A shallow neural net with model-based learning for the virtual restoration of recto-verso manuscripts

Pasquale Savino[1,*,†], Anna Tonazzini[1,†]

[1]Istituto di Scienza e Tecnologie dell'Informazione, Consiglio Nazionale delle Ricerche, Via G. Moruzzi 1, 56124 Pisa, Italy

Abstract

We propose a fast procedure based on neural networks (NN) to correct the typically complex background of recto-verso historical manuscripts, where the texts of the two sides often appear mixed. The purpose is to eliminate the interfering, shining-through text, to facilitate both the work of philologists and paleographers and the automatic analysis of the linguistic contents. We adapt the learning phase of a very simple shallow NN to exploit the information of the registered recto and verso sides of the manuscript without the need for a large class of other similar manuscripts. Hence, the training set is self-generated from the data images based on a theoretical mixing model that accounts for ink spreading through the paper fiber and for ink saturation in the text superposition areas. Operationally, we select pairs of patches containing clean text from the manuscript and then mix them symmetrically using the model with varying parameters that span the allowed range. This makes the NN able to generalize to diverse amounts of ink seeping and then classify different manuscripts. We show comparisons between the results obtained on heavily damaged manuscripts with this NN and other approaches. From a qualitative point of view, the proposed method seems quite promising.

Keywords

Ancient manuscript virtual restoration, degraded document binarization, recto-verso registration, bleed-through removal, shallow multilayer neural networks

1. Introduction

The natural degradation of materials almost always damages historical and archival manuscripts over time or by other accidental factors such as fires, floods, and poor conservation. Usually, these ancient manuscripts appear as the superposition of many different patterns or layers of information. Besides the main text and the paper texture, they may contain other informative features, such as annotations, miniatures, stamps, or non-informative interference due to the damages, such as humidity spots and molds or ink seeped from the reverse side.

An important aim of digital image processing techniques is to provide the scholars with digital versions that can help them in their work of reading and interpretation. Therefore, there is a request for algorithms of virtual restoration that attempt to put back the manuscripts to their original appearance by eliminating only the degradation without destroying the other

VIPERC2022: 1st International Virtual Conference on Visual Pattern Extraction and Recognition for Cultural Heritage Understanding, 12 September 2022

*Corresponding author.

†These authors contributed equally.

pasquale.savino@isti.cnr.it (P. Savino); anna.tonazzini@isti.cnr.it (A. Tonazzini)

0000-0002-8841-5440 (P. Savino); 0000-0001-6970-4725 (A. Tonazzini)

informative features. In this sense, the plurality of manuscript content should be analyzed and discriminated in such a way to preserve and highlight the useful patterns and remove the extra, useless patterns that can disturb or even make impossible the scholar's study [1].

Another goal of processing the digital image of a manuscript is to prepare it for automatic tasks of word spotting and/or character recognition. In this case, binarization is usually performed as the first step to extract the interesting foreground text against all other features considered, as a whole, complex background or noise. A broad interest exists in degraded document binarization [2, 3], and a large variety of methods have been proposed.

Among those, local and adaptive thresholding, or recurrent, convolutional, or deep neural networks can deal, to some extent, with degradation such as uneven illumination, image contrast variation, changes in stroke width and connection, faded or seeping ink [4, 5, 6, 7, 8].

Virtual restoration and binarization are often complementary or preparatory to each other. Indeed, a manuscript from which the strongest degradation has been removed can be binarized more effectively, as in [8] where a NN learns the degradation and iteratively refines the output, which is then binarized using Otsu's global threshold. Conversely, binary maps, in which the foreground text has been identified and extracted, can form the basis for an accurate restoration, as in [9] where the main text is mapped in a clean background obtained by inpainting.

Removing the bleed-through degradation, which occurs in manuscripts written on both sides of the paper, has raised a particular interest individually and outside the binarization context. Indeed, strong bleed-through cannot be removed entirely by binarization alone due to its significant overlap with the foreground text and the wide variation of its extent and intensity. Methods designed explicitly for bleed-through reduction have then been proposed. The so-called blind methods exploit the information of the front side alone, like the blind source separation technique proposed in [10], the recursive unsupervised segmentation suggested in [11], and the conditional random field presented in [12].

