Mathematical models and neural networks for the description and the correction of typical distortions of historical manuscripts

Pasquale Savino[1][0000−0002−8841−5440] and Anna Tonazzini[1][0000−0001−6970−4725]

Istituto di Scienza e Tecnologie dell'Informazione
Consiglio Nazionale delle Ricerche
Via G. Moruzzi 1, 56124 Pisa, Italy
{pasquale.savino,anna.tonazzini}@isti.cnr.it

Abstract. Historical manuscripts are very often degraded by the seeping or transparency of the ink from the page opposite side. Suppressing the interfering text can be of great aid to philologists and paleographers who aim at interpreting the primary text, and nowadays also for the automatic analysis of the text. We formerly proposed a data model, which approximately describes this damage, to generate an artificial training set able to teach a shallow neural network how to classify pixels in clean or corrupted. This NN has proved to be effective in classifying manuscripts where the degradation can be also widely variable. In this paper, we modify the architecture of the NN to better account for ink saturation in text overlay areas, by including a specific class for these pixels. From the experiments, the improvement of the classification and then the restoration is significant.

1 Introduction

Historical and archival manuscripts are usually damaged by a whole series of factors, primarily due to the natural degradation of the materials over time and conditions. The basic requirement for the fruition of these manuscripts is the removal of the degradations in order to make the main text fully understandable. However, this may not be sufficient for a comprehensive fruition, since these manuscripts can contain other elements, such as annotations, miniatures, watermarks, drawings, etc., that should be preserved due to their historical and informative value. Therefore, a right balance is needed between the removal of useless and harmful elements and the conservation (or enhancement) of elements which, although unrelated to the primary text, are very important for the history of the manuscript.

Degraded document binarization can be efficient in separating the main text from other patterns, which can be considered as a complex background to be removed in total [2–5,20]. However, binarization alone cannot solve cases of very strong degradation, as we will show later. Furthermore, it produces a two-class image, in black and white, which inevitably has lost interesting details of the

manuscript Thus, virtual restoration assumes the important role of preserving and highlighting the useful elements, while removing the useless ones that can disturb or even make impossible the scholar study [22].

The bleed-through degradation is the most frequent and impairing degradation in ancient manuscripts. This occurs when both sides of the paper are written. Methods specifically designed for bleed-through reduction are distinguished into blind methods, which exploit the information of the front side alone [6, 7, 16], and non-blind methods, where the often available two sides of the manuscript page are jointly exploited [8–10, 17, 28]. The non-blind methods can provide a very fine virtual restoration, with the counterpart that they require a perfect alignment of the two images [12–14].

In [29] we proposed a simple multilayer shallow neural networks with back-propagation training [23], to solve the non-blind case. We implemented the NN in such a way that it auto-adapts to the manuscript to be restored, i.e. it does not require a preliminary learning from many other manuscripts already classified. This can be realized, for example, when an existing data model can be used for generating simulated training samples. In our case, we experimented with a previously proposed data model, approximately describing the degradation affecting recto-verso manuscripts [11]. A training set was built starting from ground-truths drawn from the clean zones of the manuscript at hand, and then mixed accordingly to the model. The experimental results presented in [29] on heavily damaged manuscripts seemed encouraging in terms of degradation cancellation. We accounted for variable degradation, also very strong. This makes our NN, built on the basis of a single exemplar manuscript, to be potentially effective on other manuscripts of the same corpus but with degradation of different entity, or different pages of a same book.

The difficulty with very strong levels of bleed-through is to succeed in distinguish them from the situations in which the primary text and the opposite text overlap (we call the occlusions). In particular, a NN trained to recognize bleed-through of levels very similar to that of the main text could produce random responses in those cases.

In this paper we focus on modifications to the network architecture and learning with the aim to try to get rid of this great difficulty. With respect to the network architecture, we introduce an extra output class to classify the text overlapping pixels as occlusions rather than as mere foreground text pixels. As regards the construction of the training set, we assume a data model that explicitly includes the conditions for the occlusions to occur.

The paper is organized as follows. In Section 2 we describe the method adopted for the construction of the adaptive training set, using a specific data model. Section 3 provides some operative details about the shallow NN architecture and the learning and recall phases. Section 4 analyzes from a qualitative point of view some preliminary results, both synthetic and real, in comparison to state-of-the-art binarization methods for degraded historical manuscripts. Finally, Section 5 concludes the paper.

