Bleed-Through Removal in Degraded Manuscripts

Róisín Rowley-Brooke[†] **and Anil Kokaram**[*]

Department of Electronic & Electrical Engineering, Sigmedia Group
Trinity College Dublin

E-mail: [†]`rowleybr@tcd.ie` [*]`anil.kokaram@tcd.ie`

Abstract — **Ancient manuscripts represent important artefacts in cultural heritage, yet many are in very poor condition. One significant problem that affects the readability of some manuscripts is bleed-through degradation, where text from the recto or verso side bleeds through the folio. In the last decade there has been an increasing amount of work in this area which, in general, takes a machine learning approach. This paper presents a practical Bayesian approach to the problem, building a relatively low compute intensive algorithm starting with a linear degradation model. Results show some convincing removal of bleed-through.**

Keywords — **Bleed-through removal, Bayesian Inference, Document Restoration**

I Introduction

The use of documents as sources is central to historical study, however poor legibility due to progressive degradation is an issue often encountered. Many libraries host large collections of manuscripts which are especially vulnerable to such degradations due to the fragile nature of the media on which they were written. Physical restoration of degraded documents is a cost and time intensive process, and may affect the integrity of the original. Restoration methods using automatic image processing techniques therefore have become increasingly popular as they have the advantage of being able to make any number of alterations to the document appearance, whilst leaving the original intact.

Loss of textual information in documents may be classified into four main categories: (i) Fading of text over time due to light exposure or flaking ink. (ii) Obscured or missing text due to degradation of the writing medium. This may be caused by damp, mould, parasites, or the inherent brittleness of the medium. (iii) Bleed-through interference, where ink has seeped through from one side of a page to the other and reduces legibility. (iv) The digitisation process itself can introduce noise artifacts and degrade the textual information further. Examples of such issues are non stationary noise due to variable illumination [1], show-through caused by the scanning of double sided documents [2], and warped images as a result of of digitising documents with inherent curvature, for example due to binding.

This paper focuses on the problem of Bleed-through interference and proposes a bleed-through removal method based on a simple linear degradation model. The rest of this paper proceeds as follows. Section II discusses previous works on bleed-through removal; Section IV introduces the degradation model; Sections V and VI detail the Bayesian framework for and solution of the problem; Sections VII and VIII provide some initial results obtained and a brief discussion.

II Related Work

The area of digital document restoration, and more specifically bleed-through removal has become increasingly popular over the past decade. Approaches to bleed-through removal can be considered as belonging to one of two categories; blind and non-blind. Blind methods work with one side of the document only, whereas non-blind methods work with registered recto and verso sides, assuming that images of both are available.

Due to the relative lack of image data, blind methods often involve an intensity based thresholding step, such as hysteresis thresholding in [3], and the recursive unsupervised classification method of [4]. Tonazzini et al. [5] use independent component analysis (ICA) to separate the RGB colour channels of a single image into foreground, background, and bleed-through classes. More recently Tonazzini in [6] combines this method with different colour space representations of manuscript images for bleed-through removal and information content maximisation. Wolf in [1] addresses the problem via a dual-layer Markov Random Field (MRF), with two hidden label fields and one observation field.

Non-blind methods can exploit more data, and manifest as two stage processes. In the first stage recto and verso sides must be registered so that they are aligned and of the same resolution (this is not a trivial problem), then the second stage comprises the bleed-through removal. Tonazzini et al. extended their ICA technique to include grayscale images in [7], using the flipped verso image as one of the sources. The methods contained in [8] and [9] focus purely on early music documents. Castro et al. [8] use a combination of Sauvola's thresholding algorithm [10] and fuzzy classification, while [9] proposes extensions to two binarisation methods, namely symmetric/non-symmetric Kullback-Leibler (KL) thresholding algorithms and the binarisation algorithm of Gatos et al. [11], by adding in a second threshold level for the bleed-through interference. Huang et al. in [12] and [13] propose a framework for user assisted bleed-through reduction that takes a small set of user input training data for the background, foreground text, and bleed-through text of both recto and verso pages, and uses this to locate and remove bleed-through interference. Moghaddam et al. [14] use diffusion models for the recto and verso texts, and the background medium, and apply reverse diffusion to remove interference. More recently, they apply the diffusion methods in a unified framework, [15], using variational models for blind, non-blind, and severe bleed-through removal.