In general, images of both sides of the manuscript are available, and their joint use is recommended as it brings additional data, as informative as the primary one. This richness of information allows the design of algorithms that can selectively remove the unwanted interference alone, leaving the rest of the manuscript unaltered, thus performing a very fine virtual restoration. Examples in this respect can be found in [13], where a classification is performed by segmenting the recto-verso joint histogram with the aid of available ground truths, in [14], where a regularized energy uses a data term derived from small sets of user-labeled pixels and a smoothness term based on dual-layer Markov Random Fields, and in [15], where correlated component analysis is used to separate the information layers. In [16], the fidelity of the restored manuscript to the original one takes advantage of sparse image representation and dictionary learning.

However, the need for a perfect alignment of the two images greatly complicates the problem, especially in the presence of document skews, different image resolutions, or wrapped pages when scanning books. Registration algorithms specifically devoted to recto-verso manuscripts have been proposed in [17, 18, 19].

We adopted a simple multilayer shallow neural network with backpropagation training [20], and implemented it in such a way that it auto-adapts to the manuscript to be restored, i.e., it does not require preliminary learning from a large class of other similar manuscripts. The point of view is to automatically learn the degradation that affects the manuscript in question. The trained NN classifies the pixels of the manuscripts affected by that specific degradation as clean or noisy.

In the simplest way, degraded patches are drawn from the manuscript to learn the degradation, and then the corresponding ground truths must be somehow estimated. When an analytical data model exists that describes the degradation, the training set can be self-generated starting from ground truths drawn from the clean zones of the manuscript. Then the model can be used to generate the corresponding degraded patches. This second way of operating can be somehow more straightforward.

In our case, since we focus on the virtual restoration of recto-verso manuscripts, we use the theoretical mixing model proposed in [21], which approximates the physical phenomenon of the spreading of ink through the paper fibers and its seeping into the reverse side of the sheet. We use the model in the direct modality, i.e., for generating data consistent with the degraded manuscript that we wish to restore, and slightly modify it to correct some weaknesses we observed in correspondence of the occlusions between the front and back text. We show the performance of our NN on a real, heavily damaged manuscript, both in terms of binarization of the foreground text and virtual restoration.

The paper is organized as follows. In Section 2, we describe the data model used to build the training set from the observed recto-verso pair only. Section 3 is devoted to the details of the shallow NN architecture and the learning and recall phases. Section 4 analyzes some preliminary results on letters from the correspondence of Christoforus Clavius, conserved at the Historical Archive of the Pontificia Università Gregoriana in Rome. Finally, Section 5 concludes the paper.

2. Neural Network: building the training set by a data model

In most of the examined manuscripts, the seeped ink was also diffused through the paper fiber. Hence, the bleed-through pattern usually appeared as a smeared and lighter version of the opposite text generated. This does not mean that, on the same side, bleed-through was always lighter than the foreground text. In fact, on each side, the intensity of bleed-through is usually very variable, that is highly non-stationary, and sometimes can be as dark as the foreground text.

These considerations led us to adopt a measure of optical density, defined for each pixel t as $D(t) = -log\left(\frac{s(t)}{b}\right)$, where $s(t)$ is the intensity, and b represents the average background, and propose the description of the bleed-through degradation, to each observation channel, through the following non-stationary linear model:

$$\begin{aligned} D_r^{obs}(t) &= D_r(t) - q_v(t)log\left(\frac{h_v(t)\otimes s_v(t)}{b_v}\right) \\ D_v^{obs}(t) &= D_v(t) - q_r(t)log\left(\frac{h_r(t)\otimes s_r(t)}{b_r}\right) \end{aligned} \tag{1}$$

In eqs. (1), D^{obs} and D are the observed and the ideal optical density, with the subscripts r and v indicating the registered recto and reflected verso side, respectively, and $\otimes$ indicates convolution between the ideal intensity s and a Point Spread Functions (PSF), h, describing the smearing of ink penetrating the paper. Finally, the space-variant quantities q_r and q_v, in the range $[0, 1]$, have the physical meaning of ink penetration percentages from one side to the other.

