Noname manuscript No.
(will be inserted by the editor)

Blind Bleed-through Removal in Color Ancient Manuscripts

Muhammad Hanif · Anna Tonazzini · Syed Fawad Hussain · Usman Habib ✉ · Emanuele Salerno · Pasquale Savino · Zahid Halim

Received: date / Accepted: date

Abstract Archaic manuscripts are an important part of ancient civilization. Unfortunately, such documents are often affected by various age related degradations, which impinge their legibility and information contents, and destroy their original look. In general, these documents are composed of three layers of information: foreground text, background, and unwanted degradation in the form of patterns interfering with the main text. In this work, we are presenting a color space based image segmentation technique to separate and remove the bleed-through degradation in digital ancient manuscripts. The main theme is to improve their readability and restore their original aesthetic look. For each pixel, a feature vector is created using color spectral and spatial location information. A pixel based segmentation method using Gaussian Mixture Model (GMM) is employed, assuming that each feature vector corresponds to a Gaussian distribution. Based on this assumption, each pixel is supposed to be drawn from a mixture of Gaussian distribution, with unknown parameters. The Expectation-Maximization (EM) approach is then used to estimate the unknown GMM parameters. The appropriate class label for each pixel is then estimated using posterior probability and GMM parameters. Unlike other binarization based document restoration method where the focus is on text extraction, we are more interested in restoring the aesthetically pleasing look of the ancient documents.The experimental results validate the usefulness of proposed method in terms of successful bleed-through identification and removal, while preserving foreground-text and background information.

Keywords Bleed-through · segmentation · Gaussian mixture model · color space

This work was supported by ERCIM.

Muhammad Hanif[1,2,✉]
E-mail: muhammad.hanif@giki.edu.pk

Anna Tonazzini[1]
E-mail: anna.tonazzini@isti.cnr.it

Syed Fawad Hussain[2]
E-mail: fawadsyed@gmail.com

Emanuele Salerno[1]
E-mail: Emanuele.Salerno@isti.cnr.it

Usman Habib[2]
E-mail: Usmanhabib@giki.edu.pk

Pasquale Savino[1]
E-mail: Pasquale.Savino@isti.cnr.it

Zahid Halim[2]
E-mail: zahid.halim@giki.edu.pk

1 Istituto di Scienza e Tecnologie dell'Informazione, Consiglio Nazionale delle Ricerche, Via G. Moruzzi, 1, I-56124 PISA, Italy.
2 Faculty of Computer Science and Engineering, Ghulam Ishaq Khan Institute of Engineering Sciences and Technology, Topi 23640 Pakistan.

1 Introduction

Archival, ancient manuscripts are an important source of our historical and cultural heritage. These documents preserve our legacy from the past, what we live with today, and what we pass on to the future generations. These manuscripts and documents are an important resource of knowledge for scholars and historians, as they provide insight into the culture, lifestyle, and civilization of the distant past. Recently, digital preservation of historical documents has been the focus of intensive digitization and archiving campaigns, aimed at fostering their distribution, accessibility, and study. Digitization makes ancient documents available for public use and analysis, while keeping the original safe. Although, high-resolution images of the documents are

often available, the quality of these scanned copies mainly depends on the current status of the original documents. Unfortunately, in most cases they are severely affected by different types of degradations, occurred due to aging, improper handling, and environmental factors. These physical degradations also appear on the digitized images, thus impairing their legibility and interpretation. As an example, in the manuscripts written on both sides of the sheet, often the ink has seeped through, and appears as an unpleasant interference on the other side. This degradation, termed as bleed-through, is mainly due to aging, humidity, ink chemical properties or paper porosity [1].

In past, different chemical restoration techniques were applied to remove or reduce the physical degradations in ancient manuscripts and improve their readability, but unfortunately they are invasive, and can cause permanent damage to the original document. Nowadays, the virtual restoration of these documents using digital image processing techniques is of great interest, with the advantage of allowing to perform any number of alterations to the document image while keeping the original document intact. In addition to improve the document readability, removal of degradations is also a critical preprocessing step in many tasks such as optical character recognition (OCR), feature extraction, segmentation, and automatic transcription.

In this paper, we propose a color based segmentation method for detecting and then removing bleed-though patterns from ancient RGB manuscripts. The proposed method exploits different color spaces to highlight the differences in color profile of the various information layers present in document images. We consider color uniformity for partitioning a document image into disjoint layers (e.g., foreground text, background, and degradation) without any supervised learning. The proposed method is based on the assumption that each layer of information is defined by a homogeneous color and give rise to separated clusters in the color space. In other words, each cluster defines a class of pixels that share similar color properties. In most cases, fewer colors are present in ancient documents than in natural images, making it an easy case for color based segmentation with comparatively high color homogeneity. However, simple global intensity thresholding cannot be used for segmentation due to the high local variability of the degradation [2].

As in color based segmentation the results heavily depend on the color space, there is no single color space that can provide acceptable results for all kind of images [3]. In the literature, the choice of particular color space usually depends on the problem in hands. In this paper, we propose to perform color segmentation by jointly using different classical color spaces, including RGB, LUV and LAB, so that maximum information is conveyed to the segmentation process. At each pixel, the different color representations, along with pixel spatial location, are stacked to form a feature vector. The whole data cube is then used to perform a pixel based segmentation, where each cluster represents a class of pixels that share similar color properties. Assuming a Gaussian distribution for each class, a Gaussian mixture model (GMM) is estimated using the expectation maximization (EM) algorithm. For each pixel, the appropriate class label is then estimated using the GMM parameters. The bleed-through pixels, identified through clustering, can then be removed by inpainting them with the appropriate background texture.