The linear model presented in this work is most similar to that of Tonazzini et al. [7]. However, we do not assume constant mixing parameters across the images, nor that the extent of bleed-through is the same on both sides, nor that bleed-through can interfere with foreground text. In addition our framework is simpler and unified under a single Bayesian framework.

III Alignment/Registration

As discussed above, the recto and verso sides of the page are first registered so that the bleed-through text on one side is aligned with the originating text on the other. To do this the method proposed by Dubois et al. in [16] is modified by using a set of user selected points (minimum 3) to initialise the registration process. The user selects corresponding control points on both recto and verso images that indicate locations of the same textual features. The initialisation is an affine model derived from a least squares fit to these locations. In the experiments performed this relatively simple step improves the computation time of the algorithm and also yields much better alignment. In what follows *verso* refers to the flipped and registered image of the original un-aligned verso side.

IV Degradation Model

In the proposed model, it is assumed that the intensity of an observed (degraded) recto pixel $I_r(h,k)$, at location (h,k) in the image, is a linear combination of the clean image pixels $Y_r(h,k)$, $Y_v(h,k)$ from the recto and verso sides respectively. The combination is controlled by $\alpha_v(h,k)$, a mixing parameter and masks defined on both sides, $M_r(h,k), M_v(h,k)$. The model for each side is as follows (where we discard pixel coordinates for brevity).

$$\begin{aligned} I_r &= Y_r + M_r(1-M_v)\alpha_v Y_v + \rho \\ I_v &= Y_v + M_v(1-M_r)\alpha_r Y_r + \nu \end{aligned} \tag{1}$$

$\alpha_r,\ \alpha_v$ control the amount of bleed-through from the one side to the other. M_r and M_v here are binary masks that have value 0 where the corresponding image is foreground text, and 1 everywhere else. The noise terms ρ, ν are Gaussian $\mathcal{N}(0, \sigma^2_{\rho\rho})$, $\mathcal{N}(0, \sigma^2_{\nu\nu})$. The combination of the two mask terms ensures that bleed-through cannot interfere with foreground text, as can be seen to be the case in most bleed-through degraded documents.

V Bayesian Framework

In a Bayesian fashion the estimation of the parameters $\boldsymbol{\theta} = [\alpha_v, \alpha_r, M_v, M_r, Y_r, Y_v]$ proceeds using a Maximum A Posteriori (MAP) framework. The p.d.f. of the variables given the observed, registered data I_r, I_v, at a single location is then as follows.

$$p(\boldsymbol{\theta}|I_r, I_v, \tilde{M}, \tilde{\alpha}) \propto p(I_r, I_v|\boldsymbol{\theta}, \tilde{M}, \tilde{\alpha})p(\boldsymbol{\theta}|\tilde{M}, \tilde{\alpha}) \tag{2}$$

where $\tilde{M}, \tilde{\alpha}$ represent the existing state of the mask and linear mixing parameter (on the recto and verso sides as appropriate) in the neighbourhood of the pixel site currently being considered. The various likelihood and prior distributions are presented next.

a) Likelihood

Following the degradation model, the likelihood combines the influence of both the recto and verso sides to yield another Gaussian distribution (it is assumed that the two sides are independent).

$$p(I_r, I_v|\boldsymbol{\theta}, \tilde{M}, \tilde{\alpha}) \propto$$
$$\exp - \left\{ \frac{1}{2\sigma^2_{\rho\rho}}(I_r - Y_r - M_r(1-M_v)\alpha_v Y_v)^2 \right.$$
$$\left. + \frac{1}{2\sigma^2_{\nu\nu}}(I_v - Y_v - M_v(1-M_r)\alpha_r Y_r)^2 \right\} \quad (3)$$

b) Priors

In the usual manner, it is clear that masks and mixing parameters should be spatially smooth in some local area, and so Gibbs energy priors are employed for this purpose. These priors are defined as follows (for M_r and α_v here).

