Ink bleed-through removal of historical manuscripts based on hyperspectral imaging

Cristian da Costa Rocha, Hilda Deborah, and Jon Yngve Hardeberg

Norwegian Colour and Visual Computing Laboratory
Department of Computer Science
NTNU – Norwegian University of Science and Technology

Abstract. Old manuscripts can be degraded by different reasons over time. The ink bleed-through from the back side of the page is a common problem which compromises the readability and the aesthetics of the document. Unsupervised bleed-through removal is a challenging task as pixel intensity of bleed-through areas can be very similar to fine-details of the writing. This paper provides a brief review about state-of-the-art methods for bleed-through removal. Moreover, we propose an algorithm for the segmentation of the bleed-through areas from a manuscript collected from the National Library of Oslo using hyperspectral imaging. This work also presents the restoration of the collected manuscript applying inpainting algorithms based on our segmentation approach.

Keywords: hyperspectral, bleed-through removal, inpainting

1 Introduction

Old manuscripts are an important part of history. They provide information about the culture and people's values over the centuries. However, through the years, different types of degradation can compromise the readability and aesthetics of these ancient documents. One of the degradation types which can be observed in many cases is bleed-through of the ink, from the back side of the page to the front. With such a challenge, the importance of an ink bleed-through removal algorithm comes in its capability to improve document readability while keeping their original aesthetics.

There are two main categories of bleed-through removal methods with regards to the page sides being scanned, i.e., blind and non-blind approaches [9]. In the non-blind approach, both sides of the page (*recto-verso*) are scanned. This would give locations of the bleed-through spots, provided by the *verso* page. On the other hand, this information is not available for the blind approach, since only the *recto* is scanned. One of the main challenges for both methods is that pixel intensity from bleed-through can be really similar to fine details of the ink. In these cases, the distinction between bleed-through and ink requires a more detailed evaluation. Taking that into account, for a more accurate discrimination it is important to analyze the spectral reflectance of the document.

Hyperspectral imaging (HSI) is a growing research field which has been important in many applications, e.g., agriculture [4], medicine [5], and remote sensing [3]. Providing more accurate information of the object/ scene in different regions of the electromagnetic spectrum, HSI makes it possible to have a good overview of the important wavelengths for a given application. In this work, we are exploiting the potential of HSI for the bleed-through removal of an old manuscript. A brief state of the art of bleed-through removal methods is provided in Section 2. In Section 3, several methods are extended for use in the hyperspectral domain, and a proposal method is also presented. Experimental results and discussion are provided in Section 4 and 5, respectively, before to conclude the work in Section 6.

2 Brief state of the art

Ink bleed-through problem has been treated as the task of classifying the background (substrate), foreground (writings), and bleed-through ink. After this classification step, assuming that bleed-through spots are detected, an inpainting step would usually follow to replace these spots. In [6], a conditional random field based method was proposed. In it, the three classes are assumed to be having Gaussian distributions and further modelled as conditional probability distributions. The performance of this method was evaluated against the available ground truth images. The study also proposed *random-fill* inpainting algorithm.

Another direction in this topic pursues the task as a binarization problem. In such case, the main concern is to segment all the inks (writings and bleed-through) from the background. Assuming that the foreground pixels have lower intensity values, an iterative k-means based method was proposed to segment these pixels [2]. Combined with PCA as a decorrelating step prior to k-means, at each iteration, only darker pixels are fed as the input. As a result, only darker regions of the image, i.e., the inks, are segmented. This approach fails when pixel intensities of the bleed-through are similar to fine details of the ink or, in case of severe degradation, to the background.

A limitation of the previous approach is its failure to exploit spatial properties of the image. Assuming that the bleed-through is not as sharp as the ink, spatial properties is taken into consideration in [7], through applying an edge detection method to get the edges of the main text. To determine whether the location of a pixel is near an edge, a window of size 7×7 is used. When the condition is satisfied, the pixel is said to be part of the background, and is subsequently replaced by a relevant pixel intensity value. The drawback of this method is that the recovered text does not look natural. Moreover, if a pixel belonging to the ink does not actually fall inside the 7×7 window from an edge, it will automatically be assumed as that of the background. As a result, there will be cases when the recovered text has white spots/ holes in the middle of the ink region.

The previous approaches are focused on the segmentation part of bleed-through removal task. However, inpainting is also an important and integral part of the task. Aiming to improve readability of texts in the manuscript, in

this step, pixel values of those detected as the bleed-through are replaced with a certain value such that the final processing result looks natural. One of the most used inpainting algorithm proposes to use a weighting function of the area in the neighborhood of the bleed-through spots [8]. This weighting function is said to be able to replace the affected spots (i.e., or bleed-through in our case) smoothly, without sharp edges in the inpainted areas.