In most of the ancient document restoration methods very little attention is paid to the inpainting part. Usually, a random fill-in [4] or a substitution with the average value of the surrounding pixels [5] is suggested for the inpainting of the degraded pixels, but these kinds of pixels fill-ins result in visual imprints especially when the size of the degraded patch is large. In this work we pay special attention to the inpainting step, due to its significant effects on the quality of final restored image. Thus, in order to preserve the original look of the document and reproduce the background texture, in the proposed method the detected bleed-through pixels are inpainted based on Gaussian conditional simulation, where the surrounding context is accounted for.

The rest of the paper is organized as follows. Next section briefly presents the state of the art in the field of bleed-through removal and color image segmentation. In Section 3 we present the proposed clustering method and discuss the Gaussian texture inpainting. A set of experimental results and their quantitative evaluation is presented in Section 4. Concluding remarks are given in Section 5.

2 Related Work

2.1 Bleed-through Removal

In literature, a large number of algorithms have been proposed to remove different kind of distortions from the digital copies of ancient degraded documents. However, most of these methods depend on a certain context of use and are intended to address a precise type of degradation. Among other document degradations, bleed-through is particularly challenging due to its significant overlap with the original text and the wide variation of its extent and intensity. Bleed-through removal is usually addressed as a classification problem, where the image pixels are labeled as either background (medium),

bleed-through (noise), or foreground (original text) [1]. The large variety of methods proposed to address the bleed-through removal problem can be classified into two main categories: non-blind and blind. The non-blind methods utilize information from both sides of the page and therefore require their accurate registration, which is still an open research issue. Whereas, blind methods only require the image of one side at hand. Most of the blind methods rely on the intensity profile of the degraded document image and perform restoration based on the intensity distributions, involving thresholding. However, intensity information alone is insufficient due to the significant overlap between the bleed-through and original text intensities [6]. A comparison of several thresholding techniques to differentiate text and background in degraded ancient document images is presented in [7]. It was concluded that neither global nor local thresholding techniques perform satisfactorily. Recently, a semi-adaptive document binarization technique is presented in [8]. In [9], a blind method is outlined for multichannel images, using color decorrelation. A recursive segmentation method on the data, first decorrelated with principal component analysis (PCA), is presented in [1], where the document image is divided into two clusters: original text or other. In [10] a non-linear reaction–diffusion model is suggested for binarization of ancient documents effected by bleed-through degradation. A blind source separation approach is presented in [11], where independent component analysis (ICA) is used to separate the background, foreground and bleed-through layers in a color image. A Markov random field (MRF) based local smoothness model is used in [12], whereas a dual-layer MRF is suggested in [5], posing bleed-through removal as an image segmentation problem. A recursive unsupervised segmentation is suggested in [13], and an expectation maximization (EM) based approach is presented in [12]. Recently, a conditional random field (CRF) based method, using intensity distribution as prior, has been presented in [4]. In the non-blind category, a model based method using the difference in the intensities of recto and verso sides is presented in [14]. An extension of the same model using a variational approach and spatial smoothness in the wavelet domain is outlined in [15]. An ICA based method for grayscale document images is presented in [16]. Other methods in this category are reported in [17] and [18]. As mentioned earlier, the performance of non-blind methods depends on the registration of the recto and verso images. However, perfect registration is difficult to achieve due to document skews, different image resolutions, or wrapped pages during scanning, e.g., books.

In recent years, deep learning methods have emerged as state-of-the-art for historical document image binirization. The majority of these learning methods use the frame work of directly pixel classification, mostly in two classes. For document images, the two classes are mostly the background (white) and the foreground text (black). These deep learning binarization methods can be used to address some specific degradations in document images [19]. The convolution neural networks (CNN) is used for character recognition and large scale image segmentation [20]. A transformed version of CNN, fully convolutional network (FCN) addressed the problem of binarization as semantic segmentation, where the pixels are classified as text or background [21]. Other deep learning based segmentation methods for document images are suggested in [22][23][24]. One of the main issue with these methods is the requirement of labelled data for training [19]. In most cases, data labeling is not easy, especially for ancient documents where the quality of images are degraded with the passage of time. Secondly, the deep learning methods are mostly used for binarization, to separate text from background. With binarization alone, we cannot restore the document image in its original look.

In addition to bleed-through identification, finding a suitable replacement for the degraded pixels is also essential for a plausible restoration. The restored image in most of the above mentioned methods is either binary, pseudo-binary, or textured, i.e. the bleed-through pixels are replaced with local background mean or with random values. In [5] an estimate of the local mean background is used, but produces visible artifacts for documents with a highly textured background. A random-fill inpainting method is suggested in [4] to replace the bleed-through pixels with randomly selected pixels from the neighborhood. However, the random pixel selection produces salt and pepper like artifacts in regions with large bleed-through. In [17], as a preliminary step, a “clean” background for the entire image is estimated, but this is usually a very laborious task. Recently, a sparse representation based inpainting is used in [25] for a plausible bleed-through free image. The appropriate inpainting for the removed bleed-through pixels is important to restore the primordial look of the document.