In previous works [21, 19, 22], we proposed to invert the above model for virtually restoring the recto-verso pair. Based on the observed densities of the two sides, we first inverted the model by assuming an identically zero ideal density on the opposite side, thus obtaining estimates of the ink penetration percentages at each pixel. After some straightforward adjustments of these percentages, the system can then be solved with respect to the ideal density maps, from which the virtually restored manuscript sides are obtained.

The model approximates quite well the phenomenon of ink transparency almost everywhere, apart from the occlusion areas where the inks of the two sides overlap. Estimating the percentage of ink penetration as a ratio of the observed densities is not feasible in those areas, as the ideal density is not truly zero. Therefore, since in the background-background and foreground-foreground cases, the two observed densities are almost the same, around zero in the first case and around the maximum density in the other, small fluctuations make the value of their ratios unpredictable. Consequently, during the restoration phase, one of the two sides will have a sort of "hole" (values close to those of the background) in correspondence with the occlusion areas.

Here we propose to solve the direct problem of eq. (1) for generating the data, rather than solving the inverse problem for estimating the unknown ideal densities. Therefore, it is easier to extend the model to adequately describe the areas of occlusion, e.g., assuming that the density of the foreground text does not increase due to ink seepage. In practice, since we know the nature of each pixel this time, the density of a pixel of text on both sides can be made to saturate to the density of the original recto (original verso, respectively).

3. Neural Network: Learning and recall

We adopted a simple feedforward network with the architecture of a multilayer shallow neural network with one hidden layer and ten neurons and a backpropagation training [20]. In the specific, we used the function `patternnet` of the Matlab Deep Learning Toolbox. This net is a pattern recognition NN that can be trained to classify inputs according to target classes.

To build the training set, we select N pairs of patches containing clean text drawn from the manuscript and then symmetrically mix them using the model described in the previous section with parameters varying in the range $(0, 1)$. The patches are manually drawn, one from the recto and the other from the verso. Note, however, that they could be drawn from one side only because the net acts on a pixel-by-pixel basis so that all the mechanism is unresponsive to the character morphology and the writing style.

For each patch, the binary text is first extracted by using the Sauvola algorithm. Then, the patches are fed to the system in eq. (1) in a forward manner, with different values of the ink seepage percentage, so that we synthetically generate samples of recto-verso text with bleed-through. To account for saturation of the ink, when a pixel is foreground text in both sides, the value of the density is set to that of the recto pixel (verso pixel, respectively). For the generation of a single pair of patches, the model is taken as stationary, i.e., with fixed ink seeping percentage. However, the construction of several pairs with different percentage values