$$p(M_r|\tilde{M}) \propto \exp - \left\{ \beta_r(1-M_r) + \sum_{\tilde{M}}(M_r - \tilde{M})^2 \lambda_M \right\}$$
$$p(\alpha_v|\tilde{\alpha}) \propto \exp - \left\{ \sum_{\tilde{\alpha}}(\alpha_v - \tilde{\alpha})^2 \lambda_\alpha \right\}$$

λ_M and λ_α are smoothness weights for the masks and mixing parameters respectively, empirically set to $\lambda_M = 1$, $\lambda_\alpha = 55$ in what follows. $\tilde{\alpha}, \tilde{M}$ are values of the variables in the neighbourhood (8-connected) of the current site. β_r is used to prevent the estimated masks from *leaking* into regions that are definitely non-textual (i.e. background), it is configured at the initialisation stage (discussed below) by k-means clustering of the image patch. An Ising prior is used for the mask variables but a GMRF for the mixing parameters since these are continuous.

VI Solution

To solve for all the variables, iterated conditional modes (ICM) optimisation [17] is used. A slightly modified version is adopted here in that we draw *samples* for the underlying clean images Y, while selecting the mode of the conditionals for the remaining variables. The process is clearly iterative and we solve for $M_r, M_v, \alpha_r, \alpha_v, Y_r, Y_v$ in turn. The required expressions are discussed next.

a) Mixing Parameter Estimate

Each mixing parameter is present in only one of the observation terms, therefore the conditional probability $p(\alpha|\cdot)$ at a site is as follows (for α_v).

$$p(\alpha_v|M_v, M_r, Y_r, Y_v, I_r, I_v) \propto$$
$$\exp - \left\{ \frac{1}{2\sigma^2_{\rho\rho}}(I_r - Y_r - M_r(1-M_v)\alpha_v Y_v)^2 + \right.$$
$$\left. \sum_{\tilde{\alpha}}(\alpha_v - \tilde{\alpha})^2 \lambda_\alpha \right\} \quad (4)$$

Again, the expressions are similar for α_r since the problem is symmetric. Using ICM, an estimate of α_v at that site is obtained analytically since the expression is quadratic in α_v. Hence

$$\hat{\alpha}_v = \frac{2\sigma^2_{\rho\rho}\sum_{\tilde{\alpha}}\tilde{\alpha}\lambda_\alpha + (I_r - Y_r)M_r(1-M_v)Y_v}{2\sigma^2_{\rho\rho}\sum_{\lambda_\alpha}\lambda_\alpha + (M_r(1-M_v)Y_v)^2} \quad (5)$$

b) Mask Estimate

The estimate for the masks is generated using the following conditional at a pixel site (for M_r).

$$p(M_r|\alpha_r, \alpha_v, M_v, Y_r, Y_v, I_r, I_v) \propto$$
$$\exp - \left\{ \frac{1}{2\sigma^2_{\rho\rho}}(I_r - Y_r - M_r(1-M_v)\alpha_v Y_v)^2 \right.$$
$$+ \frac{1}{2\sigma^2_{\nu\nu}}(I_v - Y_v - M_v(1-M_r)\alpha_r Y_r)^2$$
$$\left. + \sum_{\tilde{M}}(M_r - \tilde{M})^2 \lambda_M + \beta_r(1-M_r) \right\} \quad (6)$$

In this case the estimation is performed numerically since M_r is binary. Hence both $M_r = 0, 1$ are substituted in the expression above and whichever yields the greater probability is selected.

c) Clean Image Estimate

There are no priors for Y_r, Y_v and hence the conditionals are derived directly from the likelihood, Eq.(3), to yield the expression below (for Y_r).

$$p(Y_r|\alpha_r, \alpha_v, M_r, M_v, Y_v, I_r, I_v) \propto$$
$$exp - \left\{ \frac{1}{2\sigma^2_{\rho\rho}}(I_r - Y_r - M_r(1-M_v)\alpha_v Y_v)^2 \right.$$
$$\left. + \frac{1}{2\sigma^2_{\nu\nu}}(I_v - Y_v - M_v(1-M_r)\alpha_r Y_r)^2 \right\} \quad (7)$$