3 Experimental Design

Based on the previous state of the art, the main steps of a bleed-through removal method can be categorized into segmentation and inpainting. But since our input image is a hyperspectral one, some adaptations need to be carried out to enable to use the aforementioned methods. This implementation strategies and choices will be described in more details in the following, as well as the hyperspectral image input itself. But before going into the details, workflow of the experiment to be carried out is provided in Fig. 1.

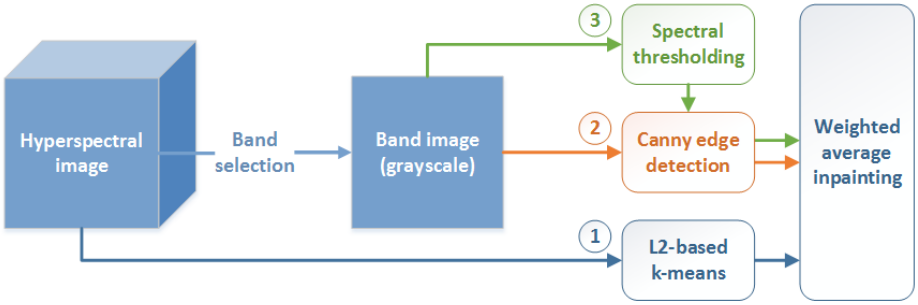
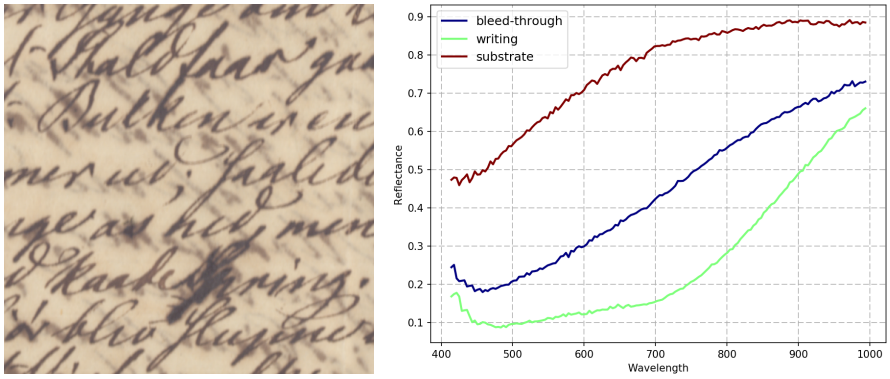


Fig. 1: Comparison workflow of bleed-through removal methods carried out in this work. The iterative k-means directly uses the hyperspectral image due to its use of L2 metric, while band selection is needed as pre-processing step in order to carry out the other methods.

3.1 Image Target

The hyperspectral image of a *Letter from Gustav Kielland to unknown* written on June 14th 1859 from the National Library of Oslo will be analyzed in this study. Its subset is shown in Fig. 2a. The hyperspectral image is of 160 bands from 414.2 to 993.7 nm, in approximately 3.6 nm intervals. Example spectra from the bleed-through, writing, and substrate groups is also shown in Fig. 2b.



(a) (b)

Fig. 2: (a) Color image of a subset of the target image and (b) samples of spectra belonging to the bleed-through, writing/ink, and substrate groups. Wavelength at approximately 727.6 nm is selected as the input to canny edge detection and spectral thresholding.

3.2 Segmentation

This step aims to detect locations of the bleed-through, to further replace their pixel intensity values in the inpainting step. The ones we are employing are iterative k-means clustering [2] and canny edge detection [7].

Iterative k-means Taking a classification approach, with three classes, the k-means algorithm was set with k=3 and the iteration limit was set to 25. Then the whole hyperspectral image is directly used as input to this method. This direct use is possible since L2 metric (Euclidean) is embedded within k-means and the metric is able to compute magnitude difference between two spectra [1].

Canny edge detection This edge detection algorithm is developed for the grayscale domain. Thus, to use it for our hyperspectral case, a specific wavelength maximizing the difference between the substrate, writings, and bleed-through is selected, i.e., around 727.6 nm. The last parameter is the window size, that is fixed as 4×4 pixels.