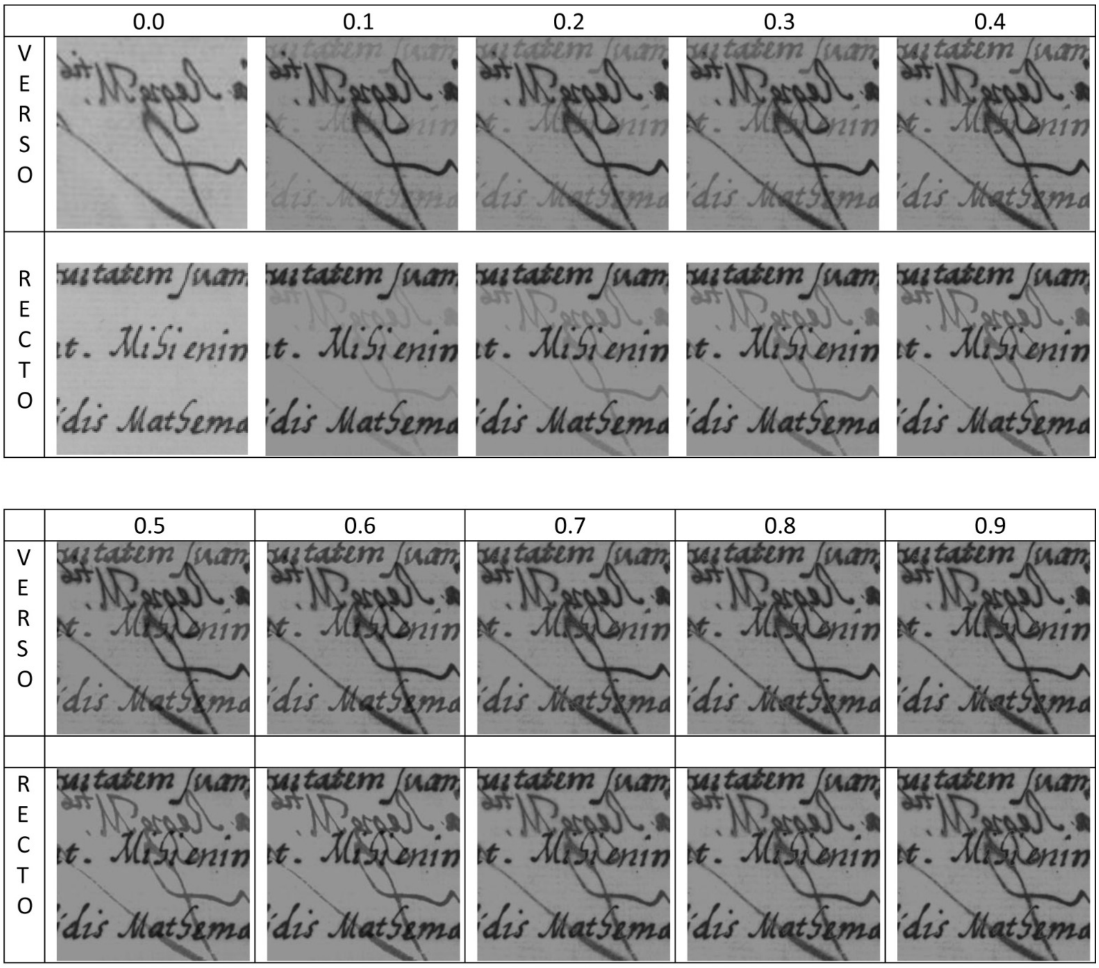

Figure 1: Example of construction of the data set based on the selection of a single pair of patches and ten values of the ink seeping percentage.

means that, as a whole, samples of non-stationary degradation will be presented to the network (Figure 1).

By construction, we know exactly the classification of each pixel of each side for these pairs of patches according to three different classes: background, foreground, and bleed-through. Thus, the ground truths of the generated samples are directly available. The data set is then randomly subdivided into the training set (the 70% of pairs) and validation set (the remaining 30%). As said, we use the Matlab `patternnet` net with a single hidden layer constituted of 10 nodes. It is possible to choose among several minimization algorithms (`training function`). Among them, we selected the scaled conjugate gradient. It is one of the most efficient for training large pattern recognition networks. We used the cross-entropy function to measure the net performance (`performance function`) during training. Indeed, Mean Square Error is the standard function used when target values are continuous. Still, the cross-entropy function

Figure 2: Application of the whole procedure: (a) original recto; (b) recto restored with NN; (c) recto restored with the procedure in [9]; (d) original verso; (e) verso restored with NN; (f) verso restored with the procedure in [9]. Original images (a) and (d): conserved at the Historical Archive of the Pontificia Università Gregoriana, APUG 529/530, c. 131r/v (Fondo Clavius).

provides better results when the targets may take discrete values - as in pattern recognition problems [20].

