Joint NN elaboration of the two sides of damaged historical documents for improving text analysis

1st Pasquale Savino
Istituto di Scienza e Tecnologie dell'Informazione
Consiglio Nazionale delle Ricerche
Pisa, Italy
pasquale.savino@isti.cnr.it
0000-0002-8841-5440

2nd Anna Tonazzini
Istituto di Scienza e Tecnologie dell'Informazione
Consiglio Nazionale delle Ricerche
Pisa, Italy
anna.tonazzini@isti.cnr.it
0000-0001-6970-4725

***Abstract*—The quality of the text analysis and transcription of historical documents, performed either automatically or manually, benefits from enhancing their digital versions, especially when the originals are ancient or degraded. Among the many and varied damages, the penetration or transparency of ink from one page to the other of the folio (see-through) is one of the most frequent and invasive degradation. Here, we focus on printed or handwritten documents affected by this degradation, and propose an enhancement technique based on pixel classification through neural networks. Our technique needs the information from both sides of the folio, but this request is commonly satisfied in the modern archives.**

We exploit a shallow neural network in order to classify into clean or corrupted all the pixels of the two sides, analyzed by pairs. Using a NN entails the availability of a suitable set of examples to be used for its training. We previously proposed a data model that roughly describes the see-through degradation in the two sides of a grayscale or color document page. In this paper, we use this model to generate an artificial training set that makes the trained network capable of generalizing to different levels of degradation. The result of pixel classification can return a binary map of the document, cleaned from the interferences, or can serve as the basis for its virtual restoration, by substituting the noisy pixels with samples of the paper support.

We test the performance of the proposed enhancement technique for improving various steps of text analysis and transcription of historical documents.

In case of printed documents, the binary map can be directly input to an OCR algorithm. We demonstrate that the joint classification of the two sides of the document can significantly improve binarization, and then character recognition, compared to the separate processing of the single sides, using the same amount of information. In case of handwritten texts, we input the virtually restored image to a popular software for text management and analysis of historical manuscripts. We show that also in this case an encouraging gain in the performance of the various tasks ca be achieved.

***Index Terms*—Historical document text analysis; degraded document binarization; optical character recognition; recto-verso documents; shallow multilayer neural networks.**

I. Introduction

Conservation and accessibility of the documentary material that preserves our memory have always been put at risk by various damages occurring over time. To guarantee the conservation of these fragile artefacts and encourage their use, it is essential to create detailed digital archives of their images, acquired in the greatest possible number of modalities. In addition to preservation and fruition, digitization also allows the non-invasive use of image processing techniques. These techniques may improve readability, repair degradation and damage, and automatically analyze writings, in order to assist scholars during transcription and critical analysis.

Although fully automatic text transcription is not yet accurate enough, many of the single tools that it requires, such as layout analysis, text segmentation, word spotting, optical character recognition (OCR) and text normalization, have achieved high levels of reliability and have proven to be very powerful instrumentsas an aid to traditional transcription. In particular, recent advances in deep learning techniques, such as convolutional neural networks (CNNs) or recurrent neural networks (RNNs), have led to significant improvements in these techniques and in the recognition accuracy of even degraded and variant texts [2]–[5], [14].

In this paper we consider digital documents whose originals have been damaged by ink seeping from one side of the paper to the other (bleed-through effect), or documents originally undegraded where the text of the opposite side is visible in the digital versions as a result of the scanning process (show-through effect). For simplicity, we call both damages "see-through effect". See-through is the most frequent and impairing degradation in ancient documents and manuscripts, and results in a complex background that must be removed for the correct analysis of the text of interest.

Binarization of the individual sides seems to be the elective pre-processing tool to separate the main text from the noisy background, also considering that many text analysis tools require binary maps as input. Unfortunately, binarization cannot solve cases of very strong interferences. Methods exploiting the information of both sides can provide more accurate suppression of even severe see-through, compared to single-side binarization [6], [7], [12], but require a perfect alignment of the two images.

We assume here that the front and back images of the degraded documents are both available in digital form, and that they are perfectly registered, since we have the instruments to do that [9]–[11]. Our goal is to show how a joint pre-processing of the two images can greatly improve text analysis

tasks, the OCR and the transcription.

In particular, we propose a simple multilayer shallow neural network with backpropagation training [15] to jointly classify the pixels of the two sides as text or noise/background. In this way, the two texts in the two sides can be more accurately identified than when a each side is processed at a time.

We implement the NN in such a way that it auto-adapts to the document at hand without requiring preliminary learning from large collections of other similar documents previously classified. Indeed, we generate simulated training samples directly from patches of the document at hand, through a data model that approximately describes the degradation affecting recto-verso documents [8].

We initially proposed this classification approach in [21], for obtaining the virtual restoration of historical manuscripts, even those heavily damaged by strong bleed-through, with satisfactory results. The goal of virtual restoration is to suppress the degradation while conserving all the natural and useful features of the original manuscript in such a way as to recover its original appearance. This can be obtained by substituting the pixels classified as noise with samples of the background paper.

The result of classification can also be used for an immediate binarization of the image for text of interest and complex background separation.

These binarization aspects of our method and their implication in improving subsequent text analysis tasks, were already emphasized in the conference paper [20]. We highlighted how the joint binarization of both sides of the page is obviously superior to single-side binarization, since it exploits all the natural available information, whereas the single-side modality neglects the information from the opposite side.

