Bleed-through Removal by Learning a Discriminative Color Channel

Mauricio Villegas
PRHLT, Universitat Politècnica de València
Camí de Vera s/n, 46022 València, Spain
mauvilsa@upv.es

Alejandro H. Toselli
PRHLT, Universitat Politècnica de València
Camí de Vera s/n, 46022 València, Spain
ahector@prhlt.upv.es

***Abstract*—This paper proposes a novel bleed-through removal technique based on learning a color channel that is optimized so that the foreground text is enhanced while at the same time the variability of the background (including the bleed-through) is diminished. The technique is intended to be part of an interactive transcription system in which the objective is obtaining high quality transcriptions with the least human effort. Thus, instead of training the bleed-through removal to work in general for any document, the technique requires a user to label regions both as foreground text and as bleed-through, with the aim that the method is adapted to the characteristics of each document. The proposal is assessed using the handwritten recognition performance on a real 17th century manuscript.**

***Keywords*-Handwritten Text Recognition; Scanned Document Noise Removal; Bleed-through; Show-through**

I. Introduction

One of the sources of noise that affects negatively the performance of handwritten text recognition (HTR) systems that take as input scanned pages, is the appearance of content from the opposite side of the page. Due to the current need of digitalization, this problem has attracted considerable attention as can be observed in the research literature. This effect is caused by two phenomena, commonly known as *show-through* and *bleed-through*: the former is due to the transparency of the paper while the latter is due to the seeping of ink from opposite side of the page. For simplicity this effect as a whole (the combined effect of transparency and seeping of ink) will be referred to as bleed-through.

Many of the proposed bleed-through removal methods, such as [1], [2], require to have both the recto and verso sides and a registration between them, that is, the alignment between the scanned front and back sides of each page. After this, these methods try to differentiate the pixels belonging to recto or verso by diverse approaches: physics models, filters, classifiers, etc. Also requiring both sides and a registration process, [3] describes an approach for grayscale images, from the perspective view of the *blind source separation* (BBS) problem. In this case *Independent Component Analysis* (ICA) and *Principal Component Analysis* (PCA) are employed to classify verso and recto pixels. A related version of this approach is presented in [4], which takes advantage of images in color and does not require the verso side or registration. However, a common shortcoming of both of these approaches is that they assume linear modeling simplifications which in practice are not adequate enough [5]. Other registration-free methods are [6], [5], where the former employs a sophisticated updating threshold technique to filter out bleed-through pixels and the latter classifies pixels in verso and recto using Markov Random Fields (MRFs).

In general most of the work done on bleed-through removal, including the before-cited ones, has been aimed at finding solutions that solve this problem without any human intervention and for any type of document. However, this is difficult to achieve in practice since many factors affect the scanned documents such as: color of the paper and ink, degradation due to age, characteristics of the scanner, etc. Therefore, a general solution might not be optimal for each type of document.

One possibility to overcome this problem is to let a user point out explicitly for some sample pages which parts are and which parts are not bleed-through. This user feedback can then be employed to adapt the bleed-though removal technique to the current document being processed to hopefully improve the recognition performance. This follows a similar idea as described in [7], where an interactive handwritten text recognition application takes advantage from the user feedback, when he/she corrects a missrecognized word, to suggest the most likely transcription of the remaining text. The final objective of a such application is to obtain a high quality transcription of the document, so instead of using only a completely automatic approach, it incorporates also the interactive supervision of the user. In this way, the efficiency of automatic HTR systems with the accuracy of the user are combined to achieve high quality transcriptions at minimum human effort. The before-mentioned user feedback consisting in annotating a few bleed-through zones can be added to this interactive scenario contributing to improve the final suggested transcriptions such that in the end the total effort for transcribing the complete document is reduced.

In this paper a bleed-through removal method is proposed which does not require the verso side or registration and is based on the assumption that the user labels a series of regions as bleed-through and clean text, a task which does not require much effort. Since the proposal is expected to work within the before-mentioned interactive scenario, the quality

of bleed-through removal is evaluated by observing the effect on the final handwritten text recognition performance.