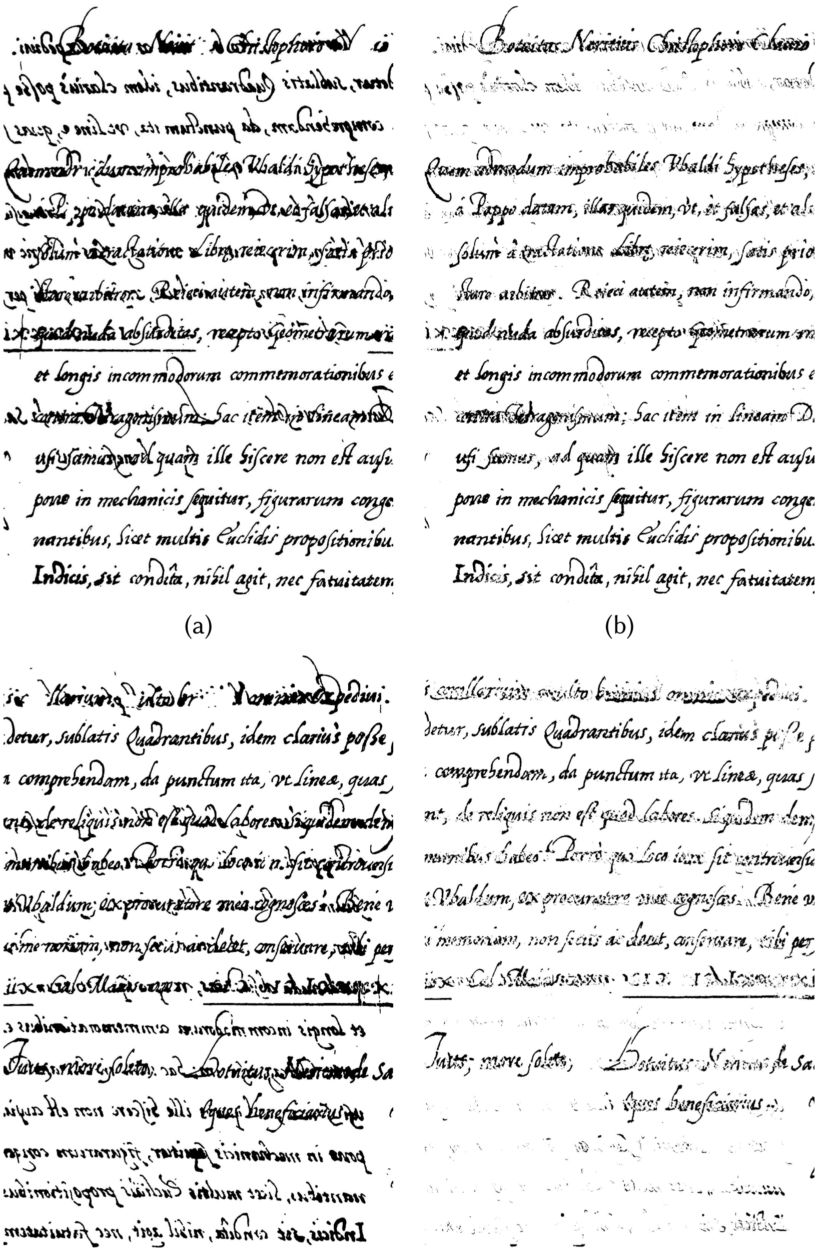

Figure 3: Binarization of the original recto-verso manuscript of Figures 2 (a) and (b): (a) recto binarized with [23]; (b) recto binarized with our NN; (c) verso binarized with [23]; (d) verso binarized with our NN.

In the experiments, the number of patches N used for constructing the data set was varying between 2 and 5, the size of the patches was chosen between 50×50 and 400×400, and the number of different values of ink seepage percentage was from 10 to 20. For the typical situation of $N = 2$, 400×400 patches, and percentage values from 0.1 to 0.9 with a step of

0.05, the execution time for constructing the net is of approximately 2 minutes.