2.2 Color image segmentation

Color space based segmentation is a critical problem in image analysis and computer vision, with a variety of applications. Image segmentation, defined as pixel classification in an image, simplify and/or change

the image representation for a meaningful and easier understanding. The segmentation process uses local image features, such as color information, boundaries, edges or textures, to form meaningful clusters, such that pixels in each cluster share certain characteristics. In color based segmentation, the clustering depends on the particular color space used and the homogeneity criterion is based on features derived from the spectral components. By definition, a color space is a tool to visualize, create, and specify the color [26]. The choice of a particular color space for image segmentation is delicate: several color spaces, such as RGB, HIS, and CIELuv have been used, but none of them is optimally suited in the same way for all kinds of images [3].

The most commonly used color space is RGB, where colors are represented in terms of red, green, and blue spectral components in an orthogonal Cartesian space [26]. However, for higher level tasks like segmentation, RGB color space fails to mimic the color perception of the human visual system, where the color is better represented in terms of hue, saturation, and intensity [27]. In addition, the R, G and B channels are highly correlated, making RGB a poor choice for segmentation. The HSI space, obtained from RGB, overcomes these problem, but both RGB and HSI are not perceptually uniform, meaning that the difference in color perceived by the human eye is not represented by similar distance in both spaces. The CIE color spaces (CIELuv, CIELab) introduce a uniform metric for assessing perceptual differences among colors using Euclidean distance. Comparatively, the CIELab color space performs better than CIELuv in image segmentation [28]. A cluster based image segmentation method that compares the performance of different color spaces is presented in [29], where the CMY color space is suggested to perform at best. A hybrid color space based approach is suggested in [30]. Using morphological analysis of color histogram, a segmentation algorithm is presented in [31], where the authors conclude that segmentation results are similar for five different color spaces. In [32], a comparison of ten color spaces for skin detection and segmentation is presented. The authors conclude that the HSV color space produces the best results. Similarly, in [33], the HSV color space is recommended for crop image segmentation. In [26] the authors visually compare segmentation results of eight different color spaces and conclude that CIELab and CIELuv produce better results. An adaptive color space selection, based on spectral color analysis, is proposed in [27], where the color space is selected according to the task at hand. Finally, a color quantization clustering method, using eigenvalues of the color covariance matrix, is suggested in [34].

From the above discussion, it is clear that different authors draw contradictory conclusions regarding the best color spaces to be used in the context of color image segmentation. In this paper, instead of searching for the best color space, we combined three different color spaces along with the spatial information of pixels to form compact clusters of pixels with similar color features, assuming that homogeneous colors in the image correspond to separate clusters. As mentioned ealier, this is likely to be a successful choice for segmenting color document images, as the variation in the color profile in different color spaces can used to exploit the difference in foreground text, bleed-through degradation and background. Moreover, in document images usually a limited set of colors are used, which make it an easy case for color based segmentation.

3 The Proposed method

The color at each image pixel can be defined by a vector $D = [d_1, d_2, d_3, ..., d_N]^T$ whose dimension N depends on the color space used, e.g., $N = 3$, $D = [R, G, B]$, for the RGB color space, or $N >> 3$ in multispectral or hyperspectral images. The color vector coordinates can be considered as a realization of a multivariate Gaussian law and a color image as a realization of a random field, where each random variable is described by a Gaussian Mixture Model (GMM) [35]. The GMM is an efficient statistical model for representing a population composed of K sub-populations, and maximum likelihood estimation via the expectation-maximization (EM) algorithm is a powerful tool to find parameters related to each sub-population in a GMM.

Recently, many studies suggested the combination of different color spaces for image segmentation tasks. Following this line, we used a combination of RGB, CIELab, and CIELuv, along with the pixel spatial location, to create a feature vector at each pixel. The CIELab and CIELuv color spaces represent perceptual uniformity and meet the psychophysical need for color based segmentation by a human observer. They are obtained through a non-linear transformation of the XYZ color space, and have an approximately uniform chromaticity scale, *i.e.*, they match the sensitivity of the human eyes. Both CIE spaces share the same L value, which defines the lightness or the intensity of a color. In our hybrid color space, composed of three color spaces, the direct color comparison can be performed based on the geometric separation within the color space. It is especially efficient in measuring even small color differences. The spatial smoothness constraint enforced by the pixel spatial information is par-

ticularly suited to describe the classes considered in our case (mainly paper texture and handwritten texts).

As mentioned earlier, the main theme here is to restore the original aesthetical look of the ancient documents. After segmentation, document restoration is completed by removing the noisy bleed-though pixels and replacing them with a suitable background pixels. The pixel removal in this case is based on the pixel classification, only the pixels identified as bleed-through degradation are removed and replaced. To this end, a befitting replacement is estimated using Gaussian texture inpainting. The inpainting step incorporates the pixel neighboring information to preserve the natural look of the manuscript.

3.1 Pixel Segmentation

Consider the observed RGB image I as an unlabeled random sample, with M pixels, $I = (i_1, i_2, \ldots, i_M)$, drawn from an independent and identically distribution (iid). In this paper, unlabeled sample means that for any pixel $i_m \in I$ the true sub-population to which i_m belongs is not known. Our main goal is to make accurate statistical inferences on properties of the sub-populations by using the information of the unlabeled observed image I only.

We first transform the original R, G and B features of the color image I into a hybrid color space, including RGB, CIELab and CIELuv. In the color feature set, we retain only one L channel, since it is the same in the CIELab and CIELuv color spaces, thus obtaining the eight distinct color features R, G, B, L, U, V, A and B. We then create a feature space S, where each pixel is associated with a feature vector containing eight color values, from three color spaces, and two spatial location information, namely x and y. As mentioned earlier, the latter are included to enforce a spatial smoothness constraint for the classes present in manuscript images, usually made of a background (paper texture) plus handwritten texts.