2 Construction of the training set

The first step of the virtual restoration process for historical recto-verso manuscripts described in this paper is to classify the pixels of each side into four different classes that we call *foreground*, *background*, *bleed-through*, and *occlusion*, respectively. These classes represent the main text, the clean paper texture with, eventually, other marks, the seeping ink and the areas where the two sides are both written and the two texts overlap. In the previous work [29] we considered three classes only, by merging the occlusion pixels with the text pixels. This reflects the appearance of only one side of the paper, as occlusions do are text and, without knowledge of the opposite side, cannot be identified with certainty. However, as we will see in the experimental results, using three classes only resulted in an overestimation of the bleed-through class.

As a classifier, we use a neural network (NN) that needs a training set with ground truths to learn how to discriminate the pixels. As mentioned, we do not use an external dataset based on similar manuscripts already classified, but our NN is trained using the same manuscript we want to classify.

Thus, to build the training set, we select N pairs of patches from the manuscript containing clean text, and then symmetrically mix them using a data model for seeping ink that describes the observed optical density of each side as the weighted sum of the ideal densities of the two sides. Defining the optical density as $D_s(t) = -log\left(\frac{s(t)}{p}\right)$, at pixel t, with $s(t)$ being the intensity, and p the mean value of the paper support, the model is expressed in the following way:

$$D_x^{obs}(t) = \begin{cases} D_x(t), & \text{if } t \text{ is text in both sides} \\ D_x(t) + q_y(t) D_{h_y \otimes s_y}(t), & \text{elsewhere} \end{cases} \quad (1)$$

where x and y indicate the two sides, which must be perfectly aligned after reflection of one of the two. Eq. (1) holds for the opposite side by exchanging the role of x and y. In eq. (1), D^{obs} and D are the observed and the ideal optical density, 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_x and q_y, whose maximum allowed range is $[0, 1]$, have the physical meaning of ink penetration percentages from one side to the other. The first condition of the model eq. (1) means that we assume that the density of the foreground text does not increase due to ink seepage, just as it happens in the majority of the cases.

In previous works [11,14,15], we neglected the ink saturation effect, and proposed to invert the equation in the second condition of the model for virtually restoring the recto-verso pair. To make the inversion possible, we assumed that the hyperparameters q and h are known in advance. Based on the observed densities of the two sides, we first inverted the model by assuming an identically zero ideal density in the opposite side, thus obtaining estimates of the ink penetration percentages at each pixel. The system can then be solved with respect to the ideal density maps, from which the virtually restored manuscript sides

are obtained. To manage the text superposition areas (whose ideal density is not zero), the obtained images were corrected using some technicalities.

Here we propose to solve the direct problem of eq. (1) for generating the data necessary for the training set, rather than solving the inverse problem for estimating the ideal densities, which are known in this case.

Operatively, each patch out of the selected N pairs containing clean text is first binarized by the Sauvola algorithm, in order to extract the map of the clean text and the map of the background. Comparing the binary map of both members of the pair allows for locate the four classes in each side, including the occlusions. Then, as said, the original, non-binary pairs of 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. The first condition in eq. (1) permits to simulate the saturation of the ink, that is, 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 means that, as a whole, samples of non-stationary degradation will be presented to the network.

3 Neural network: architecture, learning and recall

We adopted a simple feedforward network with the architecture of a multilayer shallow neural networks with one hidden layer and ten neurons, and a backpropagation training [23]. 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.

The network processes the two sides of the manuscript simultaneously, on a pixel-by-pixel basis. For each pixel, we consider as features the two density values in the two sides. As already mentioned, as target classes we consider the four different classes of background, foreground, bleed-through and occlusion.

By construction, for the pair of patches used for building the training set we exactly know the classification of each pixel of each side. Thus, the target classes of the generated samples are directly available. The data set is then randomly subdivided into training set (the 70% of pairs) and validation set (the remaining 30%). As mentioned, we use the Matlab `patternnet` net with a single hidden layer constituted of 10 nodes. As minimization algorithm (`training function`) we chose the scaled conjugate gradient, and the cross entropy for measuring the net performance (`performance function`) during training. Tests performed with a higher number of neurons did not provide significant improvement in the quality of the results.

In the experiments, the number of patches N used for constructing the data set was varying between 2 and 10, 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. The architectural simplicity of the network guarantees very

short learning times. Typical learning times are of the order of a few seconds if the indicated parameters are used.