From the output of the NN, which consists in the classification of each pixels as one of the three classes *foreground text*, *bleed-through noise* or *background*, it is immediate to obtain the binarized version of the manuscript, by merging the pixels classified as noise and background in a same class. When the goal is instead that of obtaining a virtually restored version of the manuscript, which preserves as much as possible of its original appearance and informative features, the foreground text pixels and the background pixels are given their original value, and the noisy pixels are replaced with samples drawn from the closest safe background region. To do that, in [9] we tested various state-of-the-art still image inpainting techniques and selected the best and simplest one for our purposes, the exemplar-based image inpainting technique described in [24].

4. Experimental results

In our experiments, for recto-verso registration, we used the algorithm in [19]. The manuscript is converted to grayscale for the learning phase, as the color information is unessential here for classification. Since the three RGB channels of a color manuscript share the same classes, the restored version of the color manuscript can be straightforwardly obtained.

As per virtual restoration, we make a comparison with the results of the composite practical procedure proposed in [9], which significantly improves the output of data decorrelation [25]. Concerning binarization performance, we make a comparison with the results of the segmentation method based on Laplacian energy, which was the winner of the H-DIBCO-2018 competition [2, 23, 26].

The experiments illustrated in the following were conducted in a quite challenging example, selected from the correspondence of Christoforus Clavius, conserved at the Historical Archive of the Pontificia Università Gregoriana in Rome (APUG 529/530 - Fondo Clavius). Figure 2 shows the virtual restoration of one of such letters. Figures 2 (a) and (d) show the original recto and verso, Figures 2 (b) and (e) show the results produced by our NN, and Figures 2 (c) and (f) show the results produced by the procedure described in [9]. With the NN, the results are not quantitatively perfect; however, they are correct from a qualitative point of view. The two completely overlapped texts, almost indistinguishable in the originals, have been excellently separated. Furthermore, it is apparent that the NN outperforms the procedure in [9], which is still based on a recto-verso mixing model, but stationary linear in the intensity.

In Figure 3 we compare the binarization results furnished by our NN with those obtained by the algorithm in [23]. In our results, the bleed-through pattern has been almost completely removed. On the contrary, we may observe that the method in [23], although it provides excellent results on documents with a limited amount of show-through or with other kinds of degradation, gives quite unsatisfactory results with a so heavily damaged manuscript. Nevertheless, we must point out that, to obtain the result on every single side, the NN exploits the double of information with respect to that used by the method in [23]. On the other hand, the amount of total information available and exploited by the two methods is the same to binarize the two sides. The crucial difference is that the overall information is exploited jointly in our method.

(a) (b) (c) (d)

Figure 4: Application of the same network to a different manuscript: (a) original recto; (b) recto restored; (c) original verso; (d) verso restored

The generalization capability of our net has been tested as well. Figure 4 shows the results of the application of the same NN constructed for the manuscript of Figures 2 (a) and (b) on a different recto-verso manuscript, presented in color this time.

However, our method still requires improvements. Indeed, in the binarization produced by the NN, the legibility of the extracted foreground text suffers of a sort of "corrosion" of the most compromised characters in correspondence with the occlusion areas. This is likely caused by still unsatisfactory modeling of the superposition of the recto and verso text, when they exactly overlap, and possibly by an insufficient presentation of samples in which occlusions occur or the need to include a fourth class specific for the occlusions. We plan to investigate in this respect.

5. Conclusions

We showed that by exploiting the information contained in both the recto and verso sides of an ancient manuscript affected by ink seepage, it is possible to train a very simple shallow NN to correctly classify the pixels in text, background and noise without the need of an external training set. The pairs of example-ground truth are generated from the data images themselves with the aid of a data model that describes the degradation. After classification, the output of the NN can be used to produce either a binarization of the foreground text or a virtual restoration version of the manuscript that maintains both the fullness of the informative content and the aesthetics of the original. In terms of binarization, we compare our results with those furnished by the algorithm winner of the H-DIBCO-2018 competition. The method still presents some deficiencies with respect to the correct classification of the pixels that correspond to the occlusions between the two texts. We plan to concentrate our future research on solving this residual problem, both at the data model level and the network architecture.