Let $s_m = (s_m^R, s_m^G, s_m^B, s_m^L, s_m^U, s_m^V, s_m^A, s_m^{B*}, s_m^x, s_m^y)$ be the m^{th} vector in our feature space S, where s_m^F represents the scalar value observed in the $F-th$ feature dimension. In general, the classification problem consists of finding K classes $C_1, C_2, ..., C_K$ of pixels whose feature vectors satisfy a given similarity criterion, such that every $s_m, m = 1, 2, .., M$, belongs to one of these classes and no s_m belongs to two classes at the same time, that is $\bigcup_{n=1}^{N} C_n = S$ and $C_l \bigcap C_j = \emptyset \; \forall \; l \neq j$.

One popular way to perform classification is to model the feature space S as a realization of a random field, each random variable described by a GMM [35]. More precisely, each $s_m \in S$ is supposed to be drawn from a Gaussian mixture of multivariate normal distributions given as

$$p(s|\Theta) = \sum_{j=1}^{K} \alpha_j \mathscr{G}_j(s|\theta_j) \tag{1}$$

where $\Theta = \{(\alpha_j, \theta_j)\}_{j=1,2,...,K}$, $\mathscr{G}_j(i|\theta_j)$ is a multivariate Gaussian probability density function with parameters $\theta_j = (\mu_j, \Lambda_j)$ representing the mean vector and the covariance matrix, respectively, and $\alpha_j \in \mathbb{R}$ is the mixing weight, satisfying $\sum_{j=1}^{K} \alpha_j = 1$.

Assuming that each image pixel is actually drawn from one of the K Gaussian components in the mixture of eq. (1), then the classification problem can be reformulated as finding the parameters of the K Gaussian densities of eq. (1) that best describe K homogeneous partitions of the pixels. In other words, for each i.i.d observation s_m, generated according to eq. (1), we assumed a hidden class label c_p, $p \in [1, 2, ..., K]$, indicating the density distribution in the mixture that generates it, and estimated as the maximizer of the following distribution

$$p(c_j|s_m) = \frac{p(c_j)p(s_m|c_j)}{p(s_m)} \qquad j = 1, 2, ..., K \tag{2}$$

given by the Bayes rule and representing the posterior probability that s_m belongs to class C_j.

To estimate the unknown parameter set Θ, i.e. $\theta_j = (\mu_j, \Lambda_j)$ and $p(c_j) = \alpha_j$, $j = 1, 2, ..., K$, associated with each cluster, along with probabilities (2) for each observation, we use the EM algorithm. The strength of EM algorithms is based on the concept of incomplete data and complete data, which allows to simplify parameter estimation through Maximum Likelihood. In our setting, S is considered as incomplete data, whereas $(S, p(c_j|s_m), j = 1, 2, ..., K, m = 1, 2, ..., M)$ represents the associated complete data, and the likelihood of the observed, incomplete data is the marginal over the hidden class labels of the joint distribution of the complete data.

EM is an iterative algorithm, where at each iteration it alternates between two steps, until convergence. In the first step, the E-step, by virtue of the Jensen inequality, the log likelihood of the incomplete data is equivalently substituted with the marginal over the hidden class labels of the log joint distribution of complete data, and this expectation is computed conditioned on the observed data S and the current estimates of the parameters Θ. In the M-step, the expectation is maximized to obtain an update of the parameters.

In our case, the E-step, given the current parameters Θ^{itr}, returns an estimate of the posterior distribu-

tion of eq. 2, in the form:

$$p(c_j|s_m)^{itr} = \frac{\alpha_j^{itr} p(s_m|c_j)^{itr}}{\sum_{l=1}^{K} \alpha_l^{itr} p(s_m|c_l)^{itr}}$$

where $p(s_m|c_j)^{itr}$ is the $j-th$ multivariate Gaussian distribution with parameters θ_j^{itr} computed in s_m, that is:

$$p(s_m|c_j)^{itr} = \frac{1}{\sqrt{(2\pi)det(\Lambda_j^{itr})}} \exp\left(-\frac{1}{2}(s_m-\mu_j^{itr})^T(\Lambda_j^{itr})^{-1}(s_m-\mu_j^{itr})\right)$$

In the M-step, based on the estimates $p(c_j|s_m)^{itr}$, the parameters are updated as follows

$$\alpha_j^{itr+1} = \frac{1}{M}\sum_{m=1}^{M} p(c_j|s_m)^{itr}$$

$$\mu_j^{itr+1} = \frac{\sum_{m=1}^{M} s_m p(c_j|s_m)^{itr}}{\sum_{m=1}^{M} p(c_j|s_m)^{itr}}$$

$$\Lambda_j^{itr+1} = \frac{\sum_{m=1}^{M} p(c_j|s_m)^{itr}(s_m-\mu_j^{itr+1})(s_m-\mu_j^{itr+1})^T}{\sum_{m=1}^{M} p(c_j|s_m)^{itr}}$$

The stopping criterion for the EM algorithm is set to either a maximum number of iterations or the difference between two successive iterations, i.e., $||\theta^{itr+1} - \theta^{itr}|| \leq \tau$, where τ is a small constant.