From the output of the NN, which consists in the classification of each pixel as one of the four classes, it is immediate to obtain the binarized version of the manuscript, by merging the pixels classified as text and occlusion in a same class, and, similarly, bleed-through noise and background in another single class. When the goal is instead that of obtaining a virtually restored version of the manuscript, which preserves as much as possible its original appearance and informative features, the foreground text pixels, the occlusion pixels and the background pixels are given their original value, whereas the noisy pixels are replaced with samples drawn from the closest safe background region. For this latter task, in [21] we tested various state-of-the art still image inpainting techniques, and selected as the best and simplest one for our purposes the exemplar-based image inpainting technique described in [19].

4 Experimental results

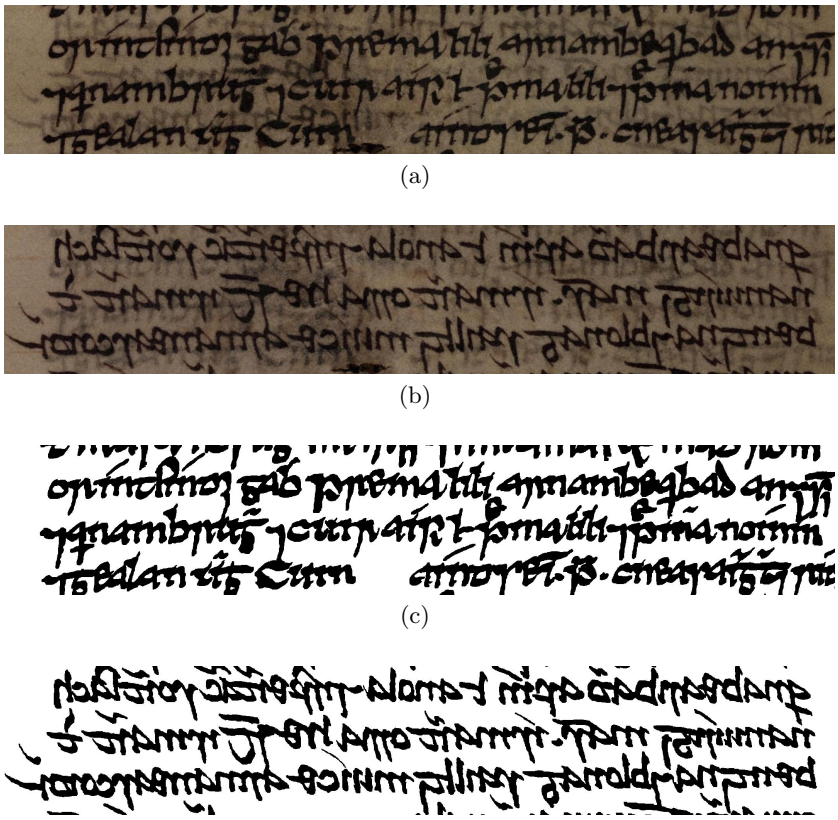
We evaluate the results of our virtual restoration method from a qualitative point of view, and comparing the NN classification result with the binarization produced by the algorithm that was the winner of the H-DIBCO-2018 competition [1]. This algorithm implements a segmentation method based on a Laplacian energy, and is described in [24, 25].

Both for the learning and classification phases, the manuscripts are converted to grayscale, as the color information is unessential here for the purpose of classification. For virtual restoration, the restored versions of the color manuscripts can be straightforwardly recovered from the classification of the grayscale versions, since the three RGB channels share the same classes.

A first experiment was totally synthetic, in the sense that the recto-verso images to restore were numerically built based on the ground-truths available for one of the recto-verso pairs contained in the database [26, 27], i.e. the 15-th pair out of a total of 25 pairs. In a second experiment, we restored the same recto-verso pair used for the synthetic case, this time as it appears in the database, i.e. with its real degradation. We processed the couple of RGB images with the NN trained on them. Since the recto-verso pair was already registered, we did not include the block alignment mechanism necessary in the case of misaligned recto-verso pairs, and described in [14, 29].

Figures 1 (a) and (b) show the recto and the reflected verso of the chosen pair, with the real degradation that affect them. Figures 1 (c) and (d) show the binary ground-truths, manually built, that accompany that recto-verso pair in the database. This ground-truths represent the correct foreground texts of the two manuscript sides, and serve as comparison to evaluate the performance of algorithms of binarization of degraded historically manuscripts, as well as, indirectly, of algorithms of virtual restoration.

