Digital document image restoration using a blind source separation method based on copulas

Amal Ourdou¹, Abdelghani Ghazdali¹, Abelmoutalib Metrane¹ and Moad Hakim²

¹ LIPIM, Sultan Moulay Slimane University, Beni-Mellal, Morocco
² LAMAI FST Marrakech, Cadi Ayyad University, Morocco

E-mail: amal.ourdou@gmail.com, a.ghazdali@gmail.com, ab.metrane@gmail.com, moad.hakim1995@gmail.com

Abstract. In the last few decades, digital image degradation issues, such as blur and noise due to the scanning process or the presence of spots, underwriting, overwriting or bleed-through/show-through effects on the image’s background has been a popular research field. To solve this problem, many background removal methods has been introduced in the literature which are based on local or adaptive filters in order to deal with the low-contrast issue. For this paper, we will be focusing on the bleed-through/show-through effects, which is already resolved in literature by an analogy between the front-ground and the background of the image, that is to say, a recognition of two images is required. To fix that problem, we suggest a new restoration method using blind source separation based on copulas theory that models the dependency structure, with the aim of improving text readability and OCR efficiency.

1. Introduction

The presence of influential artifacts in digital documents images degrades its quality significantly, whether it been antique or new scanned texts and manuscripts, which end up making the readability of the text complicated. Artifacts could perhaps emerge from several types of degradation, like scanning optic blur, underwriting, overwriting, noise or spots. The aim of this work is to deal with the so named show-through/bleed-through effects, which is detected either in antique documents due to the process of rewriting the erased parts or the intensity of the ink that appears in the reverse side of the document, or in the present double-side scanned documents due to the transparency of the paper [1, 2]. It is obvious that such intervening strokes should be at least impaired if we want the OCR system to function effectively [3, 4]. Moreover, image pre-processing steps are important to improve the OCR accuracy [5]. To deal with the bleed-through effects problem, several works have been established in the literature which stand on the use of both the front-ground and the background. In [6], a first physical model have been introduced in the literature with the aim of constituting a linear mathematical model, which have been followed by an adaptive filter [7] that employs scans from both sides of the manuscript, while in [8] a thresholding techniques is used to remove the background after a comparison at every pixel of the both sides of a gray-level documents. A wavelet procedure [9] is then used with an iterative process in order to enhance the foreground strokes and the blur in the interfering strokes. Thus, the problem with all those methods is that they necessitate an earlier registration of the both sides the foreground and the background and deal only with the

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

Published under licence by IOP Publishing Ltd
independent components, therefore as it is mentioned in [10, 11, 12], a new BSS method has been introduced to resolve the independent/dependent source elements problem, by using the statistical theory under the name of copula theory, in order to describe the dependency of the source elements, then, minimizing the Kullback-Leibler divergence between the copula density of the approximated source and the copula density of the observations.

In this current article, we used the approach described above with the aim of removing the bleed-through or show-through effects in digital images.

This paper is structured as follows. In section 2 The linear BSS problem is defined in a general manner, after that, in the third section some underlying concepts of copula theory are established, after that in section 4 we expose the proposed approach, and finally, numerical results are presented to reinforce our approach.

2. Blind source separation modelization

Blind source separation (BSS) has become an active research field over the past decades, due to its usefulness in several fields for instance, signal processing, image processing and pattern recognition .... The aim of BSS is to deal with problems where both the source and the mixing matrix are unknown, we possess only the observations which is mixtures of the signals. The linear BSS problem can be presented as follows:

\[ u(t) := Ds(t) + n(t) \in \mathbb{R}^p, \quad t \in [0, T] \]

where \( s(t) := (s_1(t), \ldots, s_p(t))^\top \) are the unknown signals to be estimated from the observations \( u(t) := (u_1(t), \ldots, u_p(t))^\top \), and \( D \in \mathbb{R}^{p \times p} \) is the unknown non-singular matrix that mixes the signals source \( s(t) \) to obtain the observations \( u(t) \). We assume that the number of signals source is equal to the observations. The BSS problem is more obscured in the presence of the noise \( n(t) \), to remedy to this issue one must use some denoising processes that are based on regularization techniques, for instance see [13]. Once the regularization method is adopted the BSS problem becomes a noise-free problem which can be written as follows:

\[ u(t) = Ds(t) \in \mathbb{R}^p \]