We chose OCR of printed texts to evaluate how and how much the improvement in the binarization quality reflects on the improvement of the text analysis process. We performed numerical tests on pairs of recto-verso images of printed documents built artificially with increasing levels of see-through interference. The binarization results obtained with our method were first qualitatively compared with those of the algorithm that was the winner of the HDIBCO 2018 competition [1]. Then, a quantitative measure was obtained by applying the `OCR` function of MathWorks, based on the technique described in [18], to the images binarized with both the two methods, and by computing the respective rates of character and word recognition.

In the present paper, we go in deeper details with respect to the advantages of joint binarization of printed, recto-verso documents, but consider also the case of historical manuscripts, naturally affected by the same kind of degradation. We then include new experiments, where our NN classification of the paired pixels faces two further main problems: the high space-variance of this kind of degradation within the same page, and the specificity of cursive handwriting, which is hardly categorizable. Since the design of general-purpose algorithms for handwritten text recognition remains an open problem, we evaluate the quality of our pre-processing technique against a broader range of text analysis tools besides transcription, that is layout analysis and word spotting.

Also in case of really degraded handwritten texts, the variability of the degradation can be successfully managed by our NN approach through the parametric nature of the degradation model used for building the training set.

To deal with transcription of manuscripts we adopted the `Transkribus` open access software [22] [23], specifically designed for handwritten ancient texts. The `Transkribus` instruments operates text analysis at various levels, and in particular enables the manual or automatic analysis of the manuscript layout and structure, in terms of text regions, text lines and even individual words. This preliminary analysis is used as the basis for text recognition, which is performed by exploiting HTR (Handwritten Text Recognition) models.

A HTR model is created by means of neural networks from manual transcriptions of parts of the documents provided by the user. HTRs relating to different historical periods and writing styles can be shared online and used by new users. They can also be adapted to different printed graphics.

We apply `Transkribus` to the color versions of both the original and the virtually restored manuscripts in a blind way, i.e. letting the system to automatically perform all the preliminary tasks that are necessary prior automatic transcription. These preliminary operations alone can hold as a valid help to scholars that tackle manual text transcription and interpretation.

The paper is organized as follows. In Section 2 we describe the general method, i.e. the NN model with its learning and classification phases, and how binary maps or virtually restored versions of the documents can be generated from the classification. Section 3 describes the experiments of optical character recognition on a printed document, artificially degraded with different levels of corruption. The results are analyzed both qualitatively and quantitatively, comparing the OCR of the binary form of the enhanced document with the OCR of the original degraded document binarized with state-of-the-art algorithms. Section 4 discusses the automatic text analysis and transcription of virtually restored real manuscripts of historical interest. Finally, Section 5 concludes the paper.

II. Joint recto-verso NN classification, binarization and virtual restoration

The basic step of our pre-processing technique, finalized to improve the results of text analysis tools on degraded recto-verso documents, is to jointly classify the pixels of both sides into four different classes that we call *foreground*, *background*, *see-through*, and *occlusion*, respectively. These classes represent the main text, the paper texture, the seeping ink, and the areas where the two sides are both written so that the two texts overlap.

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 documents already classified, but our NN is trained using the document itself that we want to classify.

To build the training set, we select from the document, no matter from which side, N pairs of patches containing clean text, and then symmetrically mix them using a data model that describes the observed optical density of each side as the weighted sum of the ideal densities of the two sides [8], [11], [12].

Figure 1 shows how the training samples and their corresponding ground-truths are generated.

Operatively, given a pair of clean patches, the two patches are first binarized, e.g. by the Sauvola algorithm [13], to extract the maps of the clean text and the maps of the real background for each of them. The comparison of the binary maps of both patches allows for locating the four classes that will appear in each side of the future degraded patches. These degraded patches are built by feeding into the model the non-binary pairs and several percentage of ink seepage. Then, we synthetically generate samples of recto-verso text with see-through of different intensity (see Figure 1, where only a single degraded pair is shown). To simulate the saturation of the ink in the occlusion areas, that is, when a pixel is foreground text in both sides, the value of the density is set to the original value of the observed pixel.

We adopted a simple feedforward network with the architecture of a multilayer shallow neural network with back-propagation training [15]. Specifically, we used the function `patternnet` of the Matlab Deep Learning Toolbox. This network is a pattern recognition NN that can be trained to classify inputs according to target classes.

The network processes the two sides of the document 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, see-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 dataset is then randomly subdivided into training set (the 70% of pairs) and validation set (the remaining 30%). The Matlab `patternnet` net is used 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, and the size of the patches was chosen between 50×50 and 400×400. To make the network learn a degradation that can be variable within the page, we set the number of different values of ink seepage intensity from 10 to 20, ranging from the allowed minimum and maximum value, i.e. 0 and 1. The architectural simplicity of the network guarantees very short learning times, of the order of 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 document, by merging the pixels classified as text and occlusion in a same class, and, similarly, see-through noise and background in another single class. The binary map can be used as a pre-processed version of the document useful for successive tasks such as word spotting, layout analysis or text transcription.

A virtual restoration of the document can be instead obtained from classification by replacing the see-through pixels with samples drawn from the identified background class. We explain this process in details in the work [21]. We will show later on that the virtual restored image of a manuscript can be directly fed to systems for handwritten text transcription.

III. Experimental results: numerical evaluation on artificially built data

Here we consider printed texts, so that text transcription can be performed by OCR algorithms applied on the binary maps. We used the `OCR` function of MathWorks, based on the technique described in [18] [24]. Specifically, the `OCR` function performs text recognition at the character level, according to the algorithm used in the Tesseract Open Source OCR Engine [25]. The input image can be graylevel or RGB, but the algorithm first binarizes it according to the Otsu method. In this paper, we provide to the `OCR` function the image already binarized according to our method or the method in [16], which are both expected to provide better binarizations.