The expression for the verso side similar. This distribution is clearly Gaussian, however instead of maximising the conditional as part of the usual ICM process, a sample is drawn from this distribution within $\pm T$ standard deviations of the mean. This is a strategy used to good effect by other authors working in video and audio restoration [18]. The idea is that using the mean tends to generate oversmooth images, while using an unconstrained random draw is visibly chaotic. By drawing samples within some distance of the mean, a textural component in the underlying signal is allowed for and the iterative process itself performs better.

The required draw is therefore $Y_r \sim \mathcal{N}(\bar{Y}_r, \sigma^2_Y)$. By completing the square in the conditional above the mean and variance are extracted as follows.

$$\bar{Y}_r =$$
$$\frac{I_r - M_r(1-M_v)\alpha_v Y_v + \sigma^2(I_v - Y_v)M_v(1-M_r)\alpha_r}{1 + \sigma^2(M_v(1-M_r)\alpha_r)^2} \quad (8)$$

where $\sigma^2 = \sigma^2_{\rho\rho}/\sigma^2_{\nu\nu}$

$$\sigma^2_Y = \frac{\sigma^2_{\rho\rho}\sigma^2_{\nu\nu}}{\sigma^2_{\nu\nu} + M_v(1-M_r)\alpha_r\sigma^2_{\rho\rho}} \tag{9}$$

The estimate for σ^2_Y clearly depends on estimates for $\sigma^2_{\rho\rho}, \sigma^2_{\nu\nu}$. In the proposed algorithm these are generated by measurement from the data. As is typical, these variance estimates tend to be overestimates. It was found that empirically adopting a different univariate sample for Y_r improves the convergence of the algorithm by reducing the overestimation effect i.e. $\tilde{Y}_r = \bar{Y}_r - \sqrt{|\mu|\sigma^2_Y}$ where $\mu \sim \mathcal{N}(0,1)$.

d) Initialisation

Good initial estimates are required for ICM to converge usefully, see Figure 1. In this work the masks are initialised using k-means clustering on the intensity channel of the observed images, using 2 or 3 clusters, dependent on the severity of the bleed-through. Assuming 3 clusters are used, the cluster with the maximum brightness is considered to be background region where there is no bleed-though. The minimum brightness cluster is then the foreground text, with the remaining cluster an initial estimate of the bleed-through. The initial mixing parameters are then obtained from the mask estimates and the observed images as follows (for α_v). Given a binary mask B coincident with the darkest two clusters from the k-means process, β is configured as $100(1-B)$. The masks and observed images are also used to estimate the noise variance for both recto and verso sides, based on the variance of regions where $B = 0$. Finally initial clean recto and verso estimates are obtained by substituting the relevant initial estimates into Eq(1).

```
At each pixel location
$\alpha_v = 0$
**if** recto is background ($M_r = 1$) **then**
    **if** verso is foreground ($M_v = 0$) **then**
        **if** recto intensity is > recto background average intensity ($b_r$) **then**
            $\alpha_v = \frac{recto - b_r}{b_r}$
        **end if**
    **end if**
**end if**
```

e) Algorithm

The final algorithm may be enumerated as follows

1. Register the recto and verso sides of the image using the process outlined in Section III.
2. Initialise all variables as described above
3. Using a checkerboard visitation pattern for sites
   (a) Generate $\hat{M_v}$, $\hat{M_r}$, $\hat{\alpha_r}$, $\hat{\alpha_v}$ using the expressions above, (across all sites in separate image passes) updating in place.
   (b) Draw samples for Y_r, Y_v as above
4. Terminate for iterations = 20.
5. Goto 3