As known, the aim of BSS is to determine the signals source \( s(t) \) from the observations \( u(t) \), hence, the predictable value \( w(t) \) is obtained as follows:

\[ w(t) = Zu(t) \]

where \( Z \) denotes a \( \mathbb{R}^{p \times p} \) matrix that de-mixes the observations \( u(t) \) to obtain the signals source \( s(t) \). To do so, we have to retrieve \( \hat{Z} \), which is assumed to be the estimate of the de-mixing matrix and also as close as possible to \( D^{-1} \). Once we retrieve such a de-mixing matrix, one can easily obtain the estimate signals source which is given by the next equation:

\[ \hat{s}(t) := \hat{Z}u(t) \approx s(t) \]

Next we will give a brief introduction on copulas, which are used to model the dependency structure between random variables.

3. Copula theory

In order to deal with the dependence between the marginal distribution and its effects, one must use the notion of copula which is firstly presented by Sklar in [14] as a function which unify the
joint distribution function with its margins. Given any random vector \( W := (W_1, \ldots, W_p) \top \in \mathbb{R}^p, p \geq 2 \), while its joint distribution function is given by:

\[
F_W(\cdot) : w \in \mathbb{R}^p \mapsto F_W(w) := F_W(w_1, \ldots, w_p) := \mathbb{P}(W_1 \leq w_1, \ldots, W_p \leq w_p)
\]

Where its marginal distribution function’s

\[
F_{W_i}(\cdot) : w_i \in \mathbb{R} \mapsto F_{W_i}(w_i) := \mathbb{P}(W_i \leq w_i), \forall i = 1, \ldots, p
\]

By the use of the Sklar theorem [14] we have the existence of a unique function \( C(\cdot) \) inversely, and also by the use of Sklar theorem [14], the function \( C \) is a multivariate distribution function, for any copula function \( C(\cdot) \) is assumed to be continuously differentiable, and it’s introduced by the next equation:

\[
C_W(\cdot) := (F_{W_1}(w_1), \ldots, F_{W_p}(w_p))
\]

The function \( C_W(\cdot) \) is a joint distribution function on \([0, 1]^p\) with uniform marginals known as a copula function. We have, \( \forall y := (y_1, \ldots, y_p) \top \in [0, 1]^p \)

\[
C_W(y) = \mathbb{P}(F_{W_1}(W_1) \leq y_1, \ldots, F_{W_p}(W_p) \leq y_p)
\]

inversely, and also by the use of Sklar theorem [14], the function \( C(F_1(\cdot), \ldots, F_p(\cdot)) \) is a multivariate distribution function, for any copula function \( C(\cdot) \), on \( \mathbb{R}^p \) and any marginal distribution function’s \( F_1(\cdot), \ldots, F_p(\cdot) \) are uniformly distributed on the interval \([0, 1]\), by the reason of the continuity of the marginal distribution function’s \( F_{W_j}(\cdot), j = 1, \ldots, p \). Furthermore, we reach the independence of the components \( W_1, \ldots, W_p \) if and only if (iff) the copula \( C_W(\cdot) \) of the random vector \( W := (W_1, \ldots, W_p) \top \in \mathbb{R}^p \) is presented as follows:

\[
C_W(y) = \prod_{j=1}^{p} y_j =: C_W(y), \forall y \in [0, 1]^p
\]

which is named the copula of independence. To retrieve the copula density, one must derive the copula function which is assumed to be continuously differentiable, and it’s introduced by the next equation:

\[
c_W(y) := \frac{\partial^p C_W(y)}{\partial y_1 \cdots \partial y_p}, \forall y \in [0, 1]^p
\]

Consequently, The copula density of independence, denoted by \( c_0(\cdot) \), is the indicator function of \([0, 1]^p\) given as follows:

\[
c_0(y) := 1_{[0,1]^p}(y), \forall y \in [0, 1]^p
\]