We evaluated our method on artificially generated data, although based on the digitized version of a real ancient printed document. The considered document comprises the two front and back pages of a same folio of an old edition of the book "Le Opere di Galileo Galilei" (Società Editrice Fiorentina, 1855). We consider this document free from see-through artefacts, although it presents other defects such as faded characters that complicate the OCR process themselves. We scanned the two pages and numerically added them a see-though pattern according to the data model in [11].

The binarization of the degraded images produced by our NN by jointly processing the two sides is evaluated visually, and compared with the binary map obtained with an algorithm based on the processing of a single side. We chose the algorithm that was the winner of the H-DIBCO-2018 competition [1]. This algorithm implements a segmentation method based on a Laplacian energy, as described in [16], [17].

To evaluate the results of the OCR applied to the binary maps described above, we measured a recognition rate with respect to the true transcription, done manually.

For both the learning and classification phases, the documents are converted to grayscale, as the color information is unessential here for the purpose of classification.

Figure 2 (a) and (b) show the original recto-verso document, where the verso has been horizontally flipped. Mixing these two images according to the data model, which is parametric concerning the percentage of ink penetration from one side to the other, we produced the three degraded recto images

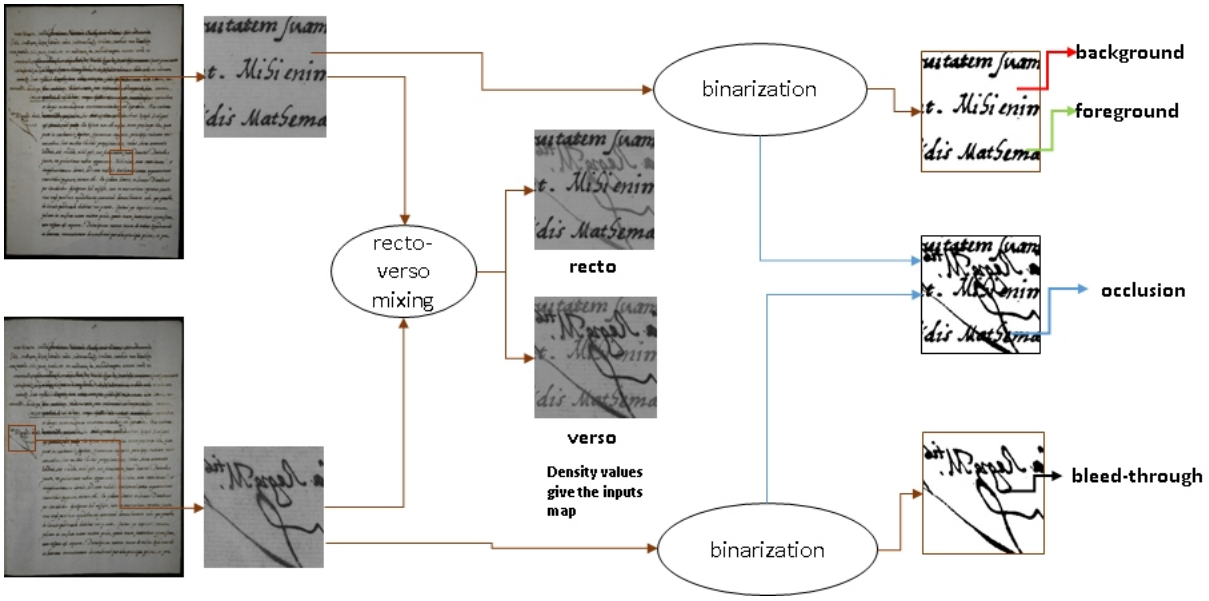


Fig. 1. Diagram illustrating the generation of the training dataset, at the level of a single pair of patches; the derived classes are shown for the recto side only.

shown in Figures 2 (c), (d) and (e), where the degradation is increasing from left to right. The corresponding verso images, not shown, are symmetric with respect to the intensity of the two texts. The details about the model are provided in [11], [12], where a patch-by-patch processing modality is also described that allow to directly work on the misaligned scansions.

We first qualitatively compare the binarizations performed on the degraded versions of the document with our NN and with the algorithm in [16]. These binarizations are shown in Figure 3, for the recto only.

It is possible to observe that, for low degradation, the two methods are competitive. When the degradation becomes heavier, it is evident that it is very difficult to discriminate noise from the text of interest, unless the information contributed by the opposite side of the page is exploited. Moreover, binarization with NN tends to be robust against increasing levels of see-through.

Note also that the text characters are more or less always corroded in all the six panels. This is probably due to the nature of the document itself that, even in its original uncorrupted form, presents many faded text characters. As a consequence, the binarization with [16] of the original document (not shown) is very similar to that of Figure 3 (a), which illustrates our binary map when the degradation is low.

On the six binary maps of Figure 3, we then apply the Matlab `OCR` function. The performance percentages reported in Table I refer to the rate of words and text characters that have been recognized, compared to manual transcription. As a baseline, we used the binary map of the undegraded recto side obtained with [16]. The percentages of corrected words and characters detected in that map are 0.80 and 0.94, respectively. These values represent somehow the best possible recognition rates, given the document characteristics and the OCR algorithm capability.

Despite the corrosion of the text in the binary maps, the recognition of the text characters is generally very good, which means that the used OCR algorithm is good enough.