EM based methods require the initialization of the parameters. In many cases, a random initialization is suggested which sometime cause unstable solution and is highly time consuming. We used the K-means++ algorithm [13] to estimate the initial values of the GMM parameters θ. The number of clusters K is set in the start and optimally adjusted according to the pixel density, i.e., clusters with small number of pixels are merged with the nearest cluster. For large size images, an optimized version of the EM algorithm presented in [36] is used to speed up the clustering process.

After optimizing the parameters of the GMM and determining the posterior probabilities $p(c|s)$ of eq. (2), the maximizers of such probabilities are used to assign a label c_p, $p \in [1,2,...,K]$ to each pixel i_m in the RGB image I.

After successful classification, the user can decide what class of pixels (clean) to keep unaltered in the original RGB image I, and what class of pixels (degradation) to discard, i.e. to inpaint with the background class.

3.2 Gaussian texture inpainting

After clustering, the pixels belonging to the degradation class are removed and replaced by a befitting value simulating the background. In this paper, we treat the degraded pixels as missing or corrupted image regions, and estimate pertinent fill-in values, which are consistent with the known non-corrupted surrounding regions. In traditional ancient document restoration methods, the noisy pixels are either replaced by a constant value or a random pixel value is assigned from the neighborhood. Unfortunately, in case of documents with textured background, these generic fill-in creates visible artifacts and destroys the original look of the document. Keeping this in mind, we perform texture inpainting using Gaussian conditional simulation [37], inferring the surrounding context to preserve the natural look of the document. For a plausible inpainting, the missing values can be estimated as conditional sampling using the known pixel values on the boundary of the missing part.

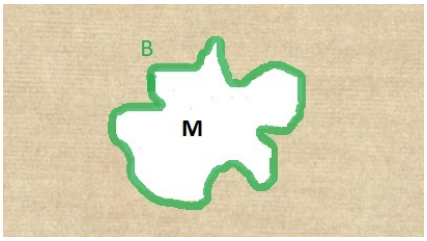

Figure. 1: Gaussian conditional inpainting: inferring the values in the mask M from the conditioning pixels in the set B located at the border of the mask.

The Gaussian conditional inpainting is especially efficient to inpaint textural content, as the case in many ancient documents background. Consider $I \in \mathbb{R}^2$ is the input image to inpaint. Let $M \in I$ represents the indices of missing pixels, i.e., pixels values of I are known except in the missing region M. Let ζ represent the outer border of the missing region M, with width w pixels as shown in Fig. 1. we wish to generate plausible values in M given a conditioning set ζ. The main idea here is to fill the missing region with a conditional sample of a Gaussian model, conditioned on the known pixels values. For our application, a Gaussian texture model is estimated using an exemplar background mask. This exemplar mask is obtained from the background pixels class, labelled as described in the previous section. The Gaussian model is then conditionally sampled to gradually replace the labelled bleed-through pixels, using the available neighboring pixels on the border. The conditional sample is obtained by combining an innovation component derived from an independent realization of the Gaussian model and a kriging component derived from the conditioning values [37]. The latter extends the long-range correlations and the for-

mer adds texture details, in a way that preserves the global covariance of the model. A detailed description of the Gaussian conditional inpainting can be found in [37]. Though designed to model microtextures, this algorithm is able to fill both small and large holes, whatever the regularity of the boundary. On the other hand, typical document backgrounds, representing the motif of the paper fiber, seem to fit well a homogeneous microtexture model.

4 Experimental Results

In this section, we present a set of experimental results to evaluate the performance of the proposed method for bleed-through removal in ancient manuscripts. We compare our results with two state-of-the-art methods, specifically designed for correcting the bleed-through distortion, proposed in [17] and [4], respectively. In [17], a non-blind bleed-though removal method is suggested using segmentation of the joint recto-verso intensity histogram. A conditional random field (CRF) based blind bleed-through removal method is suggested in [4]. For evaluation, we use images from the well-known database of ancient documents presented in [38] and available online [1]. This database contains 25 pairs of recto-verso images of ancient manuscripts affected by different levels of bleed-through, along with manually created ground truth binary images of the foreground text. It is worth mentioning that, while this database mainly focuses on bleed-through effects, our method can be used to remove also other document degradations, such as stains, folding marks, etc. For the practical implementation of the proposed method, the transformation from RGB to CIELab and CIELuv color spaces is performed using MATLAB built-in functions. The clustering algorithm is initiated with four classes, $K = 4$, to account for any possible illumination variation in the background of some images. The GMM is initialized by using the K-means function available in MATLAB. We use five iterations for the EM algorithm to find the clusters, and the Gaussian texture inpainting is initiated with pixels from the background class. The presented visual results for the method in [17] are obtained from the online database [38], whereas for the method in [4] the restored images are provided by the authors.

Evaluation of bleed-through cancellation methods is a non trivial task. Generally, the efficacy is evaluated qualitatively, as in most cases the original clean image is not available. Here, we presents both an objective evaluation via quantitative measures and a subjective evaluation using visual comparisons.