In the synthetic experiments we built an artificial clean recto-verso pair by placing the clean foreground texts on a textured background obtained by in-

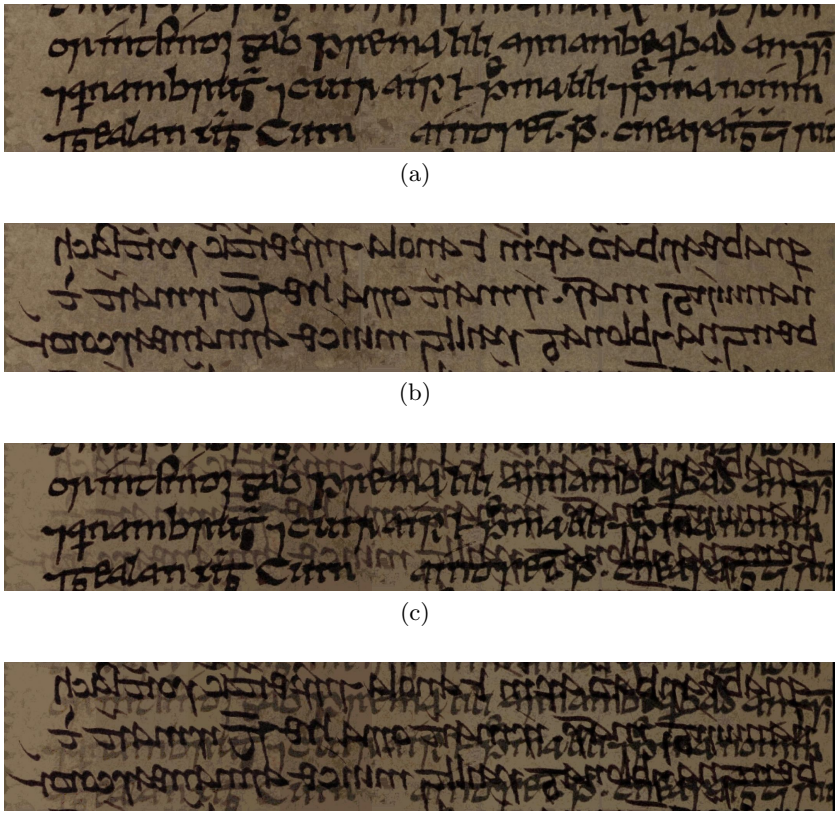


(a) (b) (c) (d)

Fig. 1. The manuscripts used for the experiments: (a) and (b) real recto and reflected verso of the 15-th pair of manuscripts in the database [27]; (c) and (d) their corresponding manually generated binary ground-truths.

painting. The foreground texts were obtained by picking up the RGB values of the real degraded images of Figures 1 (a) and (b), at the positions of the black pixels in the corresponding binary ground-truth maps (Figures 1 (c) and (d)).

Figure 2 (a) and (b) show this clean, ideal manuscript pair. An artificially degraded pair has then be obtained by mixing the ideal one through the data model of eq. 1, where the percentage of penetrating ink has been increased from 0.1 to 0.9 (left to right) (Figures 2 (c) and (d)).



(a)

(b)

(c)

(d)

Fig. 2. Generation of a synthetic manuscript pair: (a) and (b) clean, ideal recto and verso created by the images in Figure 1; (c) and (d) degraded recto and verso numerically constructed by feeding the images (a) and (b) to the data model of eq. (1).

Figure 3 shows the results of applying the NN (training and recall) on those images. The training set was constructed by selecting pairs of clean patches from the degraded images themselves, and were mixed with percentages of ink penetration spanning from 0.1 to 0.9, in such a way to cover all the range of different amounts of degradation in the data. We built two different networks,

one having as output only three classes (foreground, background and bleed-through), and the other characterized by the fourth class of the occlusions.

Figures 3 (a) and (b) show the virtually restored verso with the corresponding binary image when the number of classes of the NN was set to 3. Note how the reconstructed text sometimes appears corroded, fragmentary, with missing strokes. As already mentioned in the introduction, the fact that here the degradation is very strong, reaching up to $q = 0.9$, the lack of a specific class for the occlusions causes text pixels on both sides to be attributed to the bleed-through class and then deleted.

Figures 3 (c) and (d) show the virtually restored verso with the corresponding binary image when the number of classes of the NN was set to 4. In this case the text is reconstructed much better, the characters are complete and full. Conversely, the bleed-through cleanup is slightly less effective, especially in the more severely degraded right-hand side of the manuscript.