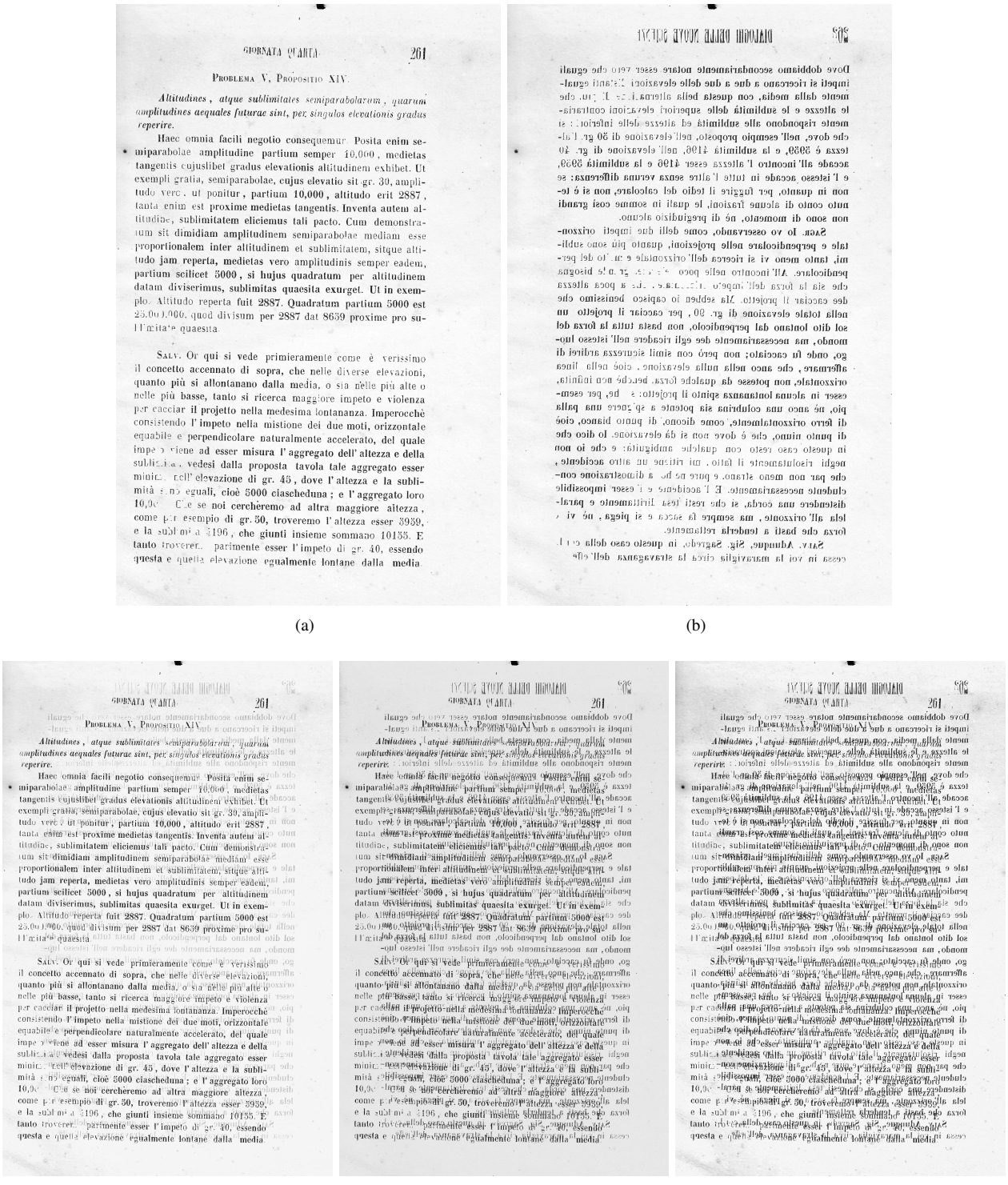
Again, it can be noticed that for the mildest level of see-through, both binarization methods lead to similar recognition performances, which correspond to the best possible ones for that document and those algorithms. The performance of the OCR when binarization is done with [16] tend to degrade with increasing levels of see-through, whereas it is rather stable when binarizazion is based on the NN.

Finally, just as an example, Figure 4 shows the results of the virtual restoration of the degraded RGB recto-verso pair, whose recto in grayscale is shown in Figure 2 (e). This has been obtained as described in [21], still based on the joint recto-verso classification by the same NN.

IV. Experimental results on real historical manuscripts

In this section, we moved to the experimentation of our virtual restoration method on real manuscripts, i.e. ancient handwritten documents. In this case too, we assume the availability of the registered recto and verso sides of the folio.

We tested the method on letters of the correspondence of Christophorus Clavius (1538-1612), conserved in the Historical Archives of the Pontifical Gregorian University in Rome. This manuscript corpus includes some letters that Clavius exchanged with Galileo Galilei. In majority of the letters, the see-through degradation is particularly strong, and it is caused by a real penetration of the ink from one side to the other (bleed-through).



(a) (b)

(c) (d) (e)

Fig. 2. A degraded recto-verso document numerically generated: (a) and (b) the original recto and the mirrored verso; (c), (d) and (e) the recto after mixing the two sides with increasing amount of ink seepage intensity.

	words [16]	words NN	characters [16]	characters NN
mild degradation	0.79	0.80	0.94	0.94
moderate degradation	0.15	0.80	0.54	0.94
strong degradation	0.09	0.79	0.44	0.94

TABLE I
QUALITY INDICES OF THE USED OCR FOR THE SIX BINARY MAPS IN FIGURE3.