References

[1] A. Tonazzini, P. Savino, E. Salerno, M. Hanif, F. Debole, Virtual restoration and content analysis of ancient degraded manuscripts, International Journal of Information Science and Technology 3 (2019) 16–25. URL: http://www.innove.org/ijist/index.php/ijist/article/view/133.

[2] I. Pratikakis, K. Zagori, P. Kaddas, B. Gatos, Icfhr 2018 competition on handwritten document image binarization (h-dibco 2018), in: Proc. 16th International Conference on Frontiers in Handwriting Recognition (ICFHR), 2018, pp. 489–493. doi:https://doi.org/10.1109/ICFHR.2014.141.

[3] I. Pratikakis, K. Zagoris, X. Karagiannis, L. Tsochatzidis, T. Mondal, I. Marthot-Santaniello, Icdar 2019 competition on document image binarization (dibco 2019), in: 2019 International Conference on Document Analysis and Recognition (ICDAR), 2019, pp. 1547–1556. doi:10.1109/ICDAR.2019.00249.

[4] Y.-T. Pai, Y.-F. Chang, S.-J. Ruan, Adaptive thresholding algorithm: Efficient computation technique based on intelligent block detection for degraded document images, Pattern Recognition 43 (2010) 3177–3187. URL: https://www.sciencedirect.com/science/article/pii/S0031320310001342. doi:https://doi.org/10.1016/j.patcog.2010.03.014.

[5] F. Westphal, N. Lavesson, H. Grahn, Document image binarization using recurrent neural networks, in: 2018 13th IAPR International Workshop on Document Analysis Systems (DAS), 2018, pp. 263–268. doi:10.1109/DAS.2018.71.

[6] C. Tensmeyer, T. R. Martinez, Document image binarization with fully convolutional neural networks, 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR) 01 (2017) 99–104. doi:10.1109/ICDAR.2017.25.
[7] Q. N. Vo, S. H. Kim, H. J. Yang, G. Lee, Binarization of degraded document images based on hierarchical deep supervised network, Pattern Recognition 74 (2018) 568–586. URL: https://www.sciencedirect.com/science/article/pii/S0031320317303394. doi:https://doi.org/10.1016/j.patcog.2017.08.025.
[8] S. He, L. Schomaker, Deepotsu: Document enhancement and binarization using iterative deep learning, Pattern Recognition 91 (2019) 379–390. URL: https://www.sciencedirect.com/science/article/pii/S0031320319300330. doi:https://doi.org/10.1016/j.patcog.2019.01.025.
[9] P. Savino, A. Tonazzini, A procedure for the routinary correction of back-to-front degradations in archival manuscripts, in: N. T. Nguyen, B. H. Hoang, C. P. Huynh, D. Hwang, B. Trawiński, G. Vossen (Eds.), Computational Collective Intelligence, Springer International Publishing, Cham, 2020, pp. 838–849. doi:https://doi.org/10.1007/978-3-030-63007-2_66.
[10] A. Tonazzini, L. Bedini, E. Salerno, Independent component analysis for document restoration, IJDAR 7 (2004) 17–27. URL: http://dx.doi.org/10.1007/s10032-004-0121-8. doi:10.1007/s10032-004-0121-8.
[11] D. Fadoua, F. Le Bourgeois, H. Emptoz, Restoring ink bleed-through degraded document images using a recursive unsupervised classification technique, in: H. Bunke, A. L. Spitz (Eds.), Document Analysis Systems VII, Springer Berlin Heidelberg, Berlin, Heidelberg, 2006, pp. 38–49.
[12] B. Sun, S. Li, X.-P. Zhang, J. Sun, Blind bleed-through removal for scanned historical document image with conditional random fields, IEEE Transactions on Image Processing 25 (2016) 5702–5712. doi:10.1109/TIP.2016.2614133.
[13] R. Rowley-Brooke, F. Pitié, A. C. Kokaram, A non-parametric framework for document bleed-through removal, in: 2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, June 23-28, 2013, 2013, pp. 2954–2960. URL: http://dx.doi.org/10.1109/CVPR.2013.380. doi:10.1109/CVPR.2013.380.
[14] Y. Huang, M. S. Brown, D. Xu, User-assisted ink-bleed reduction, IEEE Transactions on Image Processing 19 (2010) 2646–2658. URL: http://dx.doi.org/10.1109/TIP.2010.2048971. doi:10.1109/TIP.2010.2048971.
[15] A. Tonazzini, L. Bedini, Restoration of recto–verso colour documents using correlated component analysis, EURASIP Journal on Advances in Signal Processing 2013 (2013). URL: https://doi.org/10.1186/1687-6180-2013-58. doi:10.1186/1687-6180-2013-58.
[16] M. Hanif, A. Tonazzini, P. Savino, E. Salerno, Non-local sparse image inpainting for document bleed-through removal, Journal of Imaging 4 (2018). URL: https://www.mdpi.com/2313-433X/4/5/68. doi:10.3390/jimaging4050068.
[17] J. Wang, C. L. Tan, Non-rigid registration and restoration of double-sided historical manuscripts, in: 2011 International Conference on Document Analysis and Recognition, ICDAR 2011, Beijing, China, September 18-21, 2011, 2011, pp. 1374–1378. URL: http://dx.doi.org/10.1109/ICDAR.2011.276. doi:10.1109/ICDAR.2011.276.
[18] R. Rowley-Brooke, F. Pitié, A. C. Kokaram, Nonrigid recto-verso registration using page