If \( f_W(\cdot) \), exists, which is the joint probability density on \( \mathbb{R}^p \) of the random vector \( Y := (W_1, \ldots, W_p) \top \), and, additively, the associated marginal probability densities \( f_{W_1}(\cdot), \ldots, f_{W_p}(\cdot) \). Then, for all \( w := (w_1, \ldots, w_p) \top \in \mathbb{R}^p \), the relation between \( f_W(\cdot) \), and the copula density is expressed by the next relation:

\[
f_W(w) = \left( \prod_{i=1}^{p} f_{w_i}(w_i) \right) c_W(F_{W_1}(w_1), \ldots, F_{W_p}(w_p))
\]

Several models have been introduced for copulas in the literature, namely semiparametric copula models [15, 16], which present an important category of models that measures the dependency
Table 1: Examples of Semiparametric copulas

| Copula models | Θ | θ₀ |
|---------------|---|----|
| Gaussian      | ] −1, 1[p^{p−1})/2 | 0 |
| AMH           | [−1, 1] | 0 |
| Clayton       | [−1, +∞]\{0\} | 0 |
| Frank         | R\{0\} | 0 |
| Gumbel        | [1, +∞[ | 0 |

structure, and are based on a parametric copulas \{\mathcal{C}(\cdot; \theta); \theta \in \Theta\} that are indexed by a parameter \theta that belongs to a certain subset \Theta. Every single model is recognized by its \theta_0 named as the particular value of \theta that allows us to reach independence given by the next equation:

\[ \mathcal{C}(\mathbf{y}, \theta_0) := \mathcal{C}_0(\mathbf{y}) := \prod_{i=1}^{p} u_i, \forall \mathbf{y} \in [0, 1]^p \]

we recall in the next table some example of semiparametric copulas, and for its expressions see [16, 15]

### 4. Proposed approach

In this section we assume that a document which is affected by the bleed-through/show-through effect is viewed as the superposition of three sources called respectively background, underwriting, and overwriting, consequently, In our BSS problem, we have three different sources that in one way or another fused to provide the studied image. Simultaneously, we may believe, that there were three observed charts that were divided into red, green and blue components. As stated above, we assume having an equality between the number of sources and observations, we may also denote that the color of each of the three sources is indexed as follow (r₁, g₁, b₁) for the background, (r₂, g₂, b₂) for the overwriting, and (r₃, g₃, b₃) for the underwriting. Therefore, the BSS problem is stated as follows:

\[
\begin{bmatrix}
  u_r(t) \\
  u_g(t) \\
  u_b(t)
\end{bmatrix} =
\begin{bmatrix}
  r_1 & r_2 & r_3 \\
  g_1 & g_2 & g_3 \\
  b_1 & b_2 & b_3
\end{bmatrix}
\begin{bmatrix}
  s_1(t) \\
  s_2(t) \\
  s_3(t)
\end{bmatrix}
\]

Where \( u(t) \) denotes the observations, \( D = \begin{bmatrix}
  r_1 & r_2 & r_3 \\
  g_1 & g_2 & g_3 \\
  b_1 & b_2 & b_3
\end{bmatrix} \) the mixing matrix that belongs to \( \mathbb{R}^{3 \times 3} \), and \( s(t) \) is the source that we want to deduce in the end of this process by the next equation:

\[ w(t) = Zu(t) \]

Where \( w(t) \) denotes the estimate value of \( s(t) \). Notice that \( W, U \) and \( S \) are three random variables of \( \mathbb{R}^3 \) whom realizations are respectively \( w, u \) and \( s \). Under the dependency condition of the source components, we denote by \( c_S(\cdot) \) the copula density of the random variable \( S \) which is assumed unknown. To retrieve the de-mixing matrix which is assumed to be as close as
possible to $D^{-1}$, one must introduce the following objective function $Z \mapsto KL(c_W, c_S)$, as stated in [10] where

$$KL(c_W, c_S) := \int_{[0,1]^3} \log \left( \frac{c_W(y)}{c_S(y)} \right) c_W(y)dy$$

$$= \mathbb{E} \left[ \log \frac{c_W(F_{W_1}(W_1), \ldots, F_{W_3}(W_3))}{c_S(F_{W_1}(W_1), \ldots, F_{W_3}(W_3))} \right]$$