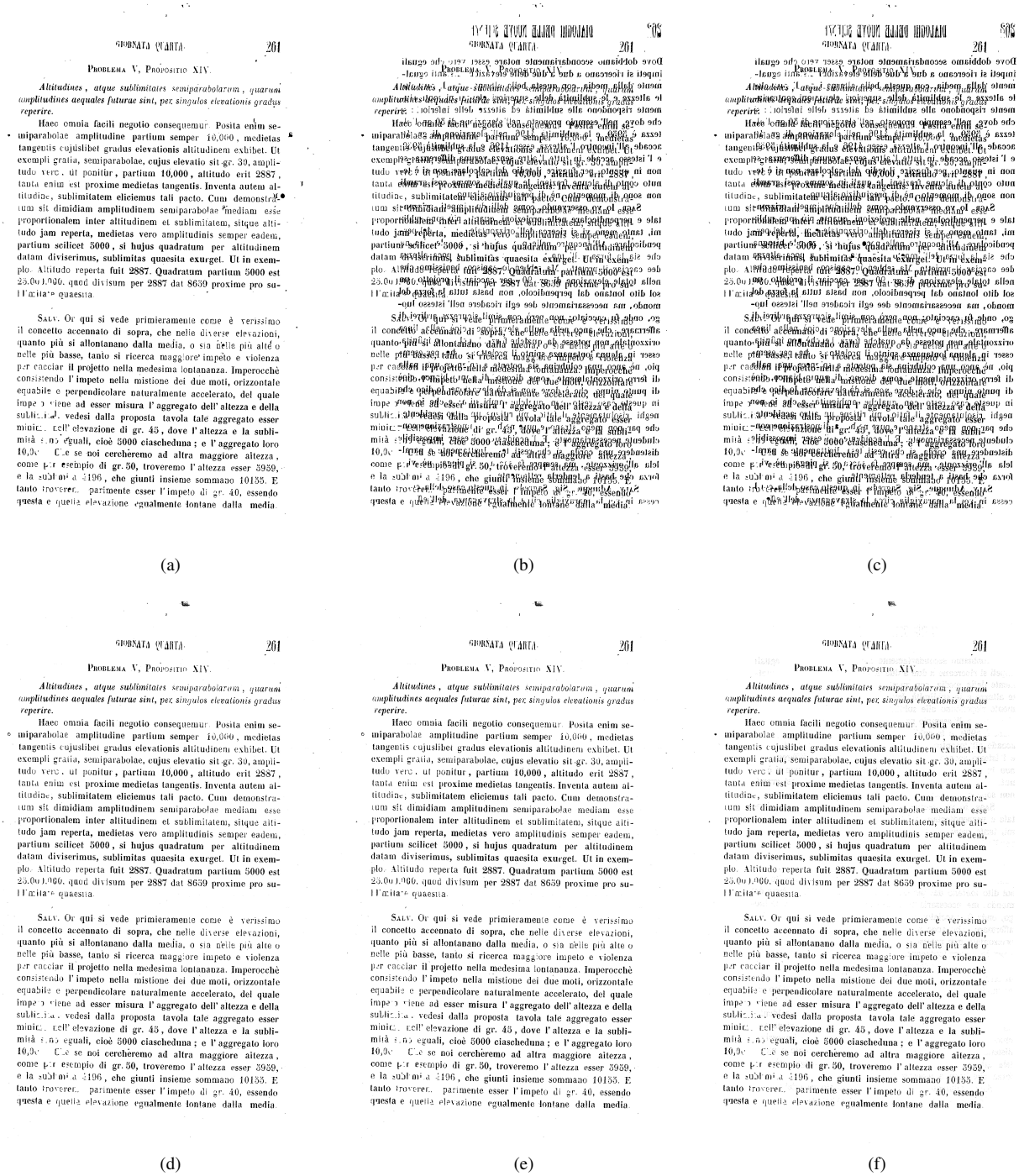
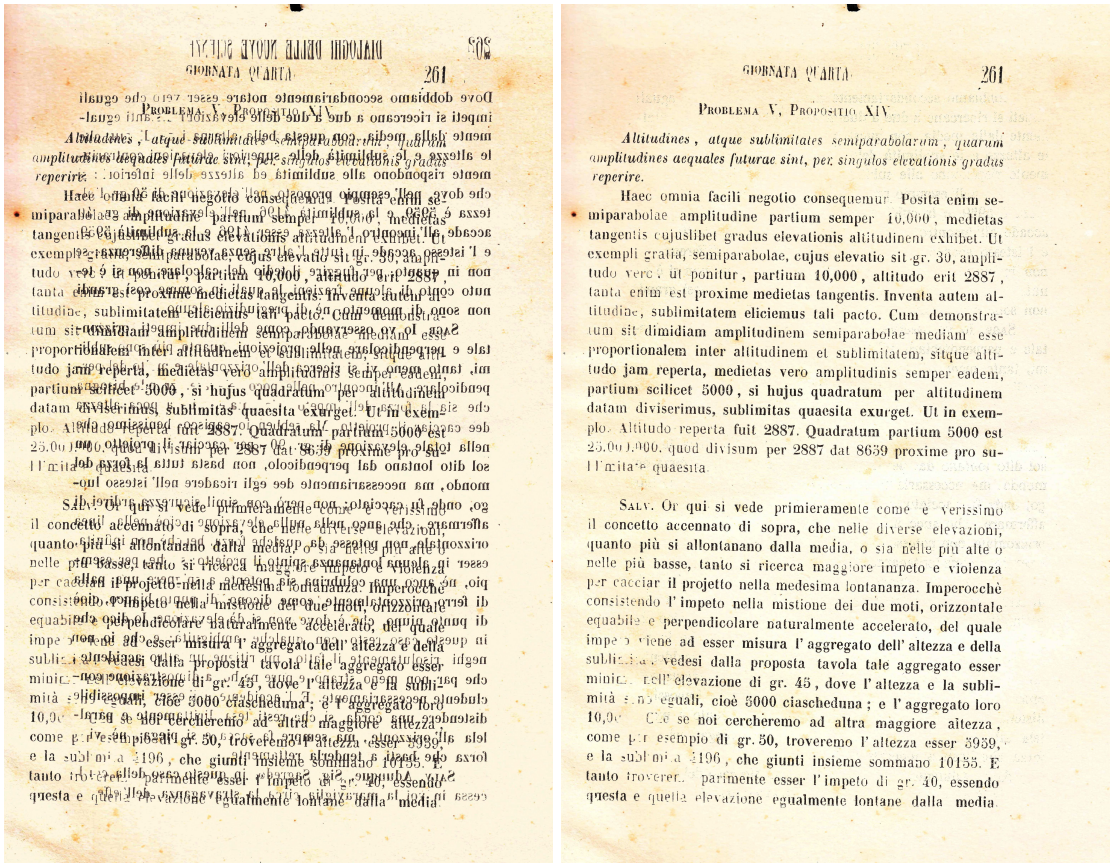