II. Bleed-through Removal Method

The proposed technique for removing bleed-through is based on learning a discriminative color channel by considering a set of labeled local image patches. The labeled data is human supervised, obtained by the selection of example image regions for two classes: bleed-through and clean handwritten text. In the case of the clean text it is not too important that bleed-through be completely absent, although the cleaner the better. However, for the bleed-through class it is important that no text strokes be included in the data. The method is intended to work only for a single document such that between the pages there are similar conditions (paper and ink color, degradation, scanning environment, etc.), for which it is safe to assume that the bleed-through can be discriminated in the color space.

Figure 1 presents example images of what the labeled regions can be. The top two rows show simple bounding box regions which are extremely simple to label, however, depending on the document it may or may not be easy to find large regions of only bleed-through or clean text. So for hard cases, these bounding box regions would be composed of single lines or narrow columns. Alternatively the regions could be defined by curves, e.g. the third row of Figure 1. A more elaborate technique could be computer assisted. For example, based on the result of a line detector, the regions of possible text and bleed-through could be tentatively marked which would then be corrected by a person, and finally obtaining data such as the bottom row of Figure 1.

In any case the human effort required for labeling data is relatively low particularly because from a single labeled region generally a large amount of local image patches can be extracted.

Figure 1. Example image regions from the Plantas corpus labeled as clean text (left column) and bleed-through (right column). Depending on the data, the selected regions can be simple bounding boxes (e.g. top two rows) or more complex defined regions (e.g. bottom two rows). From each labeled region a large amount of local patches is extracted.

A. Color Channel Learning

The objective of the technique is to find a parametrized transformation function $f_{\boldsymbol{\theta}}$ that maps each pixel from the original image color space into a single discriminative color channel, being $\boldsymbol{\theta}$ the set of parameters. In mathematical terms this is $f_{\boldsymbol{\theta}} : \mathbb{R}^C \rightarrow \mathbb{R}$, where C is the dimensionality of the input color space. The simplest model would be for $f_{\boldsymbol{\theta}}$ to take into account only the pixel's C color values, however, other sources of information could also be useful. For instance it could consider context properties such as the mean and standard deviation for each of the original color channels for a window whose center is the pixel in question.

In order to find a function $f_{\boldsymbol{\theta}}$ which reduces the effect of bleed-through while at the same time preserving the information of the foreground handwritten text, the chosen strategy was to maximize the ratio of expected patch variances, where in the numerator is the expectation for patches of clean text and in the denominator the expectation for patches of bleed-through. Formally this objective can be expressed as

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \frac{\mathsf{E}\left[\mathrm{var}(\boldsymbol{f}_{\boldsymbol{\theta},\mathsf{x}})\right]}{\mathsf{E}\left[\mathrm{var}(\boldsymbol{f}_{\boldsymbol{\theta},\mathsf{y}})\right]} \ , \tag{1}$$

where x, y represent random patches for clean text and bleed-through, respectively, $\mathsf{E}[]$ is the expectation operator, $\mathrm{var}()$ is a function that gives the variance for the elements of a vector and $\boldsymbol{f}_{\boldsymbol{\theta},\mathsf{z}}$ is a vector whose elements are obtained by applying individually the function $f_{\boldsymbol{\theta}}$ to each of the pixels of patch z, thus for patches of N pixels then $\boldsymbol{f}_{\boldsymbol{\theta},\mathsf{z}} \in \mathbb{R}^N$.

Optimizing Eq. 1 for any $f_{\boldsymbol{\theta}}$ is undesirable since some functions can give singularities and others can overfit the training data, a case in which the solution is unlikely to generalize well to unseen data. In this paper it is only considered the following family of transformation functions which can avoid the aforementioned problems:

$$f_{\boldsymbol{b}}(\boldsymbol{p}) = \left[\boldsymbol{g}(\boldsymbol{p})\right]^{\mathsf{T}} \boldsymbol{b} \ , \tag{2}$$

where $\boldsymbol{p}$ is a vector representation of a pixel, $\boldsymbol{b} \in \mathbb{R}^D$ is a projection vector which is the only parameter to optimize by Eq. 1 and the function $\boldsymbol{g} : \mathbb{R}^C \rightarrow \mathbb{R}^D$ is a fixed pixel transformation that will be commented later. Now, let

$$\boldsymbol{f}_{\boldsymbol{b},\mathsf{z}} = \boldsymbol{G}_{\mathsf{z}} \boldsymbol{b} \ , \tag{3}$$

where the rows of matrix $\boldsymbol{G}_{\mathsf{z}} \in \mathbb{R}^{N \times D}$ correspond to the output of function $\boldsymbol{g}$ for each of the pixels of the patch.