Proposed approach: Spectral thresholding + Canny edge detection Spectral reflectances of the analyzed document was also explored and to be exploited in this work. By observing the spectra shown in Fig. 2, differences between the three groups (ink, bleed-through, and substrate) is evident. As shown in the figure, spectral region where the bleed-through is more distant from the ink is in the near infrared domain, around 727.6 nm. Thus, a threshold is set

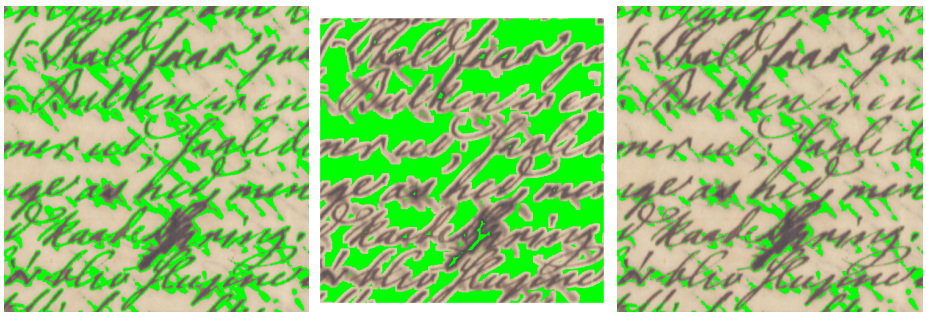
for segmenting out the bleed-through in this specific wavelength, i.e., pixels with values between [0.40, 0.65]. However, the thresholding step would remove not only the bleed-through, but also some fine details of the ink. To tackle this issue, Canny edge detection is applied to the same infrared band image. The resulting edge information is then used as reference, to reconstruct the lost fine details of the ink in the thresholding result. If a given pixel is within a 2×2 window distance from any edge, it is assumed to be part of the ink. Thus, the pixel value is recovered from the original image and it will not be segmented out. As a result, the final segmentation will have the values selected in the thresholding step and which were not in a 2×2 window distance from the edges.

3.3 Inpainting

In this experiment, the inpainting method to employ picks a weighted-sum of the neighborhood of the bleed-through spots instead [8]. And in this experiment, the weighting function is the average taken from a 3×3 pixel window.

4 Experimental Results

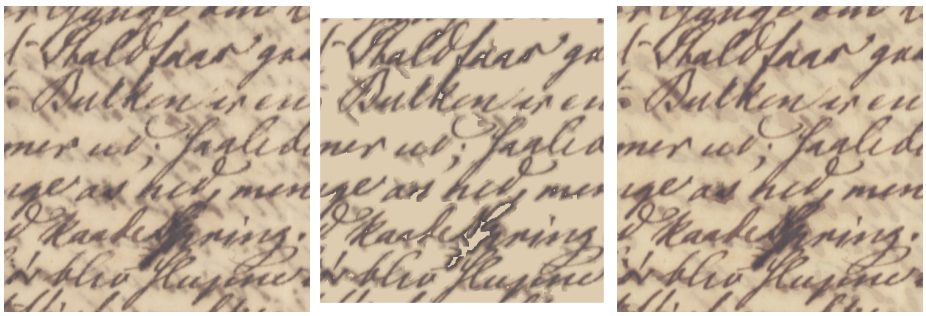
The resulting image for the segmentation of the bleed-through is illustrated in Fig. 3. In the k-means approach, it is clear that most of the bleed-through is detected in the segmentation step. On the other hand, fine details of the ink are also segmented. For the Canny edge detection method, the writing is preserved, but the bleed-through spots nearby the text are not segmented. The combined approach proposed in this paper segments most of the bleed-through spots smoothly while preserving the fine details from the writing.



(a) k-means (b) canny (c) proposed approach

Fig. 3: Segmentation results from the three compared methods.

In Fig. 4, the results for the inpainting step are shown. In this case, the proposed approach is compared with k-means using the inpainting algorithm from Telea, *et al.* [8]. As the bleed-through segmentation using Canny edge detection removes most of the background, the inpainting algorithm does not have good results. Taking that into account, the inpainting algorithm for the Canny edge detection segmentation is just the replacement of the segmented spots by the average of the background.



(a) k-means (b) canny (c) proposed approach

Fig. 4: Results of inpainting step, with results from Fig. 3 as inputs.

5 Discussion

For the k-means based bleed-through segmentation, we found a similar problem to [2] where we were able to remove most of the bleed-through but the readability was compromised as the fine-details of the ink were also being extracted. The main issue is that the pixel intensity on the final details of the ink is very similar to the intensity on most of the bleed-through. With the k-means algorithm, the bleed-through and the least intense parts of the ink will be clustered in the same group. The k-means algorithm was also tested using just one spectral band on different wavelengths of the cube but the problem with the fine-details of the ink was found in these cases as well. Moreover, the bleed-through is not totally extracted with this method. The degradation of the document is so severe in some cases that the intensity in some regions in the bleed-through has similar intensity to the ink and, as a result, part of the bleed-through is not removed.