Fig. 3. Binarization of the three degraded recto images for increasing degradation strength (from left to right): first row, algorithm in [16]; second row: our NN.



(a) (b)

Fig. 4. Example of virtual restoration of the RGB document: (a) heavily degraded recto; (d) virtually restored recto.

As previously mentioned, after virtual restoration, we attempted text transcription of these handwritten documents by means of the Transkribus platform [22], specifically designed for the digitization, text detection, transcription and searching of historical documents. Transkribus performs text recognition at the line level and exploits a HTR (Handwritten Text Recognition) model, possibly user provided. The HTR model is generated by a recurrent NN learning process based on manual transcriptions of some parts of the documents (ground truth transcriptions). These user-provided HTR models can be shared online and exploited by other users.

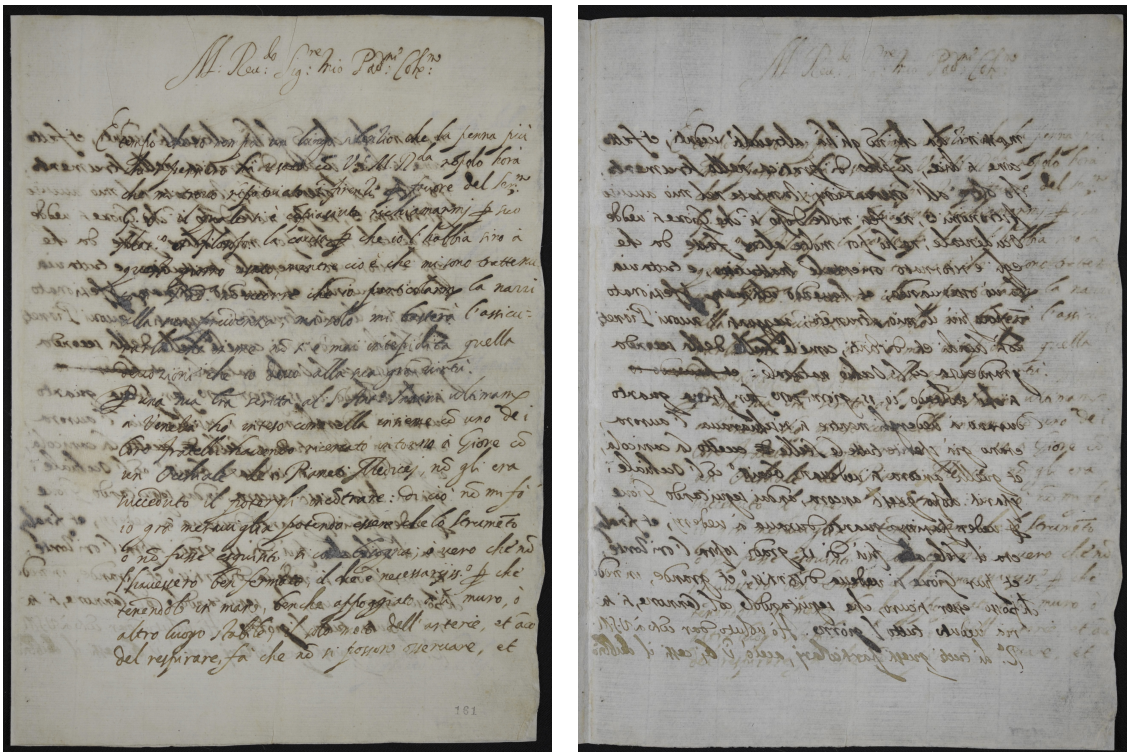
Our application of Transkribus was simplified in the sense that we did not exploit all the functionalities of the system, which would allow us to take advantage of prior knowledge about the document characteristics through the setting of specific parameters. For instance, we did not generate a HTR model for our manuscripts, but chose a model among those proposed by the user interface, and provided to the system by previous users. The model that we chose was trained on Italian administrative manuscripts of the XVI-XVII centuries, which fit quite well with the language and the historical period when our documents were produced.

In the following, we show the results obtained on the autograph letter that Galileo wrote to Clavius on 17 September 1610 from Florence. This letter is identified in the Archives as Apug 530 cc. 161r-162v (see Figures 5 (a) and (b)). The algorithm described in [11] obtained the registration of the originals.