Using Eq. 3 it can be shown that the empirical variance for a dataset of patches $\mathcal{Z}$ can be expressed as

$$\mathsf{E}\left[\mathrm{var}(\boldsymbol{f}_{\boldsymbol{b},\mathsf{z}})\right] \approx \underbrace{\boldsymbol{b}^{\mathsf{T}}\left[\frac{1}{|\mathcal{Z}|N}\sum_{\forall \boldsymbol{z}\in\mathcal{Z}} \boldsymbol{G}_{\boldsymbol{z}}^{\mathsf{T}}\left(\boldsymbol{I}-\frac{\mathbf{1}_{N\times N}}{N}\right)\boldsymbol{G}_{\boldsymbol{z}}\right]\boldsymbol{b}}_{\boldsymbol{H}_{\mathsf{z}}\in\mathbb{R}^{D\times D}}, \quad (4)$$

where $\mathbf{1}_{N\times N}$ is a square matrix of N rows/columns with all elements equal to one. Using Eq. 4, the objective of Eq. 1 becomes the well known generalized Rayleigh quotient:

$$\hat{\boldsymbol{b}} = \arg\max_{\boldsymbol{b}} \frac{\boldsymbol{b}^{\mathsf{T}}\boldsymbol{H}_{\mathsf{x}}\boldsymbol{b}}{\boldsymbol{b}^{\mathsf{T}}\boldsymbol{H}_{\mathsf{y}}\boldsymbol{b}} . \quad (5)$$

The solution to Eq. 5 is given by the generalized eigenvector associated to the largest generalized eigenvalue λ of $\boldsymbol{H}_{\mathsf{x}}\boldsymbol{b} = \boldsymbol{H}_{\mathsf{y}}\boldsymbol{b}\lambda$. Since obtaining training samples is relatively cheap and the dimensionality D tends to be low, the matrices $\boldsymbol{H}_{\mathsf{x}}$ and $\boldsymbol{H}_{\mathsf{y}}$ can be expected to be full rank, thus there are no infinite generalized eigenvalues.

Among the family of functions in Eq. 2, the simplest form would be $\boldsymbol{g}(\boldsymbol{p}) = \boldsymbol{p}$, in which case $f_{\boldsymbol{b}}$ is a linear combination of the input color values. This linear model is exceedingly simple, although it is capable of providing bleed-through reduction as can be observed in Section IV. In essence the function $\boldsymbol{g}$ was introduced to allow more complex nonlinear transformations which can be important to handle for example a larger color variability of the text ink. Apart from the linear model, in this paper two other options for $\boldsymbol{g}$ have been considered, specifically general second and third order models, i.e., all product combinations of pixel values up to second or third order.

B. Discretization and Gamma Correction

The optimization presented in Section II-A does not guaranty that the output of the color transformation be in a specific range and furthermore the output is real valued, however, a discretized output (e.g. 8-bit) might be required for subsequent image processing steps. This discretization requires to have a limited interval, i.e., minimum and maximum values, and a mapping function to the discrete space. The minimum and maximum are chosen by analyzing the values obtained for the clean text training patches $\mathcal{X}$, although discarding a small percentage at the extremes to account for possible outliers, e.g. 0.1%. Preliminary results showed that a uniform mapping function was not adequate, especially for the nonlinear models. However, it was observed that a gamma correction tends to provide a good enough solution.

The technique to obtain the gamma value automatically which gives adequate discretized images, is based on the fact that the histogram of handwritten text documents has a well known distribution. As illustrated in Figure 2, the histograms

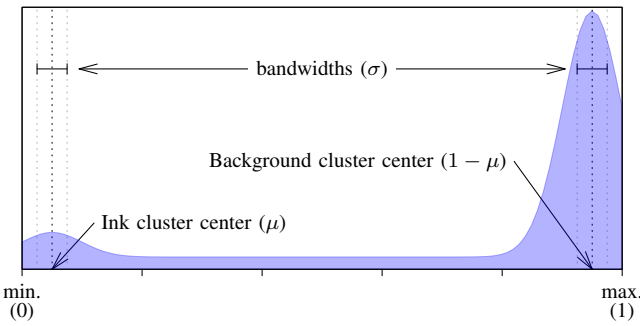

Figure 2. Plot of Eq. 6 used as the reference histogram of a handwritten text document employed for the estimation of the gamma parameter.