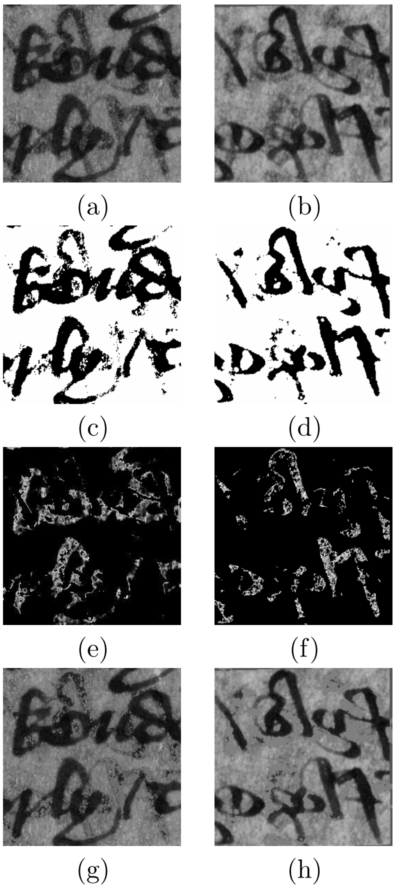

Fig. 1: (a),(b) degraded recto and verso images of the Piers example, (c),(d) initial mask estimates using k-means clustering ($k = 3$), (e),(f) initial estimates of α_r and α_v (white represents 1), (g),(h) initial clean recto and verso estimates.

VII Results

Small 255x255 patch pairs from larger high resolution (600dpi) images were extracted for testing. The manuscript images used were from *Piers Plowman* 'B' text from the late 14th century, fol.22, MS.201, Corpus Christi Library, Oxford (*Piers*), and a Welsh dictionary from 16th/17th Century, fol.1, MS.16 Jesus College Library, Oxford (*Welsh*). Figure 1(a,b) show the recto and verso sides respectively for Piers, and Figure 3(a,b) show the Welsh samples. These images are already registered using the technique outlined in Section III. Initial parameter estimates are also shown in Figure 1. The initialization for the clean images show some bleed-through removal already but the initialization for α is far from optimal and certainly too active.

Figure 2 shows output over 15 iterations on Piers. As can be seen the bleed-through is reasonably well removed and the mask estimates match quite well to the area of the foreground text on the recto and verso sides. The mixing parameters also coincide well with the non-binary nature of the bleed-through. Figure 3 shows two iterations from a sequence of 5 on Welsh. The bleed-through is, again, reasonably removed.

In both examples the estimated image is more noisy than the original. This is because at each iteration a *sample* of the underlying clean image is generated and *not* some optimal estimate. These estimates could be averaged after some *burn in* period to generate the MMSE estimate for the clean image, but the samples here are shown to give the reader an impression of the actual estimates in use by the algorithm.

VIII Conclusion

This paper has presented a relatively simple algorithm for bleed-through removal that relies on estimates generated from local pixel neighbourhoods in an ICM scheme. The process converges and does yield bleed-through removal. However, in the Piers example the verso image is certainly more blurry than the recto image and our degradation model does not model this sharpness difference. This means that the spatial smoothness on both sides is not the same and that could influence estimation detrimentally. Furthermore it is noted that the algorithm does not perform optimally in cases where the bleed-through regions are much larger than the originating text due to the porosity of the medium (as in the Welsh verso example). These issues are being addressed currently

Acknowledgment

This research has been funded by the Irish Research Council for Science, Engineering and Technology.