Figure 6 shows the virtually restored version of the recto side (Figure 5 (a)), when our NN classification is applied to the original recto-verso pair. We uploaded to the Transkribus platform both the original and restored recto images, for automatic layout analysis and subsequent transcription.

When Transkribus is used in a blind way, i.e.without intervention from the user, layout analysis can be performed in automatic. Page layout detection is a text analysis tool consisting in the detection of text regions, lines and words. Through the user interface, it is possible to display the output of the layout analysis at different levels. For instance, Figures 7 (a) and (b) show the segmentation of the original recto and of the restored recto into text region and lines. Text regions are identified by the green boxes, and the text lines, contained in each region, are marked by the cyan polygons. This segmentation is produced as output along with the transcription. Various options allow different formats for the layout analysis rendering.

From the Figures 7 (a) and (b), the improvement in the layout analysis, when our NN based enhancement is performed,



(a) (b)

Fig. 5. Real manuscript affected by bleed-through: (a) heavily degraded recto; (b) heavily degraded verso. The two images shown are already registered by the algorithm in [11]. Misaligned originals provided by courtesy of the Historical Archives of the Pontifical Gregorian University in Rome.

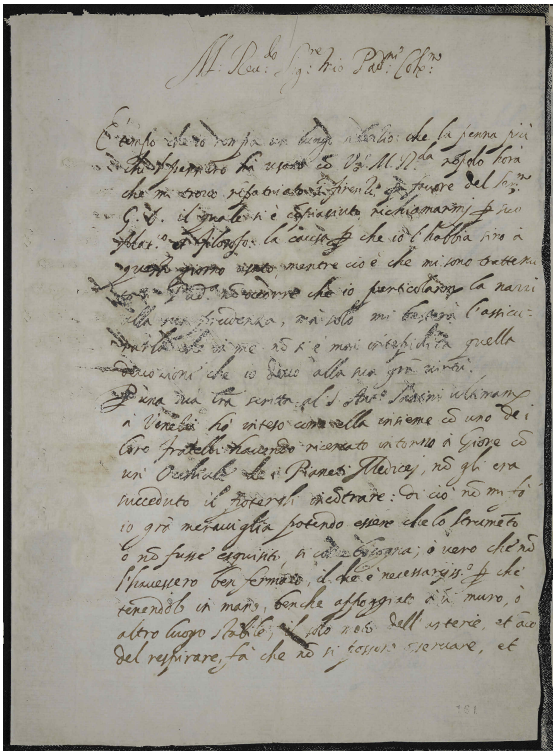
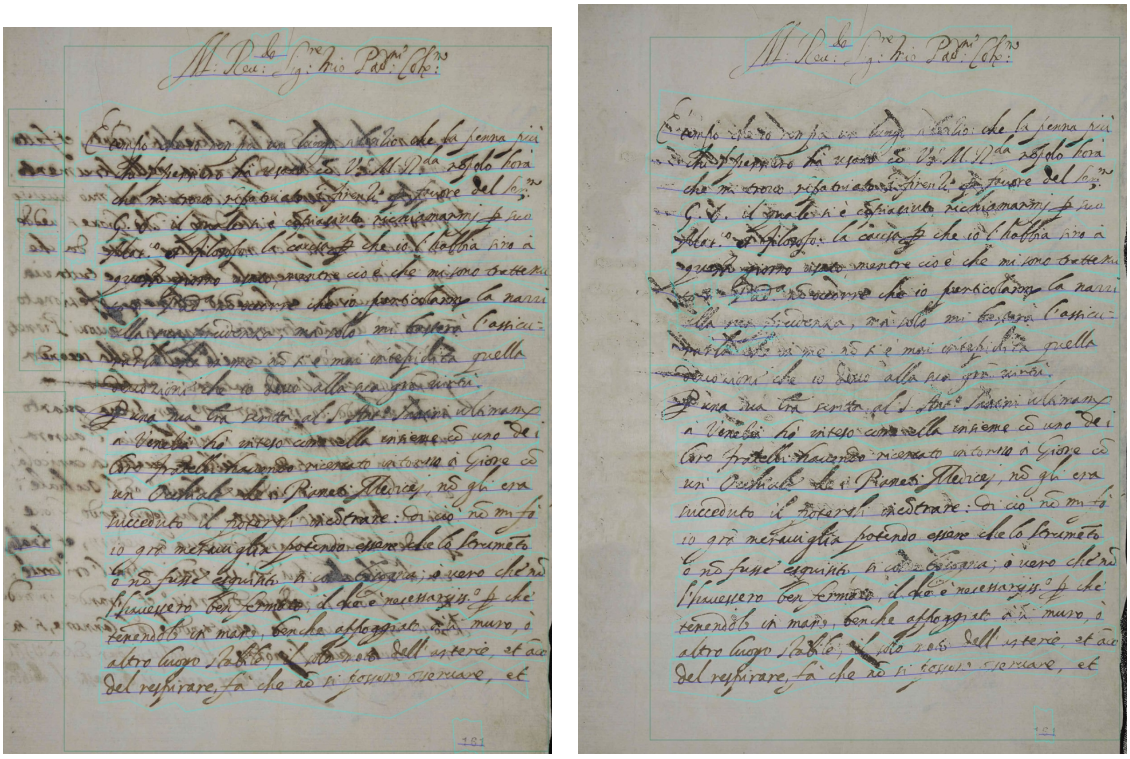


Fig. 6. Virtual restoration of the recto side of Figure 5 (a) obtained by our NN applied to the degraded recto-verso pair.