tend to have two clusters, one for the color of the ink and a much larger one for the color of the background. The gamma values are sampled and the one chosen is the one that gives the least Jensen-Shannon divergence between the histogram of the training data $\mathcal{X}$ and a prototypical histogram (see Figure 2) defined by the function

$$y(x) = \frac{\overbrace{\mathcal{N}(x,\mu,\sigma) + A\mathcal{N}(x,1-\mu,\sigma) + B}^{z(x)}}{\int_0^1 z(x')\,dx'} , \quad (6)$$

where $\mathcal{N}$ is the normal distribution probability density function, A is the amplitude of the background cluster w.r.t. the ink cluster, B is a constant to account for pixels between the clusters, and the denominator simply normalizes to guaranty a sum to one. Note that this gamma is adjusted and fixed for the optimized color channel. It is not readjusted for each input image to process.

The knowledge regarding the histogram is also used to invert the direction of the projection $\boldsymbol{b}$ if required (since the direction is irrelevant in the eigenvalue decomposition) so that in the resulting space the background corresponds to white and text to black.

III. HTR System Overview

The quality of the proposed bleed-through removal method will be indirectly assessed through the obtained HTR system performance on text images with removed bleed-through. For carrying out this, a conventional off-line HTR architecture is adopted, composed of three modules: *preprocessing, feature extraction* and *recognition*; see [8].

The preprocessing is aimed at correcting image degradations and geometry distortions: skew, slant corrections, and size normalization. On the other hand, the feature extraction process transforms a preprocessed text line image into a sequence of 60-dimensional feature vectors. For more details about preprocessing and feature extraction refer to [8].

The recognition process is based on Hidden Markov Models (HMMs), that is, characters are modeled by continuous density left-to-right HMMs. The Gaussian mixture

is a probabilistic approach to model the emission of feature vectors in each HMM state.

Each lexical word is modeled by a stochastic finite-state automaton, which represents all possible concatenations of individual characters to compose the word. On the other hand, text sentences are modeled using bi-grams with Kneser-Ney back-off smoothing [9], estimated directly from the training transcriptions of the text line images. In practice, bi-gram language models are affected by the so-called *grammar scale factor* and by the length-dependent factor *"word insertion penalty"*. Both are tuned empirically to better balance the contribution of morphological HMM characters and bi-gram models [10].

All these finite-state (HMM character, word and sentence) models can be easily integrated into a single global model, on which a decoding process is efficiently performed by the Viterbi algorithm [10].

IV. Experimental Evaluation

In the results, the proposed method is referenced as LDCC (Learning a Discriminative Color Channel). As baseline two methods have been employed. The first one is a standard conversion of the images to grayscale, analog to the proposed technique since it also takes as input a color image and produces a single channel. The second baseline is the Double Markov Random Field (MRF) method [5], which was the only comparable technique (that handles color images and not requiring the verso side) for which we were able to obtain the corresponding software. The parameters of the Double MRF were the same as in [5].

A. Assessment Measures

As commented in Section I, quality evaluation of bleed-through removal from text images will be done under the perspective of the improvement level achieved in the recognition of them by employing a HTR system.

The quality of the automatic recognition obtained by a HTR system is measured by means of the *Character Error Rate* (CER). It is defined as the minimum number of characters that need to be substituted, deleted, or inserted to match the recognition output with the corresponding reference ground truth, divided by the total number of characters in the reference transcriptions.

B. Dataset Description

The manuscript collection *Historia de Las Plantas* (Plantas for short) was written by Bernardo de Cienfuegos, one of the most outstanding Spanish botanists of the 17th century. As this collection came to be handwritten using a quill-pen, the bleed-through effect becomes quite evident in several of its pages (see Figures 1 and 3). This collection consists of seven volumes, which are currently held at the Biblioteca Nacional de España. Moreover, a digital reproduction can be found at the Biblioteca Digital Hispánica.[1]

Table I
Basic statistics of the Plantas Prologue chapter.

Num. Pages	38
Num. Lines	1,206
Running Words	11,642
Lexicon	3,899
Running Chars	61,973
Num. Chars	71

For the assessment of the proposed approach, experiments were carried out only on the Prologue chapter belonging to the first volume of Plantas. Table I shows basic statistics of this chapter.