Finally, Figure 3 (e) shows the binarization of the degraded verso of Figure 2 (d) with the algorithm in [24], which was the the winner of the H-DIBCO-2018 competition [1]. This algorithm works using only the information of the side to be processed. Clearly, the extreme degradation of the manuscript makes it impossible to discriminate noise from the text of interest without information contributed by the opposite side of the page.

In the real experiment, we compared the performance of the 3-class NN and the 4-class NN on the degraded real images of Figures 1 (a) and 1 (b). At present, the training phase requires a rough estimate of the maximum amount of degradation within the manuscript, in order to make the NN work best. An estimate of the parameters q can be done as in [14].

Thus, in this cases the training set was constructed by limiting the maximum value of the ink penetration percentage to 0.5, as the degradation is not as extreme as in the synthetic case. Our results still demonstrate clear superiority of the 4-classes NN with respect to the 3-classes NN, as shown for the verso side in Figure 4. Indeed, again, the 3-class network is able to recognize the bleed-through pixels, so that most of them can be removed. However, because the learning phase has associated pixels that are text on both sides with the foreground class rather than with the specific occlusion class, some ambiguity remains between the foreground class and the bleed-through class.

With respect to the binary version of the restored manuscript (verso side shown in Figure 4 (d)), this time the one obtained through the algorithm in [24] ((verso side shown in Figure 4 (e)) is slightly cleaner. Note however that it also presents big local defects, such as the lack of entire characters and the excessive thickness of others, for example in the area highlighted with the red box. This area is shown enlarged in Figure 5 for both methods, in comparison with the binary ground-truth provided in the public database.

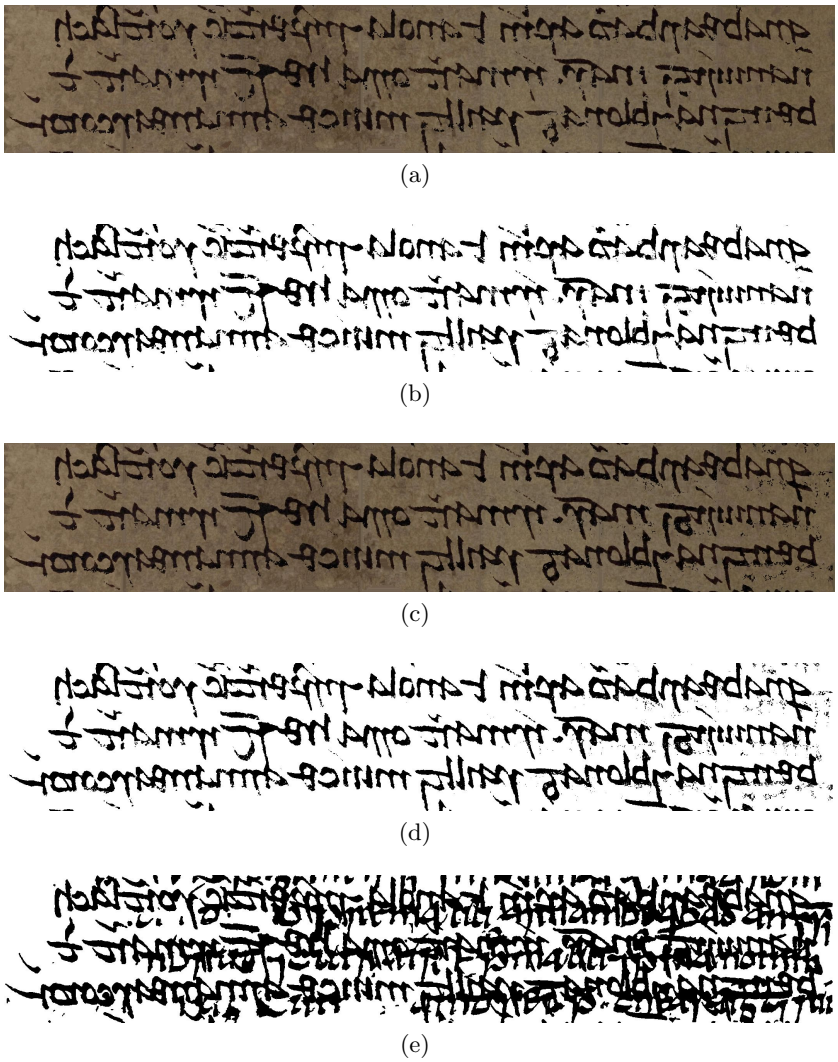


Fig. 3. Virtual restoration of the synthetic pair shown in Figures 2 (c) and 2 (d): (a) and (b) verso restored with the 3-classes NN and the corresponding binary version; (c) and (d) verso restored with the 4-classes NN and the corresponding binary version; (e) degraded verso of Figure 2 (d) binarized with the algorithm in [24].