[1] http://www.isos.dias.ie

Table 1: Objective evaluation

Method	Precision	Recall	F-measure
Rowley-Brooke [17]	0.92	0.87	0.89
CRF [4]	0.88	0.84	0.86
Proposed	0.92	0.88	0.90

4.1 Numerical Comparison

In the objective comparisons, we employed the familiar metrics of precision, recall and F-measure, calculated for the foreground text pixels. The detected foreground text is compared against the manually created binary ground truth image provided in [38]. All the resulting images are converted to a similar format, i.e., binary, in order to compare them objectively with the related ground truth image. The MATLAB function *imbinarize* is used with adaptive threshold to get the binary images. The comparison metrics are computed as below [4]

$$Precision = \frac{Sum(FT_R \cap FT_{GT})}{Sum(FT_R)}$$
$$Recall = \frac{Sum(FT_R \cap FT_{GT})}{Sum(FT_{GT})}$$
$$F-measure = \frac{2\times(Precision)(Recall)}{Precision+Recall}$$

where FT_R is the foreground text in the restored image and FT_{GT} is the foreground text in the related ground truth image. Table **??** shows the average values for the 25 recto images in the Irish database. It can be observed that the proposed method obtains the best results compared to the blind CRF method, and competitive results compared to the non-blind method in [17]. This numerical comparison shows that the proposed method accurately labels the foreground text, using the color feature vector, without exploiting the extra information provided by the verso side of the document.

4.2 Visual Comparison

The objective comparison, presented in the previous section, neglects the overall quality of the restored document, while we are interested in restoring the original look of the document, including the background texture. Most of the ancient document restoration methods proposed in the literature focus on the removal of the degradation without paying much attention to the estimation of a befitting replacement for the removed pixels. This sometimes results in visual imprints and disturbs the background texture of the document.

Figure 2 shows a visual comparison of images restored by different methods considered. Note that, although our results are in color, we show them in grayscale for a fair comparison with the grayscale results of the

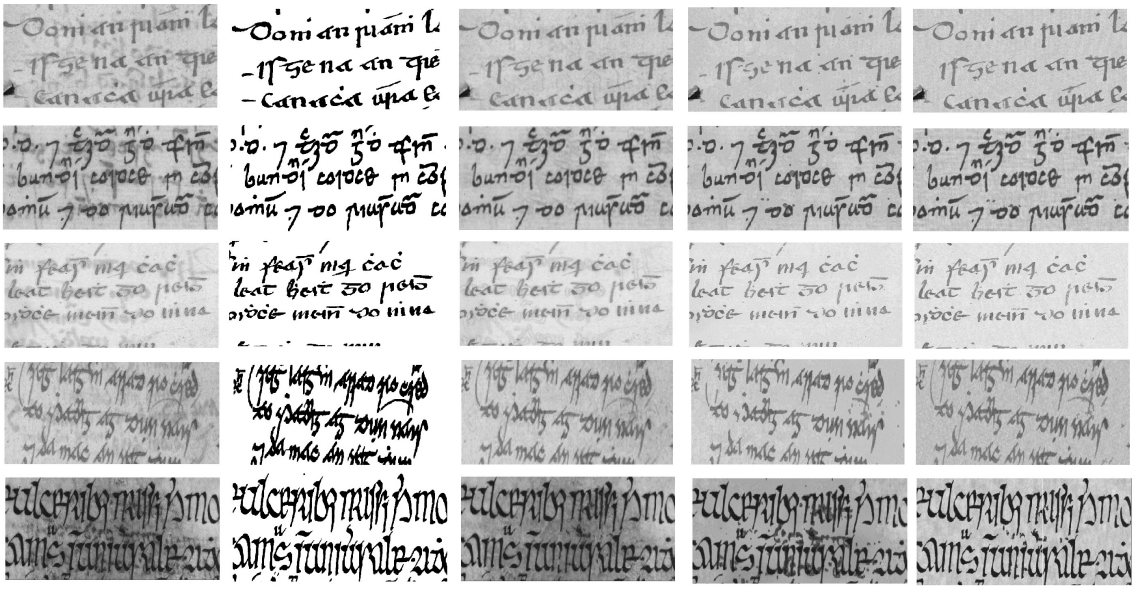

Figure. 2: Visual comparison of bleed-through removal methods: (a) input degraded image, (b) manually created binary ground truth image, (c) restored image by [17], (d) restored image by [4], (e) restored image using the proposed method.

other two methods. The main goal here is to remove bleed-through while leaving intact the foreground text and replacing the bleed-through pixels with suitable values, so that the original look of the document is preserved. The proper inpainting of the bleed-through areas is critical in preserving the paper texture in the background. As can be seen, the proposed method (Fig. 2 (e)) produces comparatively better results in this respect. It efficiently removes the bleed-through degradation, leaving intact the foreground text, and reproducing well the background texture. The nonparametric method of [17] (Fig. 2 (c)) retains the foreground text and the texture of the clean background, but the bleed-through imprints are clearly visible. The method recently proposed in [4] (Fig. 2 (d)) produces better results, but some strokes of the foreground text are missing and some bleed-through strokes are still visible. It is worth to note that, when the background is uniformly illuminated, as in the first three rows of Fig. 2, the EM algorithm was able to reduce the number of clusters from four to three. For the fourth row document, the use of four classes helped in discriminating the dark background in the left part of the document from the foreground and the bleed-through strokes. We then obtained two background classes, a dark one and a light one (right side of the document). When reconstructing the image of Fig. 2 (e) bottom, we merged the two clusters, by inpainting the dark background with the light background. This served to correct the illumination defect present in the document.

A set of color bleed-through free manuscripts, obtained by using the proposed method, is shown in Fig.3. The top row shows the input images and the bottom row displays the correspondent restored images.

Other examples of restored color manuscripts are presented in Fig.4. Note that, in the left-hand side manuscript, in addition to the bleed-through removal, the stain in the top part is also removed. In the right-hand side manuscript of Fig.4, the proposed method successfully retained the blue stamp, along with the black foreground text.