C. Experimental Setup

For exprimental evaluation, a 10-fold cross-validation has been performed on the 38 pages of the Prologue chapter, which contains $1,206$ text line images (see Table I). For each cross-validation run, using only the training samples, an open-vocabulary dictionary was built, and a Kneser-Ney smoothed n-gram model was trained.

In the preprocessing no slant correction was applied and the skew parameters were obtained for the grayscale images and kept fixed for all methods being compared. After bleed-through removal, the images were further cleaned by a Sauvola-like [11] algorithm which preserves grayscale information. Finally feature extraction was carried out on preprocessed line images obtaining the 60-dimensional feature vector sequences according to Section III. On the other hand, the 71 different characters were modeled by continuous density left-to-right HMMs with the same number of states and Gaussian mixture components per state.

The meta-parameter values of the HTR system were tuned empirically using the cross-validation on the baseline grayscale images (see Table II) and also kept fixed for the comparison of the remaining different methods. The final parameters used for recognition were HMMs of 12 states, Gaussian mixtures of 32 components per HMM state, grammar scale factor of 108 and word insertion penalty -15.

For the LDCC optimization, image regions from other chapters (not the Prologue) were selected and labeled. In total we used 11 regions for clean text and 9 regions for bleed-through. The parameters of the prototypical histogram in Eq. 6 were varied, but it was observed that they did not affect the results significantly, so we kept them fixed as $\mu = 5\%$, $\sigma = 5\%$, $A = 10$ and $B = 2$.

D. Results and Discussion

Figure 3 presents some example images comparing the original with the result after applying bleed-through removal