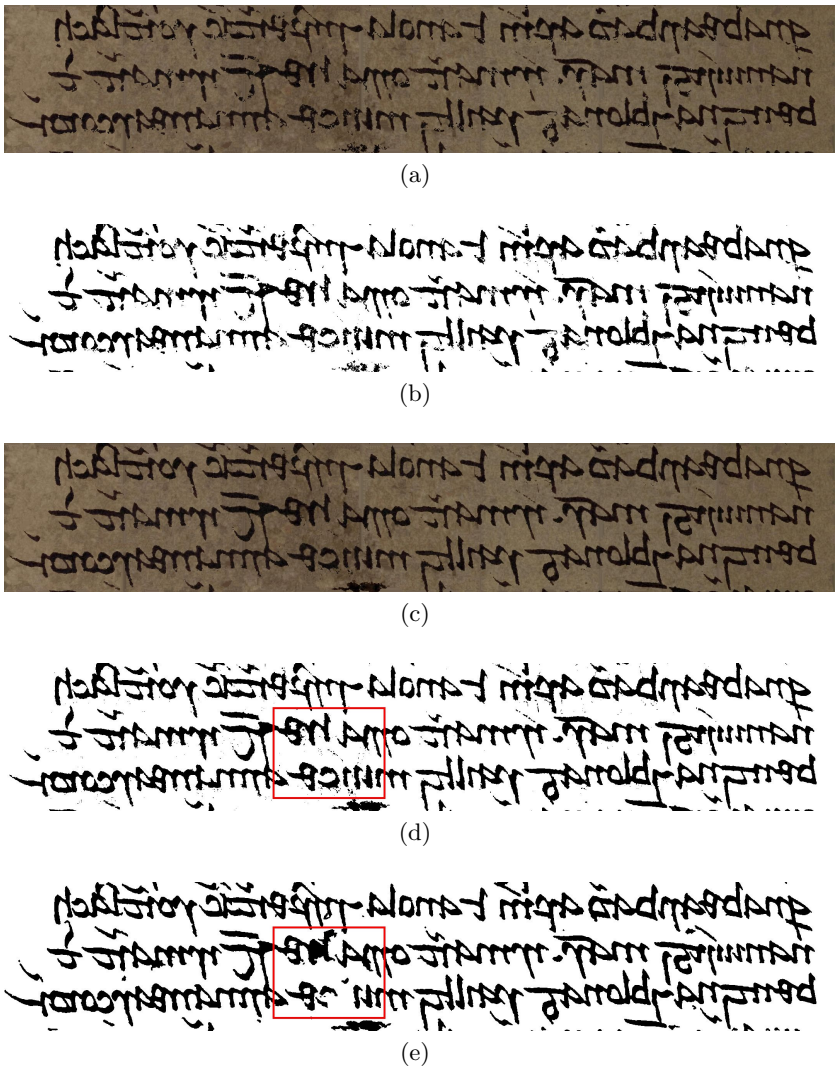
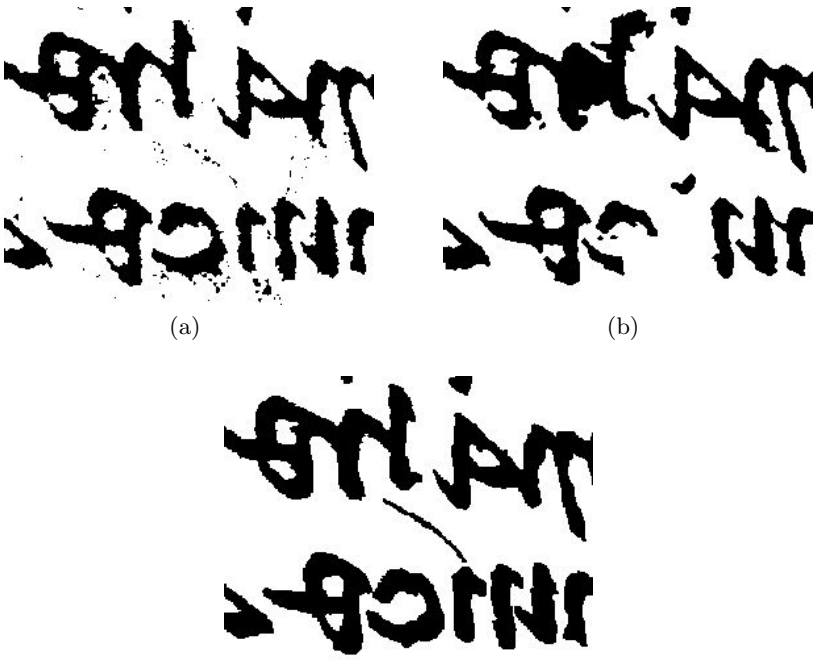