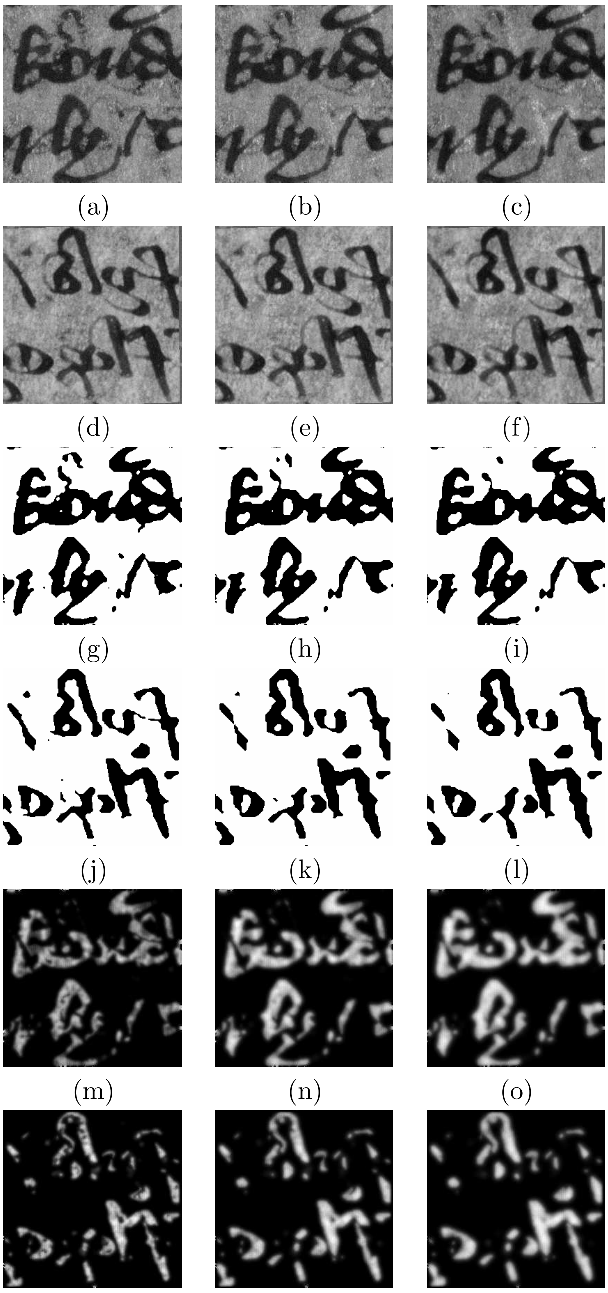

Fig. 2: Iterations 5, 10, 15 for the Piers example. (a-c) Y_r, (d-f) Y_v, (g-i) M_r, (j-l) M_v, (m-o) α_r, (p-r) α_v. For iteration 15, the estimate for Y_r was obtained by setting $M_r = 1$ and $M_v = 0$ to increase smoothness on the final result, similarly for Y_v.

References

[1] 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*, 32(3):431–447, 2010.

[2] G. Sharma. Cancellation of show-through in duplex scanning. In *Image Processing, 2000. Proceedings. 2000 International Conference on*, volume 2, pages 609–612 vol.2, 2000.

[3] R. Estrada and C. Tomasi. Manuscript bleed-through removal via hysteresis thresholding. In *Document*

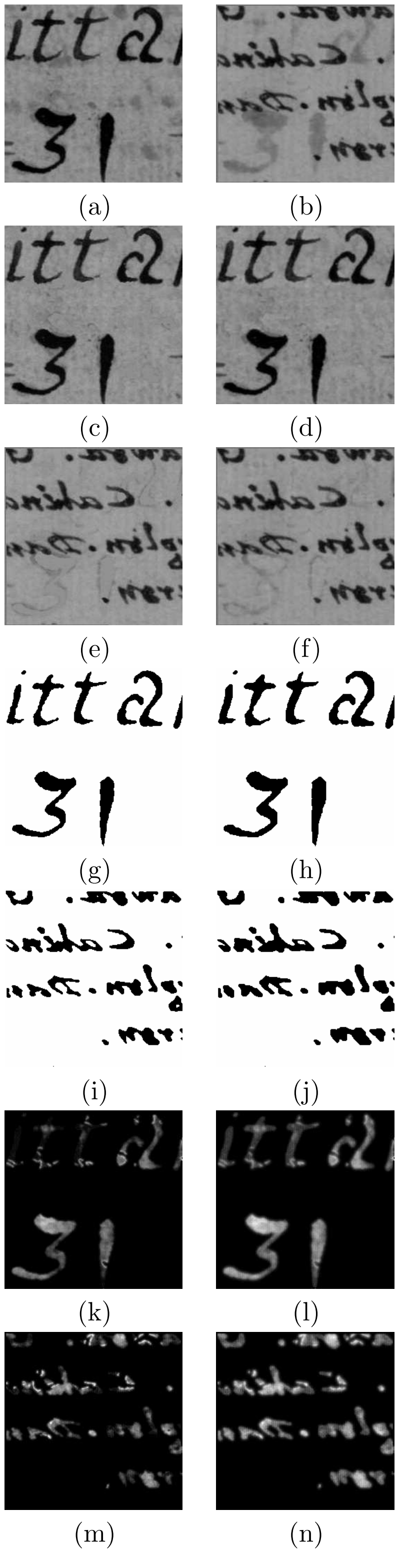
(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)
(i)
(j)
(k)
(l)
(m)
(n)