[1] http://www.bne.es/es/Catalogos/BibliotecaDigitalHispanica/Inicio/

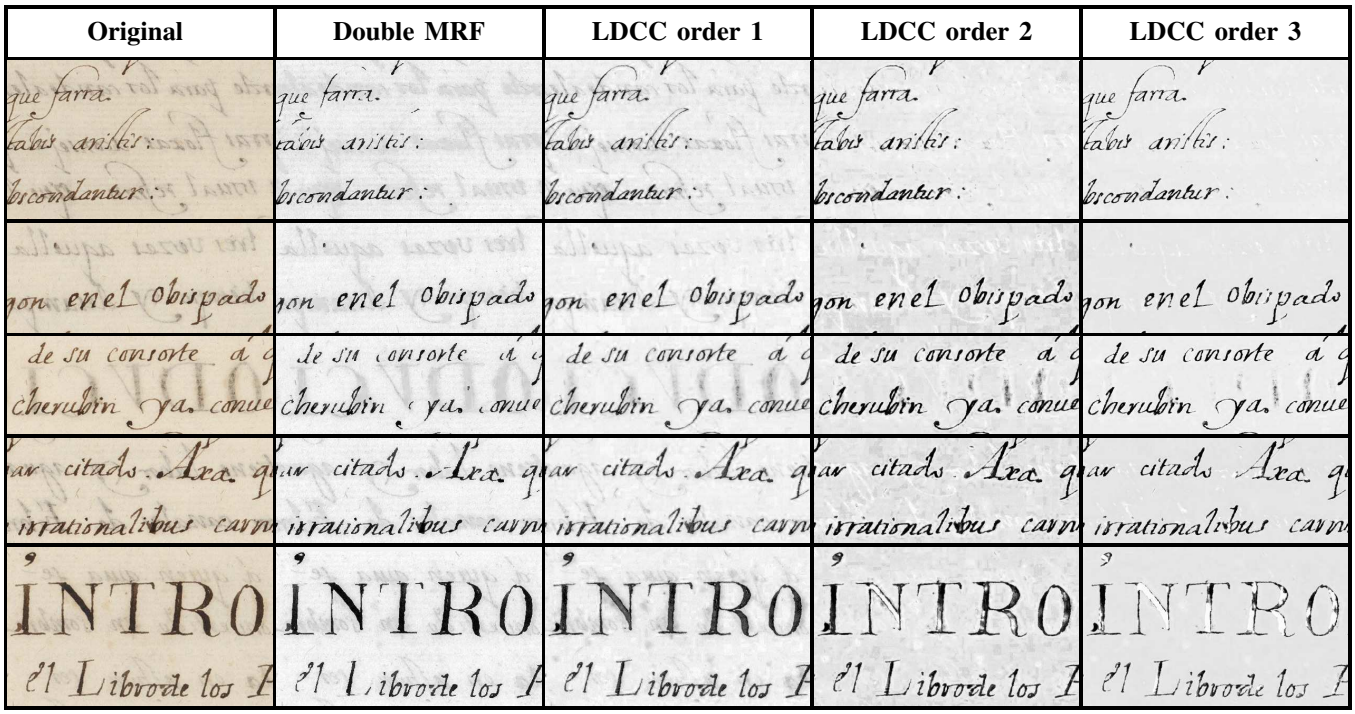

Figure 3. Example images extracted from the Plantas Prologue chapter comparing the original with the results after applying different bleed-through removal methods.

with each of the techniques. Analyzing the proposed LDCC for the different orders of the model, it can be noted that the optimization objective Eq. 5 works as expected. The text is preserved while the variance of the background tends to be reduced. As the order of the model increases, the bleed-through is reduced further, however, a negative effect is also observed. Some parts of the text strokes are removed (see Figure 3 LDCC order 3, bottom two rows), an effect that seems to happen to wide strokes or large dots. At the moment we do not have a definite explanation for this. However, a possibility is that the local patches obtain a higher variance if strokes with large area are hollowed (make the center white), thus if there is a difference in color between the borders and the center of the strokes, then the optimization will lead to this effect. In future works this could be analyzed further and propose possible improvements to the objective functions so that this is avoided.

The Double MRF does not seem to reduce much the bleed-through. Furthermore, by looking closer at the images, it can be observed that there are several text strokes classified incorrectly (i.e. removed), see Figure 3: row 1 letter *f* of word *farra*, row 3 letter *c* of word *consorte*, row 5 letter *L* of word *Libro*. This result suggests that probably the parameters of this technique need to be fine tuned for it to work correctly on new datasets.

The text recognition results over the ten cross-validation runs for the grayscale baseline and the four bleed-through removal techniques are presented in Table II. Even though

Table II
HANDWRITTEN TEXT RECOGNITION PERFORMANCE ON THE PLANTAS CORPUS COMPARING THE BASELINE TECHNIQUES WITH THE BEST PERFORMANCE OF LDCC FOR THE DIFFERENT ORDERS OF THE MODEL.

Method	CER (%)	95% Conf. Int.
Grayscale	27.61	27.25 – 27.97
Double MRF [5]	29.09	28.73 – 29.45
LDCC order 1	**25.39***	25.04 – 25.74
LDCC order 2	**26.04***	25.69 – 26.39
LDCC order 3	27.70	27.34 – 28.06

*Statistically significantly better than Grayscale for a confidence level of 99% using a two-proportion z-test.

in Figure 3 the Double MRF seems similar to LDCC order 1, its recognition performance is considerably worse than both LDCC order 1 and grayscale. This is possibly due to the unwanted removal of strokes by the Double MRF that was commented before. On the other hand, among the LDCCs, the best performance is for LDCC order 1 which is somewhat unexpected considering the visual appearance of the images.

Figure 4 shows a plot of the behavior of the CER performance for the LDCCs as the size of the local patches is varied. The plot also includes the performance for grayscale and Double MRF and the 95% confidence intervals estimated using Wilson's method. As can be observed, the performance of both LDCC order 1 and order 2 is relatively stable and consistently better by a wide margin than grayscale. In

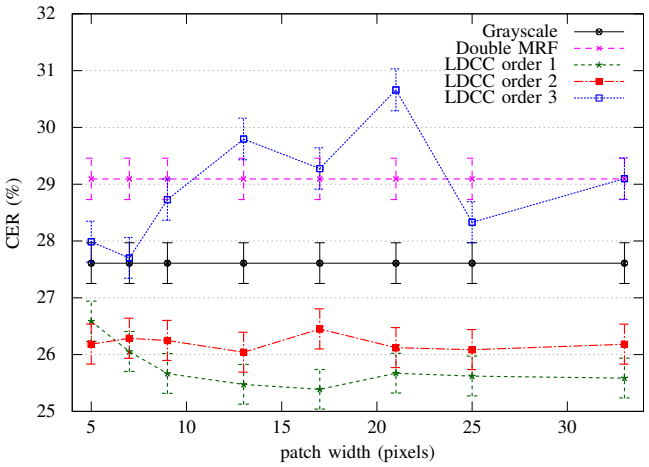

Figure 4. Plot comparing the CER recognition performance of the grayscale and Double MRF [5] baselines with the proposed LDCC for different patch sizes and orders of the model.

the case of LDCC order 3, the behavior is a bit erratic, suggesting that a third order model is too general possibly with problems of overfitting or affected greatly by the hollow effect mentioned earlier.