On the other hand, the canny edge detection algorithm does not preserve the aesthetics of the manuscript, as we are just replacing the background by its mean value. Furthermore, the bleed-through spots which are nearby the ink are also reconstructed. Another problem found using canny edge detection is that ink pixels which are too far away from the edges are not reconstructed. As a result, there are many bright spots in the middle of some letters. While it would

be possible to recover these bright spots in the ink using a bigger window size, bleed-through spots which are around the edges would also be recovered due to the big window size.

To the best of our knowledge, this paper presents the first approach which combined the segmentation based on thresholding with Canny edge detection for bleed-through removal. As discussed in this paper, relying only on pixel intensity is a risky way for bleed-through segmentation, as there are many bleed-through spots with similar intensity to the fine details of the ink. Taking that into account, the main idea of our proposed method is to firstly remove the bleed-through regions using the spectral threshold, which is also going to remove the final details of the ink. Then, using Canny edge detection, we recover the regions removed from the ink in the first step. Both methods are complimentary.

The inpainting algorithm used in this experiment replaces the bleed-through regions based on the nearby pixels. One of the problems of this method is that the ink can also be considered when inpainting segmented spots, which is going to give a darker appearance to some processed areas. However, this method keeps the readability and the original aesthetics of the document, while reducing the intensity of the bleed-through areas.

6 Conclusion

In this article, we present our work in the task of removing bleed-through inks from historical manuscripts based on hyperspectral imaging. A state of the art of bleed-through removal algorithm is also presented. A contribution has been proposed for unsupervised bleed-through removal, by exploiting the availability of more accurate information from the hyperspectral image. This method is able to improve the readability of the document while keeping its original aesthetics - differently from other methods which remove part of the ink and do not preserve the aesthetics of the original manuscript [2, 7, 9].

As a future work, in addition to improving both the segmentation and inpainting steps by exploiting more of the spectral image, it is important to develop a quality evaluation measure for the results. Two directions can be taken, i.e., objective or subjective through psycho-visual experiments. Despite the unavailability of ground-truth information, an objective evaluation is still possible thanks to the no-reference approach of image quality assessment.

References

1. Deborah, H., Richard, N., Hardeberg, J.Y.: A comprehensive evaluation on spectral distance functions and metrics for hyperspectral image processing. IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens. 8(6), 3224–3234 (2015)
2. Fadoua, D., Le Bourgeois, F., Emptoz, H.: Restoring ink bleed-through degraded document images using a recursive unsupervised classification technique. In: Document Analysis Systems VII. pp. 38–49. Springer Berlin Heidelberg (2006)

3. Loggenberg, K., Strever, A., Greyling, B., Poona, N.: Modelling water stress in a shiraz vineyard using hyperspectral imaging and machine learning. Remote Sensing 10(2), 202 (2018)
4. Ravikanth, L., Jayas, D.S., White, N.D.G., Fields, P.G., Sun, D.W.: Extraction of spectral information from hyperspectral data and application of hyperspectral imaging for food and agricultural products. Food Bioprocess Tech. 10(1), 1–33 (2017)
5. Regeling, B., Thies, B., Gerstner, A.O.H., Westermann, S., Müller, N.A., Bendix, J., Laffers, W.: Hyperspectral imaging using flexible endoscopy for laryngeal cancer detection. Sensors 16(8), 1288 (2016)
6. Sun, B., Li, S., Zhang, X.P., Sun, J.: Blind bleed-through removal for scanned historical document image with conditional random fields. IEEE Trans. Image Process. 25(12), 5702–5712 (2016)
7. Tan, C.L., Cao, R., Wang, Q., Shen, P.: Character extraction from interfering background - analysis of double-sided handwritten archival documents. In: Advances in Pattern Recognition – ICAPR 2001. pp. 93–102. Springer Berlin Heidelberg (2001)
8. Telea, A.: An image inpainting technique based on the fast marching method. Journal of Graphics Tools 9(1), 23–34 (2004)
9. Tonazzini, A., Salerno, E., Bedini, L.: Fast correction of bleed-through distortion in grayscale documents by a blind source separation technique. Int. J. Doc. Anal. Recog. 10(1), 17–25 (2007)