(a) (b)

Fig. 7. Layout analysis performed by Transkribus on color manuscripts: (a) application to the degraded recto of Figure 5 (a); (b) application to the restored recto of Figure 6.

	words	words - degraded	words - restored	characters	characters - degraded	characters - restored
absolute value	207	108	127	1143	839	874
percentage	1	0.52	0.62	1	0.73	0.77

TABLE II
QUALITY OF THE TRANSCRIPTION IN THE MANUSCRIPT OF FIGURE 5 (ORIGINAL DEGRADED) AND OF FIGURE 6 (VIRTUALLY RESTORED), COMPARED TO THE HUMAN TRANSCRIPTION.

can be clearly appreciated, especially at the higher level of text regions identification. Of course, the region identification influences the subsequent phase of line identification and then of both words spotting and character recognition.

To measure the quality of the automatic transcription performed by Transkribus on the basis of this layout analysis, we exploit the handmade transcriptions provided by the archive that conserves the collection which the manuscript belongs to. The result is that text transcription is still largely unsatisfactory in both the original and the restored images. The recognition rate resulting from the Transkribus software in both cases is very low, as reported in Table II.

We deem that this bad performance of the transcription is due to the almost perfect superposition of the foreground text and bleed-through pattern for this manuscript, which is normally the worst possible situation for the bleed-through degradation. The suppression of most of the spurious characters obtained with our enhancement method seems not to be sufficient to improve automatic transcription.

Better results could be probably obtained if the platform was used in its full potential, first of all, by training it to learn the specific HTR model corresponding to our corpus of manuscripts.

However, in this specific case virtual restoration/binarization of the enhanced manuscript can be more satisfactorily exploited for helping the manual transcription, which is undoubtedly very difficult, heavy and tiring if performed on the original degraded manuscript.

V. CONCLUSIONS

We have shown that, by exploiting the information contained in both the recto and verso sides of ancient documents affected by ink penetration or transparency, it is possible to train a very simple shallow NN to correctly classify pixels into primary text, paper background, see-through noise and overlaid texts, without the need for an external training set. The example-target 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 main text, so that OCR algorithms can be applied for character recognition in

printed documents, or virtually restored copies of the original manuscripts, for use with automatic text transcription systems.

In terms of binarization, we compare our results with those provided by the winning algorithm of the H-DIBCO-2018 [1] competition. Our results are clearly superior, especially when the degradation is strong. Of course, this is possible because we use twice as much information. However, this extra information is normally available in the libraries and archives, and must be exploited to binarize the opposite side. We then processed the two sides jointly, demonstrating that this makes a great difference, for the same amount of information used.

We provided both qualitative and quantitative results of our pre-processing technique and the subsequent text recognition. On the binary maps of the enhanced printed documents, we used the `OCR` MathWorks function derived from the Tesseract Open Source OCR Engine. In this case, we quantified the character recognition results on artificially degraded texts, starting from the scans of a real ancient document. For real degradations on historical manuscripts, we used the `Transkribus` platform developed for treating handwritten texts.

Further work on this subject will be mainly devoted to implementing different, more complex network architectures, aiming to obtain a more accurate classification in the presence of stronger degradations. We will also pay attention at ameliorating the mathematical degradation model for the see-through and other typical degradations.

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] F. Lombardi, and S. Marinai, "Deep Learning for Historical Document Analysis and Recognition—A Survey," *Journal of Imaging*, vol. 6, 110, 2020.

[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] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] J. Sauvola and M. Pietikäinen, "Adaptive document image binarization," *Pattern Recognition*, vol. 33, p. 225–236, 2000.

[14] S. He, and L. Schomaker, " DeepOtsu: Document Enhancement and Binarization using Iterative Deep Learning," *Pattern Recognition*, vol. 9, pp 379–390, 2019.

[15] Hagan, M.T., H.B. Demuth, and M.H. Beale, Neural Network Design, Boston, MA: PWS Publishing, 1996.

[16] 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.

[17] 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.

[18] Smith, R., D. Antonova, and D. Lee, "Adapting the Tesseract Open Source OCR Engine for Multilingual OCR," Proceedings of the International Workshop on Multilingual OCR, 2009.

[19] P. Savino, A.Tonazzini, "Mathematical models and neural networks for the description and the correction of typical distortions of historical manuscripts," Workshop on Mathematical Methods for Image Processing and Understanding MMIPU 2023, O. Gervasi et al. (Eds.): ICCSA 2023 Workshops, LNCS 14108, pp. 545–557, 2023.

[20] P. Savino, A.Tonazzini, "Preprocessing of recto-verso printed documents based on neural networks for text analysis," CiSt-DPWH 2023, Agadir, December 2023, to be published in IEEE Xplore, 2024.

[21] P. Savino, A.Tonazzini, "Training a shallow NN to erase ink seepage in historical manuscripts based on a degradation model," Neural Computing and Applications, https://doi.org/10.1007/s00521-023-09354-7,"

[22] "Transkribus - Unlock historical documents with AI", www.transkribus.eu

[23] S. Colutto, P. Kahle, G. Hackl, G. Muhlberger, "Transkribus. A Platform for Automated Text Recognition and Searching of Historical Documents", 15th International Conference on eScience (eScience), pp. 463-466, 2015.

[24] R. Smith, "An Overview of the Tesseract OCR Engine", 9th Int. Conf. on Document Analysis and Recognition, 2007, pp. 629-633.

[25] "Tesseract Open-Source OCR", http://code.google.com/p/tesseract-ocr