Where $KL$ denotes the Kullback-Liebler divergence between the copula density of the source components and the copula density of the observations, which is also the mathematical expectation operator. As stated above the copula density of the source components is assumed unknown, to resolve this problem one must use the semiparametric models for copulas, that is to say, we estimate the unknown copula density by the semiparametric models of copulas which is a set of $L$ candidate semiparametric models, presented as follows:

$$M_l := \{ c_l(\cdot; \theta_l); \theta_l \in \Theta_l \subset \mathbb{R} \}, l = 1, \ldots, L$$

We have then:

$$Z \mapsto \min_{l=1, \ldots, L} \inf_{\theta_l \in \Theta_l} KL(c_W, c_l(\cdot; \theta_l))$$

this kullback-leibler criterion is nonnegative and reach its minimum value zero if and only if $Z = D^{-1}$ (up to scale and permutation indeterminacies). Thus, the de-mixing matrix is approximated as follows:

$$\hat{Z} := \arg \inf_{Z} \inf_{\theta_l \in \Theta_l^*} \bar{KL}(c_W, c_l^*(\cdot; \theta_l^*))$$

where

$$l^* = \arg \min_{l=1, \ldots, L} \inf_{\theta_l \in \Theta_l} \bar{KL}(c_W, c_l(\cdot; \theta_l))$$

$\bar{KL}(c_W, c_l(\cdot; \theta_l))$ is the approximated value of $KL(c_W, c_l(\cdot; \theta_l))$, illustrated as follows:

$$\bar{KL}(c_W, c_l(\cdot; \theta_l)) := \frac{1}{N} \sum_{n=1}^{N} \log \left( \frac{\hat{c}_W(\hat{F}_{W_1}(w_1(n)), \ldots, \hat{F}_{W_3}(w_3(n)))}{c_l(\hat{F}_{W_1}(w_1(n)), \ldots, \hat{F}_{W_3}(w_3(n)); \theta_l)} \right)$$

where, $\hat{c}_W$ is the kernel estimate of $c_W$, that is given $\forall y \in [0, 1]^3$, by:

$$\hat{c}_W(u) := \frac{1}{N H_1 \cdots H_3} \sum_{n=1}^{N} \prod_{j=1}^{3} k \left( \frac{\hat{F}_{W_j}(w_j(m)) - y_j}{H_j} \right)$$

$\hat{F}_{W_j}(x), j = 1, \ldots, 3$, is the approximate of the marginal distribution function $F_{W_j}(x)$ of the random variable $W_j$, at any real value $x \in \mathbb{R}$, given by:

$$\hat{F}_{W_j}(x) := \frac{1}{N} \sum_{m=1}^{N} K \left( \frac{w_j(m) - x}{h_j} \right)$$

where $K(\cdot) : x \in \mathbb{R} \mapsto K(x) := \int_{-\infty}^{x} k(t)dt$, is the primitive of a kernel $k(-)$,. According to [17] one can use the triangular kernel to better estimate the copula density, in order to cope with the boundary effects, presented by:

$$k(x) := (1 - |x|)1_{[-1,1]}(x), \forall x \in \mathbb{R}$$
The choice of the bandwidth parameters $H_1, \ldots, H_3$ and $h_1, \ldots, h_3$ is crucial, therefore, according to Silverman’s rule of thumb, see [18] i.e., for all $j = 1, \ldots, 3$, we take $H_j = \left( \frac{4}{3} \right)^{\frac{1}{4}} \frac{1}{N^{-\frac{1}{4}}} \hat{\Sigma}_j$, and $h_j = \left( \frac{4}{3} \right)^{\frac{1}{4}} \frac{1}{N^{-\frac{1}{5}}} \hat{\sigma}_j$, where $\hat{\sigma}_j$ and $\hat{\Sigma}_j$ denote, respectively, the empirical standard deviation of the data $w_j(1), \ldots, w_j(N)$ and $\hat{F}_{W_j}(w_j(1)), \ldots, \hat{F}_{W_j}(w_j(N))$. Finally, to retrieve the minimum $Z$, we attempt to use the gradient descent algorithm on both $Z$ and $\theta$ of the objective function

$$(Z, \theta) \mapsto \hat{K}L(c_W, c_l(\cdot; \theta_1))$$

Finally the minimizing process leads to the following estimates of the source :

$$\hat{s}(n) := \hat{Z}u(n), n = 1, \ldots, N$$

The following algorithm summarize the proposed approach:

**Algorithm 1** The BSS algorithm for Images

**Data:** the image $I$.

**Result:** the estimated sources $\hat{w}(n), n = 1, \ldots, 3$.

**Initialization:** $u_r = \text{RedComponent}(I)$, $u_g = \text{GreenComponent}(I)$, $u_b = \text{BlueComponent}(I)$

$Z^{(0)} = I_p$, $w^{(0)} = Z^{(0)}u$. Given $\varepsilon > 0$, $\nu > 0$.

**Do:** Update $Z$ and $w$:

$$Z^{(q+1)} = Z^{(q)} - \nu \frac{dKL(c_W, c_S)(Z)}{dZ}$$

$$w^{(q+1)} = Z^{(q+1)}u.$$  

**Until** $\|Z^{(q+1)} - Z^{(q)}\| < \varepsilon$

$$\hat{s} = w^{(q+1)}.$$

### 5. Numerical results

In this section, we will show the efficiency of our approach dealing with the bleed-through/show-through effect [19] by exposing the results using the BSS process based on copulas. To do so, we used a real image of ancient document as shown in figure 1. As we can notice on the four zooming regions, the image is truly affected by the bleed-through effect, this image is the superposition of three observations $u_r$, $u_g$ and $u_b$ given by figure 2. As stated above in section 4 the dependency structure is unknown between the source elements, that’s why we will use our

Figure 1: (a) Image 1, (b) Zoom on four regions
6. Conclusion

The aim of this article is to present a separation of linear BSS problem where the dependency structure is unknown, and applying it to remove the bleed-through/show-through effect, and as we noticed our approach has given tremendous results in the separating process.
[1] Rowley-Brooke R and Kokaram A 2012 Bleed-through removal in degraded documents *Document Recognition and Retrieval XIX* vol 8297 (International Society for Optics and Photonics) p 82970T

[2] De S and Rawal A

[3] Rabex V, Journet N and Domenger J P 2011 Ancient documents bleed-through evaluation and its application for predicting ocr error rates *Document Recognition and Retrieval XVIII* vol 7874 (International Society for Optics and Photonics) p 78740Q

[4] Gupta M R, Jacobson N P and Garcia E K 2007 *Pattern Recognition* 40 389–397

[5] Bieniecki W, Grabowski S and Rozenberg W 2007 Image preprocessing for improving ocr accuracy 2007 *International Conference on Perspective Technologies and Methods in MEMS Design* (IEEE) pp 75–80

[6] Sharma G 2001 *IEEE Transactions on Image Processing* 10 736–754

[7] Franke K and Köppen M 2001 *International Journal on Document Analysis and Recognition* 3 218–231

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

[9] Tan C L, Cao R and Shen P 2002 *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24 1399–1404

[10] Keziou A, Fenniri H, Ghazdali A and Moreau E 2014 *Signal Processing* 104 319–324

[11] Ghazdali A, El Rhabi M, Fenniri H, Hakim A and Keziou A 2017 *Signal Processing* 131 502–513

[12] Keziou A, Mamouni N and Fenniri H 2020 Separation of recto-verso documents using copula based dependent source separation 2020 *IEEE 11th Sensor Array and Multichannel Signal Processing Workshop (SAM)* (IEEE) pp 1–5

[13] El Rhabi M, Fenniri H, Keziou A and Moreau E 2013 *Signal Processing* 93 818–827

[14] Sklar M 1959 *Publ. inst. statist. univ. Paris* 8 229–231

[15] Ali M M, Mikhail N and Haq M S 1978 *Journal of multivariate analysis* 8 405–412

[16] Joe H 1997 *Multivariate models and multivariate dependence concepts* (CRC Press)

[17] Omelka M, Gijbels I, Veraverbeke N *et al.* 2009 *The Annals of Statistics* 37 3023–3058

[18] Silverman B W 1986 *Density estimation for statistics and data analysis* vol 26 (CRC press)

[19] Tonazzini A, Bedini L and Salerno E 2004 *Document Analysis and Recognition* 7 17–27