Fig. 4. Virtual restoration of the real pair shown in Figures 1 (a) and 1 (b): (a) and (b) verso restored with the 3-classes NN and the corresponding binary version; (c) and (d) verso restored with the 4-classes NN and the corresponding binary version; (e) binarization of the verso of Figure 1 (b) with the algorithm in [24].



(a) (b)

(c)

Fig. 5. Enlarged detail of the binary maps highlighted in red in Figure 4: (a) our result; (b) the H-DIBCO-18 result [24]; (c) the binary ground-truth provided in the public database [27].

5 Conclusions

We have shown that, exploiting the information contained on both the recto and verso of an ancient manuscript affected by ink penetration, it is possible to train a very simple shallow NN to correctly classify pixels in primary text, paper background, bleed-through noise and overlaid texts, without the need for an external training set. The example-target class pairs are generated from the data images themselves with the help of a data model that describes the degradation. After classification, the output of the NN can be used to produce a binarization of the foreground text or a virtual restoration version of the manuscript that maintains both the fullness of information content and the aesthetics of the original. The method improves on our previous proposals regarding the correct classification of the pixels corresponding to the occlusions between the two texts. In terms of binarization, we compare our results with those provided by the winning algorithm of the H-DIBCO-2018 [1] competition. The superiority of our method is evident in a synthetic case constructed in such a way as to cover the extent of degradation from almost zero to the maximum allowed. For moderate, rather uniform, real degradation, the binarization method performs slightly better, albeit with large local errors. Since the data model used is independent of the neural network paradigm, we intend to test our approach with other more sophisticated neural networks. We will also try to resolve the residual ambiguity between the two classes bleed-through and occlusion by using more descriptors.

References

1. I. Pratikakis, K. Zagori, P. Kaddas, and 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.
2. Y. Pai, Y. Chang, and S. Ruan, "Adaptive thresholding algorithm: Efficient computation technique based on intelligent block detection for degraded document images," *Pattern Recognition*, vol. 43, p. 3177–3187, 2010.
3. F. Westphal, N. Lavesson, and H. Grahn, "Document image binarization using recurrent neural networks," in *13th IAPR Int. Workshop on Document Analysis Systems (DAS2018), Proceedings*, 2018, p. 263–268.
4. R. Tensmeyer and T. Martinez, "Document image binarization with fully convolutional neural networks," in *14th IAPR Int. Conf. on Document Analysis and Recognition (ICDAR 2017), Proceedings*, 2017, pp. 99–104.
5. Q. Vo, S. Kim, H. Yang, and G. Lee, "Binarization of degraded document images based on hierarchical deep supervised network," *Pattern Recognition*, vol. 74, p. 568–586, 2018.
6. D. Fadoua, F. L. Bourgeois, and H. Emptoz, "Restoring ink bleed-through degraded document images using a recursive unsupervised classification technique," *Document Analysis Systems VII, Lecture Notes in Computer Science, vol. 3872. Springer*, pp. 27–38, 2006.
7. B. Sun, S. Li, X. P. Zhang, and J. Sun, "Blind bleed-through removal for scanned historical document image with conditional random fields," *IEEE Trans. Image Process.*, pp. 5702–5712, 2016.