outline structure and content preserving warps, in: Proceedings of the 2nd International Workshop on Historical Document Imaging and Processing, ACM, 2013, pp. 8–13. doi:https://doi.org/10.1145/2501115.2501124.

[19] P. Savino, A. Tonazzini, Digital restoration of ancient color manuscripts from geometrically misaligned recto-verso pairs, Journal of Cultural Heritage 19 (2016) 511–521. URL: https://www.sciencedirect.com/science/article/pii/S1296207415001843. doi:https://doi.org/10.1016/j.culher.2015.11.005.

[20] H. B. Demuth, M. H. Beale, M. T. Hagan, Neural Network Design, PWS Publishing, 1996.

[21] A. Tonazzini, P. Savino, E. Salerno, A non-stationary density model to separate overlapped texts in degraded documents, Signal, Image and Video Processing 9 (2015) 155–164. URL: https://doi.org/10.1007/s11760-014-0735-3. doi:10.1007/s11760-014-0735-3.

[22] P. Savino, A. Tonazzini, L. Bedini, Bleed-through cancellation in non-rigidly misaligned recto–verso archival manuscripts based on local registration, International Journal on Document Analysis and Recognition (IJDAR) 22 (2019) 163–176. URL: https://doi.org/10.1007/s10032-019-00323-2. doi:10.1007/s10032-019-00323-2.

[23] W. Xiong, X. Jia, J. Xu, Z. Xiong, M. Liu, J. Wang, Historical document image binarization using background estimation and energy minimization, in: 2018 24th International Conference on Pattern Recognition (ICPR), 2018, pp. 3716–3721. doi:10.1109/ICPR.2018.8546099.

[24] A. Criminisi, P. Perez, K. Toyama, Region filling and object removal by exemplar-based image inpainting, IEEE Transactions on Image Processing 13 (2004) 1200–1212. doi:10.1109/TIP.2004.833105.

[25] A. Tonazzini, E. Salerno, L. Bedini, Fast correction of bleed-through distortion in grayscale documents by a blind source separation technique, IJDAR 10 (2007) 17–25. URL: http://dx.doi.org/10.1007/s10032-006-0015-z. doi:10.1007/s10032-006-0015-z.

[26] W. Xiong, L. Zhou, L. Yue, L. Li, S. Wang, An enhanced binarization framework for degraded historical document images, EURASIP Journal on Image and Video Processing 2021 (2021). URL: https://doi.org/10.1186/s13640-021-00556-4. doi:10.1186/s13640-021-00556-4.