V. Conclusions and Future Work

This paper has presented a new bleed-through removal technique based on an optimized pixel-by-pixel transformation from color to a single channel. The method has been designed to be part of an interactive transcription system, in which the bleed-through technique is adapted to each particular document such that it is safe to assume that the conditions between pages is similar. The adaptation is done by a user labeling a series of example regions as either clean text or bleed-through, a very simple task that has the potential of reducing the overall effort of interactively transcribing a large document thanks to the improvement of the suggested transcriptions.

Experimental results on a realistic 17th century manuscript show a significant recognition improvement in comparison to a basic grayscale conversion and the Double MRF bleed-through removal technique. By visually inspecting the resulting images, it can be observed that the proposed optimization objective effectively diminishes the bleed-through, and as the representation capability of the model is increased, the reduction of bleed-through is each time more evident. However, for the highest order model tried the text recognition performance decreased significantly.

Future research should be focused on finding why the performance of the higher models affects performance, and proposing improved optimization criteria that account for it. Also, contextual information (not just the pixel's color values) could be used for further improving the bleed-through removal performance. Experimentation must also be performed on other datasets, explore better how it performs in comparison with other techniques, and analyze the behavior as the users add more labeled regions.

Acknowledgments

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under the tranScriptorium project (#600707) and from the Spanish MEC under the STraDA project (TIN2012-37475-C02-01).

References

[1] G. Sharma, "Show-through cancellation in scans of duplex printed documents," *Image Processing, IEEE Transactions on*, vol. 10, no. 5, pp. 736–0754, 2001. 1

[2] E. Dubois and A. Pathak, "Reduction of bleed-through in scanned manuscript documents," in *PICS*, 2001, pp. 177–180. 1

[3] A. Tonazzini, E. Salerno, and L. Bedini, "Fast correction of bleed-through distortion in grayscale documents by a blind source separation technique," *International Journal of Document Analysis and Recognition (IJDAR)*, vol. 10, no. 1, pp. 17–25, 2007. [Online]. Available: http://dx.doi.org/10.1007/s10032-006-0015-z 1

[4] A. Tonazzini, L. Bedini, and E. Salerno, "Independent component analysis for document restoration," *International Journal of Document Analysis and Recognition (IJDAR)*, vol. 7, no. 1, pp. 17–27, 2004. 1

[5] C. Wolf, "Document ink bleed-through removal with two hidden markov random fields and a single observation field," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, no. 3, pp. 431–447, 2010. 1, 4, 5, 6

[6] R. Estrada and C. Tomasi, "Manuscript bleed-through removal via hysteresis thresholding," in *Document Analysis and Recognition, 2009. ICDAR '09. 10th International Conference on*, 2009, pp. 753–757. 1

[7] A. H. Toselli, V. Romero, M. Pastor, and E. Vidal, "Multimodal interactive transcription of text images," *Pattern Recognition*, vol. 43, no. 5, pp. 1824–1825, 2009. 1

[8] A. H. Toselli, A. Juan, D. Keysers, J. González, I. Salvador, H. Ney, E. Vidal, and F. Casacuberta, "Integrated Handwriting Recognition and Interpretation using Finite-State Models," *Int. Journal of Pattern Recog. and Artificial Intelligence*, vol. 18, no. 4, pp. 519–539, 2004. 3

[9] R. Kneser and H. Ney, "Improved backing-off for N-gram language modeling," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP '95)*, vol. 1. Los Alamitos, CA, USA: IEEE Computer Society, 1995, pp. 181–184. 4

[10] F. Jelinek, *Statistical methods for speech recognition*. MIT Press, 1998. 4

[11] J. Sauvola and M. Pietikäinen, "Adaptive document image binarization," *Pattern Recognition*, vol. 33, no. 2, pp. 225–236, 2000. 4