8. R. Rowley-Brooke, F. Pitié, and A. Kokaram, "A non-parametric framework for document bleed-through removal," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2013, pp. 2954–2960.
9. Y. Huang, M. S. Brown, and D. Xu, "User assisted ink-bleed reduction," *IEEE Transactions on Image Processing*, vol. 19, no. 10, pp. 2646–2658, 2010.
10. M. Hanif, A. Tonazzini, P. Savino, and E. Salerno, "Non-local sparse image inpaintig for document bleed-through removal," *Journal of Imaging*, vol. 4, p. 68, 2018.
11. A. Tonazzini, P. Savino, and E. Salerno, "A non-stationary density model to separate overlapped texts in degraded documents," *Signal, Image and Video Processing*, vol. 9, pp. 155–164, 2015.
12. R. Rowley-Brooke, F. Pitié, and A. C. Kokaram, "Non-rigid recto-verso registration using page outline structure and content preserving warps," in *2nd International Workshop on Historical Document Imaging and Processing, proceedings*, 2013, pp. 8–13.
13. J. Wang and C. L. Tan, "Non-rigid registration and restoration of double-sided historical manuscripts," in *Proc. Int. Conf. on Document Analysis and Recognition (ICDAR)*, 2011, p. 1374–1378.
14. P. Savino and A. Tonazzini, "Digital restoration of ancient color manuscripts from geometrically misaligned recto-verso pairs," *Journal of Cultural Heritage*, vol. 19, pp. 511–521, 2016.
15. P. Savino, A. Tonazzini, and L. Bedini, "Bleed-through cancellation in non-rigidly misaligned recto-verso archival manuscripts based on local registration," *Int J. on Document Analysis and Recognition*, vol. 22, p. 163–176, 2019.
16. A. Tonazzini, L. Bedini, and E. Salerno, "Independent component analysis for document restoration," *Int. Journal on Document Analysis and Recognition*, vol. 7, pp. 17–27, 2004.
17. A. Tonazzini and L. Bedini, "Restoration of recto-verso colour documents using correlated component analysis," *EURASIP Journal on Advances in Signal Processing*, p. 2013:58, 2013.
18. A. Tonazzini, E. Salerno, and L. Bedini, "Fast correction of bleed-through distortion in grayscale documents by a blind source separation technique," *Int. Journal on Document Analysis and Recognition*, vol. 10, pp. 17–25, June 2007.
19. A. Criminisi, P. Pérez, and K. Toyama, "Region filling and object removal by exemplar-based image inpainting," *IEEE Trans. on Image Processing*, vol. 13, pp. 1200–1212, 2004.
20. S. He, and L. Schomaker, " DeepOtsu: Document Enhancement and Binarization using Iterative Deep Learning," *Pattern Recognition*, vol. 9, pp 379–390, 2019.
21. P. Savino, A. Tonazzini, "A procedure for the routinary correction of back-to-front degradations in archival manuscripts", Proc. IWCIM 2020, ICCCI 2020, in: Computational Collective Intelligence, N.T. Nguyen, B.H. Hoang, C.P. Huynh, D. Hwang, B. Trawinski and G. Vossen Eds., pp. 838-849, Springer, 2020.
22. A. Tonazzini, P. Savino, E. Salerno, M. Hanif, and F. Debole, "Virtual restoration and content analysis of ancient degraded manuscripts", *Int. J. of Information Science and Technology (iJIST)*, vol. 3, pp. 16-25, 2019
23. Hagan, M.T., H.B. Demuth, and M.H. Beale, Neural Network Design, Boston, MA: PWS Publishing, 1996.
24. W. Xiong, X. Jia, J. Xu, Z. Xiong, M. Liu, J. Wang, "Historical document image binarization using background estimation and energy minimization," in Proc. 24th International Conference on Pattern Recognition (ICPR 2018), Beijing, CHINA, 2018, pp. 3716-3721.

25. W. Xiong, L. Zhou, L. Yue, L. Li and S. Wang, "An enhanced binarization framework for degraded historical document images," EURASIP Journal on Image and Video Processing, Vol. 2021, 2021.
26. R. Rowley-Brooke, F. Pitié and A. C. Kokaram, A ground truth bleed-through document image database, in P. Zaphiris, eds. G. Buchanan, E. Rasmussen, and F. Loizides, Theory and Practice of Digital Libraries, Lecture Notes in Computer Science **7489** (2012) 185–196.
27. Irish Script On Screen Project (2012), www.isos.dias.ie.
28. M. Hanif, A. Tonazzini, Syed Fawad Hussain, Usman Habib, E. Salerno, P. Savino and Zahid Halim, "Blind Bleed-through Removal in Color Ancient Manuscripts," Multimedia Tools and Applications, published online 27 September 2022, https://doi.org/10.1007/s11042-022-13755-6
29. P. Savino, A.Tonazzini, "A shallow neural net with model-based learning for the virtual restoration of recto-verso manuscripts," 1st Int. Virtual Conference on Visual Pattern Extraction and Recognition for Cultural Heritage Understanding VIPERC 2022, https://ceur-ws.org/Vol-3266/paper3.pdf, 2022