Fig. 3: Iterations 2,5 for the Welsh example (a) I_r, (b) I_v, (c-d) Y_r,(e-f) Y_v,(g-h) M_v, (i-j) M_v, (k-l) α_r, (m-n) α_v.

Analysis and Recognition, 2009. ICDAR '09. 10th International Conference on, pages 753–757, 2009.

[4] F. Drira, F. Le Bourgeois, and H. Emptoz. Restoring ink bleed-through degraded document images using a recursive unsupervised classification technique. In Horst Bunke and A. Spitz, editors, *Document Analysis Systems VII*, volume 3872 of *Lecture Notes in Computer Science*, pages 38–49. Springer Berlin / Heidelberg, 2006.

[5] A. Tonazzini, L. Bedini, and E. Salerno. Independent component analysis for document restoration. *International Journal on Document Analysis and Recognition*, 7(1):17–27, 2004.

[6] A. Tonazzini. Color space transformations for analysis and enhancement of ancient degraded manuscripts. *Pattern Recognition and Image Analysis*, 20(3):404–417, 2010.

[7] A. Tonazzini, E. Salerno, and L. Bedini. Fast correction of bleed-through distortion in grayscale documents by a blind source separation technique. *International Journal on Document Analysis and Recognition*, 10(1):17–25, 2007.

[8] P. Castro, R. J. Almeida, and J. R. C. Pinto. Restoration of double-sided ancient music documents with bleed-through. In *Progress in Pattern Recognition, Image Analysis and Applications*, volume 4756 of *Lecture Notes in Computer Science*, pages 940–949. Springer Berlin/Heidelberg, 2007.

[9] J. A. Burgoyne, J. Devaney, L. Pugin, and I. Fujinaga. Enhanced bleedthrough correction for early music documents with recto-verso registration. In *International Conference on Music Information Retrieval*, pages 407–412, Philadelphia, USA, 2008.

[10] J. Sauvola and M. Pietikinen. Adaptive document image binarization. *Pattern Recognition*, 33(2):225–236, 2000.

[11] B. Gatos, I. Pratikakis, and S. J. Perantonis. Adaptive degraded document image binarization. *Pattern Recognition*, 39(3):317–327, 2006.

[12] Y. Huang, M. S. Brown, and D. Xu. A framework for reducing ink-bleed in old documents. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–7, 2008.

[13] Y. Huang, M. S. Brown, and D. Xu. User-assisted ink-bleed reduction. *Image Processing, IEEE Transactions on*, 19(10):2646–2658, 2010.

[14] R. F. Moghaddam and M. Cheriet. Low quality document image modeling and enhancement. *International Journal on Document Analysis and Recognition*, 11(4):183–201, 2009.

[15] R. F. Moghaddam and M. Cheriet. A variational approach to degraded document enhancement. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(8):1347–1361, 2010.

[16] E. Dubois and A. Pathak. Reduction of bleed-through in scanned manuscript documents. In *IS&T Image Processing, Image Quality, Image Capture Systems Conference (PICS2001)*, volume 4, pages 177–180, Montreal, Canada, 2001.

[17] J. Besag. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society. Series B (Methodological)*, 48(3):259–302, 1986.

[18] J. K. O Ruanaidh and W J Fitzgerald. *Numerical Bayesian Methods Applied to Signal Processing.* Statistics and Computing. Springer, 1996.