5 Conclusion

This paper presents a color based segmentation method for degradation removal from ancient documents. Segmentation is based on feature vectors created using different color space representations and the pixel spatial location. A GMM based algorithm is used to classify pixels into foreground, background, and noise, i.e. the unwanted, interfering degradation. A texture inpainting method based on Gaussian conditional simulation is used to replace the detected degraded pixels. The method is tested to remove the commonly encountered bleed-through distortion in ancient documents. Our future plan is to extend the method to the correction of other typical degradations affecting ancient documents, such as spots of humidity or mold, and to generalize it to the selective extraction of the different layers of information usually present in the doc-

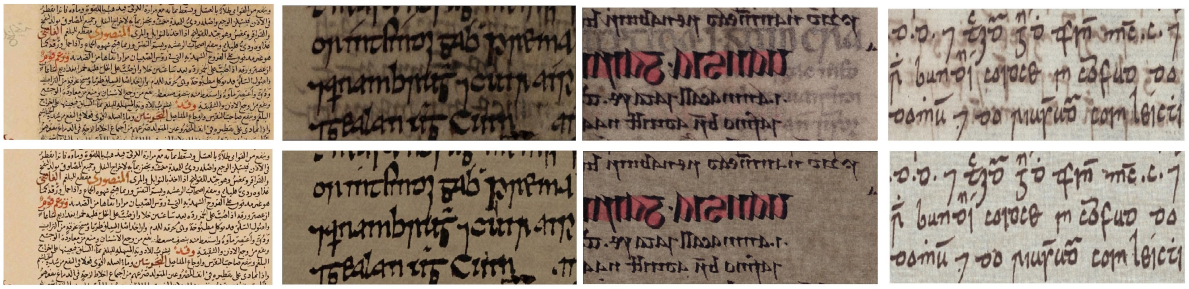

Figure. 3: Visual examples of bleed-through removal: input degraded image (top row), restored image using the proposed method (bottom row).

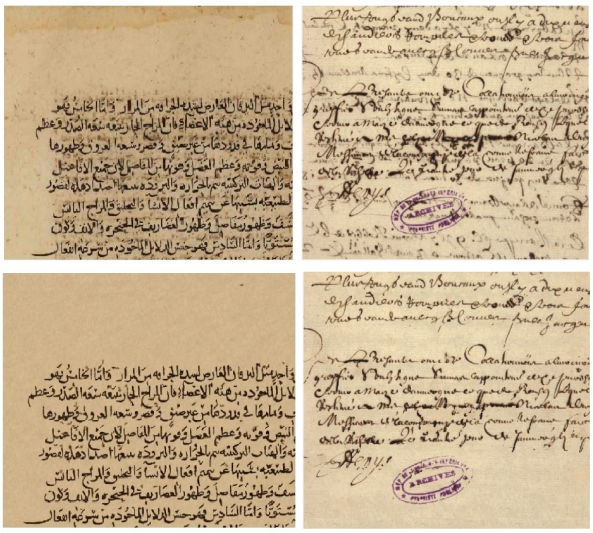

Figure. 4: Visual examples of ancient document image restoration: input image (top row), restored image using the proposed method (bottom row).

uments (stamps, pencil annotations, decorations, etc.). This could be exploited for tasks such as document layout analysis and indexing.

Declarations

Funding and/or Conflicts of interests/Competing interests

No funds or grants were received.
The authors declare no conflict of interest.

Data Availability Statement

The data used in the experimental section of this paper is publicly available on https://www.isos.dias.ie/.

References

1. 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*, vol. 3872, pp. 27–38, 2006.
2. C. Rotaru, T. Graf, and J. Zhang., "Color image segmentation in hsi space for automotive applications." *Journal of Real-Time Image Processing*, vol. 3, 2008.
3. O. Alata and L. Quintard, "Is there a best color space for color image characterization or representation based on multivariate gaussian mixture model?" *Computer Vision and Image Understanding*, vol. 113, pp. 867–877, 2009.
4. 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.
5. C. Wolf, "Document ink bleed-through removal with two hidden markov random fields and a single observation field," *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 431–447, 2010.
6. Z. Shi and V. Govindaraju, "Historical document image enhancement using background light intensity normalization," *Proc. Int. Conf. Pattern Recogn.*, pp. 473–476, 2004.
7. G. Leedham, S. Varma, A. Patankar, and V. Govindaraju., "Separating text and background in degraded document images a comparison of global thresholding techniques for multi-stage thresholding." *IEEE Trans. Neural Netw.*, pp. 244–249, 2002.
8. N. S. Rani, B. J. B. Nair, M. Chandrajith, G. H. Kumar, and J. Fortuny, "Restoration of deteriorated text sections in ancient document images using a tri level semi-adaptive thresholding tech-

