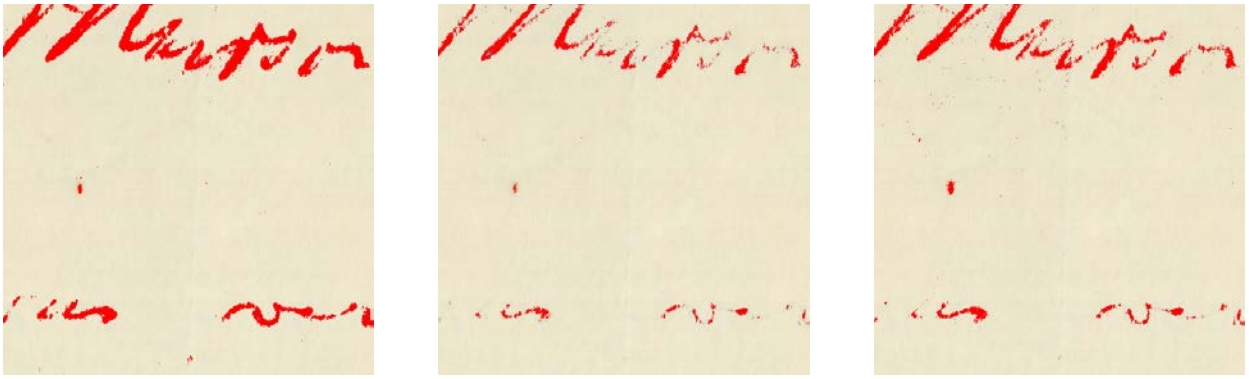
(c) Value channel of HSV

Fig. 5. Detection of two different inks on document M2. The spectral approach is able to minimize misclassified pixels due to corrosion of ink.



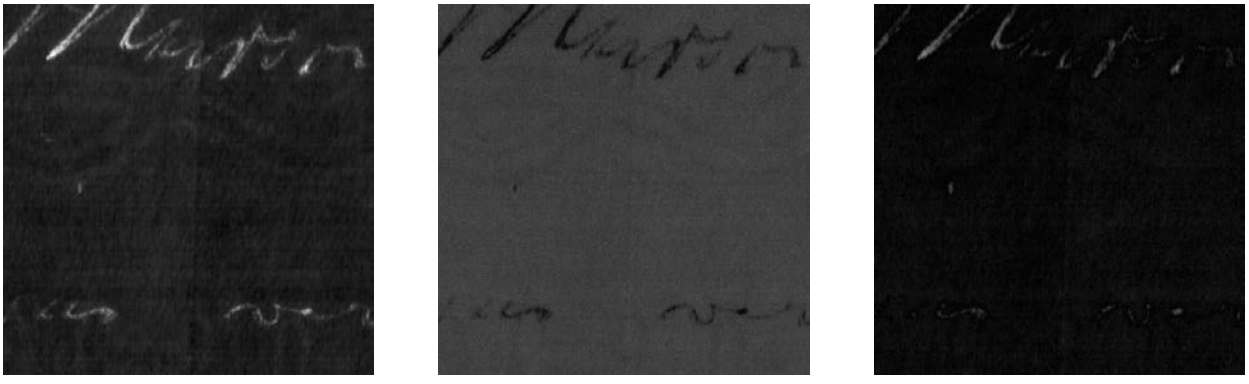
(a) Spectral (b) CIELAB (c) Value channel of HSV

Fig. 6. Detection of ink on document M4, discarding the bleed-through inks. Regions inside the blue circles are where examples of where the color approaches failed to separate the bleed-through and front side inks.



(a) Spectral (b) CIELAB (c) Value channel of HSV

Fig. 7. Detection of ink on document M4 using a spectral and two color approaches.



(a) Spectral (b) CIELAB (c) Value channel of HSV

Fig. 8. Three different approaches were used to try to bring out the watermark found on the substrate of document M5. The information given by the spectral approach were obtained from the near-infrared (NIR) region, and indeed it gives more visibility to the watermark.