nique," *Automatika*, vol. 63, pp. 378–398, 2022.
9. A. Tonazzini, I. Gerace, and F. Martinelli, "Multichannel blind separation and deconvolution of images for document analysis," *IEEE Trans. Image Process.*, vol. 19, pp. 912–925, 2010.
10. X. Zhang, C. He, and J. Guo, "Selective diffusion involving reaction for binarization of bleed-through document images," *Applied Mathematical Modelling*, vol. 81, pp. 844–854, 2020.
11. A. Tonazzini, L. Bedini, and E. Salerno, "Independent component analysis for document restoration," *Int. J. Doc. Anal. Recogn.*, vol. 7, pp. 17–27, 2004.
12. ——, "A markov model for blind image separation by a mean-field em algorithm," *IEEE Trans. Image Process.*, pp. 473–482, 2006.
13. F. Drira, F. L. Bourgeois, and H. Emptoz, "Restoring ink bleed-through degraded document images using a recursive unsupervised classification technique," *Proc. DAS*, pp. 38–49, 2006.
14. R. F. Moghaddam and M. Cheriet, "Low quality document image modeling and enhancement," *Int. J. Doc. Anal. Recogn.*, vol. 11, pp. 183–201, 2009.
15. ——, "A variational approach to degraded document enhancement." *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, pp. 1347–1361, 2010.
16. A. Tonazzini, E. Salerno, and L. Bedini, "Fast correction of bleed-through distortion in grayscale documents by a blind source separation technique," *Int. J. Doc. Anal. Recogn.*, vol. 10, pp. 17–27, 2007.
17. R. Rowley-Brooke, F. Pitié, and A. C. Kokaram, "A non-parametric framework for document bleed-through removal," *Proc. CVPR*, pp. 2954–2960, 2013.
18. H. Yi, M. S. Brown, and X. Dong, "User-assisted ink-bleed reduction," *IEEE Trans. Image Process.*, vol. 19, pp. 2646–2658, 2010.
19. C. Tensmeyer and T. Martinez, "Historical document image binarization: A review," *SN Computer Science*, vol. 1, 05 2020.
20. J. Pastor-Pellicer, S. Espa na Boquera, F. Zamora-Martínez, M. Afzal, and M. Castro-Bleda, "Insights on the use of convolutional neural networks for document image binarization," *International work-conference on artificial neural networks, Springer*, vol. 1, pp. 115 – 126, 2015.
21. J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Computer vision and pattern recognition (CVPR)*, pp. 3431 – 3440, 2015.
22. P. X., W. C., and C. H., "Document binarization via multi-resolutional attention model with DRD loss," *IEEE International conference on document analysis and recognition (ICDAR)*, pp. 45 – 50, 2019.
23. V. GD and P. C., "Document binarization via multi-resolutional attention model with DRD loss," *Pattern Recognition*, vol. 81, pp. 224 – 239, 2018.
24. Z. J, S. C, J. F, W. Y, and X. B, "Document image binarization with cascaded generators of conditional generative adversarial networks." *Pattern Recognition*, vol. 96, 2019.
25. M. Hanif, A. Tonazzini, P. Savino, and E. Salerno, "Non-local sparse image inpainting for document bleed-through removal," *Journal of Imaging*, vol. 4, p. 68, 2018.
26. H. Cheng, X. Jiang, Y. Sun, and J. Xan, "Color image segmentation: advances and prospects," *Pattern Recognition*, vol. 34, pp. 2259–2281, 2001.
27. L. Busin, N. Vandenbroucke, and L. Macaire, "Color spaces and image segmentation," *Advances in Imaging and Electron Physics*, vol. 151, pp. 65–168, 2008.
28. X. Cai, R. Chan, M. Nikolova, and T. Zeng., "A three stage approach for segmenting degraded color images: Smoothing, lifting and thresholding (slat)." *Journal of Scientific Computing*, vol. 72, pp. 1313–1332, 2017.
29. A. Jurio, M. Pagola, M. Galar, C. Lopez-Molina, and D. Paternain., "A comparison study of different color spaces in clustering based image segmentation." *Information Processing and Management of Uncertainty in Knowledge-Based Systems*, vol. 81, pp. 532–541, 2010.
30. N. Vandenbroucke, L. Macaire, and J.-G. Postaire., "Color image segmentation by pixel classification in an adapted hybrid color space. application to soccer image analysis." *Computer Vision and Image Understanding*, vol. 90, pp. 190–216, 2003.
31. S. Park, I. Yun, and S. Lee., "Color image segmentation based on 3d clustering morphological approach." *Pattern Recognition Pattern Recognition*, vol. 31, pp. 1061–1076, 1998.
32. J. M. Chaves-González, M. A. Vega-Rodríguez, J. A. Gómez-Pulido, and J. M. Sánchez-Pérez., "Detecting skin in face recognition systems: a colour spaces study." *Digital Signal Processing*, vol. 20, pp. 806–823, 2010.
33. G. Ruiz-Ruiz, J. Gómez-Gil, and L. M. N. Gracia., "Testing different color spaces based on hue for the environmentally adaptive segmentation algorithm easa." *Computers and Electronics in Agriculture*, vol. 68, pp. 88–96, 2009.
34. M. Orchard and C. Bouman., "Color quantization of images." *IEEE Trans. on Signal Processing.*,

vol. 39, pp. 2677–2698, 1991.
35. K. Blekas, A. Likas, N. Galatsanos, and I. Lagaris., "A spatially constrained mixture model for image segmentation." *IEEE Trans. Neural Netw.*, vol. 16, pp. 494–498, 2005.
36. E. Cappe, O.and Moulines, "On-line expectation-maximization algorithm for latent data models." *Journal of the Royal Statistical Society*, vol. 71, pp. 593–613, 2009.
37. B. Galerne and A. Leclaire, "Texture inpainting using efficient gaussian conditional simulation," *SIAM Journal on Imaging Sciences*, vol. 10, pp. 1446–1474, 2017.
38. R. Rowley-Brooke, F. Pitié, and A. C. Kokaram, "A ground truth bleed-through document image database," in *Theory and Practice of Digital Libraries*, ser. LNCS, P. Z. adn G. Buchanan, E. Rasmussen, and F. Loizides, Eds., vol. 7489. Springer, 2012, pp. 185–196.