Document Ink bleed-through removal with two hidden Markov random fields and a single observation field

Christian Wolf

Document Ink bleed-through removal with two hidden Markov random fields and a single observation field

Christian Wolf

***Abstract*— We present a new method for blind document bleed through removal based on separate Markov Random Field (MRF) regularization for the recto and for the verso side, where separate priors are derived from the full graph. The segmentation algorithm is based on Bayesian Maximum a Posteriori (MAP) estimation. The advantages of this separate approach are the adaptation of the prior to the contents creation process (e.g. superimposing two hand written pages), and the improvement of the estimation of the recto pixels through an estimation of the verso pixels covered by recto pixels; Moreover, the formulation as a binary labeling problem with two hidden labels per pixels naturally leads to an efficient optimization method based on the minimum cut/maximum flow in a graph. The proposed method is evaluated on scanned document images from the 18^{th} century, showing an improvement of character recognition results compared to other restoration methods.**

***Index Terms*— Markov Random Fields, Bayesian estimation, Graph cuts, Document Image Restoration.**

I. Introduction

General image restoration methods which do not deal with document image analysis have mostly been designed to cope with sensor noise, quantization noise and optical degradations as blur, defocussing etc. (see [31] for a survey). Document images, however, are often additionally subject to further and stronger degradations:

1) non stationary noise due to illumination changes.
2) curvature of the document.
3) ink and coffee stains and holes in the document.
4) ink bleed through : the appearance of the verso side text or graphics on the scanned image of the recto side. This is an important problem when very old historical documents are processed.
5) low print contrast.
6) errors in the alignment of multiple printing or imaging stages.

Christian Wolf is with Université de Lyon, CNRS; and with INSA-Lyon, LIRIS, UMR5205, F-69621, France. email: christian.wolf@liris.cnrs.fr

In this paper we concentrate on ink bleed through removal, i.e. the separation of a single scanned document image into a recto side and a verso side. We assume that a scan of the verso side is *not* available (blind separation). In this case, the task is equivalent to a segmentation problem: classify each pixel as either *recto*, *verso*, *background*, or eventually *recto-and-verso* (simultaneously), making immediately available the vast collection of widely known segmentation techniques. However, document images are a specific type of images with their own properties and their own specific problems.

At first thought it might be a good idea to interpret the task as a blind source separation problem similar to the "cocktail party" problems successfully dealt with by the (audio) signal processing community. The widely used technique independent components analysis (ICA) has been applied to document bleed through removal, mainly by Tonazzini et al. [34]. However, ICA assumes a linear model which makes this formulation questionable: $\boldsymbol{d}_s = \boldsymbol{A}\boldsymbol{f}_s$ where $\boldsymbol{d}_s$ is the observation vector, $\boldsymbol{f}_s$ is the source vector and $\boldsymbol{A}$ is the mixing matrix. The source vectors, corresponding to the pixels at sites s, are mostly chosen to be three dimensional: the recto signal, the verso signal and an additional signal adding the background color [34]. In this case, the column vectors of the mixing matrix become the color vectors for, respectively, recto pixels, verso pixels and background pixels, as can be seen setting $\boldsymbol{f}_s = [\ 1\ 0\ 0\ ]^T$,$[\ 0\ 1\ 0\ ]^T$ and $[\ 0\ 0\ 1\ ]^T$ and $\boldsymbol{d}_s$ to the respective color vector and solving for $\boldsymbol{A}$. We can easily verify that the linear hypothesis cannot be justified for ink bleed through by calculating the color of an observed pixel created by a source pixel which contains overlapping recto and verso pixels

($\boldsymbol{f}_s = [\ 1\ 1\ 0\ ]^T$), thus the sum of the color vectors for the recto and the verso pixel, which is unlikely.

The same authors present a non-blind technique also applicable to grayscale images [37], the different components corresponding to the recto scan and the verso scan. The inverse of the mixing matrix is calculated using orthogonalization justified by several assumptions on the degradation process. In [35] the same authors introduce a double MRF model similar to our proposition combined with a likelihood term consisting of a linear mixing model. However, whereas our graphical model is directly employed for classification, the MRF in [35] guides an EM algorithm estimating the inverse of the mixing matrix. As with the other algorithms based on mixing, the biggest weakness is the linearity of the model. In [36], the model is extended to convolutive mixtures.

Sharma presents a non-blind restoration algorithm, i.e. a method which requires a scan of the recto as well as the verso side of the document [32]. The two images are aligned using image registration techniques. A reflectance model taking into account a bleed-through spread function is created, approximated and corrected with an adaptive linear filter. Another non-blind method is proposed by Dubois and Pathak [14], where the emphasis is set to the image registration part, the restoration itself is performed using a thresholding-like heuristic.

Tan et al. propose a non-blind method where the alignment is done manually [33]. Foreground strokes are detected using both images and enhanced using a wavelet decomposition and restoration. The same authors also present a directional blind wavelet method exploiting the hypothesis that handwriting is (very) slanted, and therefore that the strokes of the recto and the verso side are oriented differently [41].

Nishida and Suzuki describe a method based on the assumption that high frequency components of the verso side are cut off in the bleeding process [28]. Their restoration process uses a multi-scale analysis and edge magnitude thresholding. Leydier et al. propose a serialized (i.e. adaptive) version of the k-means algorithm [24]. Drira et al. propose an iterative algorithm which recursively applies the k-means algorithm to the image reduced with principal components analysis [13].

The method presented by Don [11] is justified by a noise spot model with Gaussian spatial distributions and Gaussian gray value distributions. Under this assumption, thresholds near the class means produce spurious connected components. A histogram of connected component counts is created and thresholded using standard techniques.

MRF regularization has already been used for this kind of problem. For instance, Tonazzini et al. present a document recognition system which restores selected difficult parts of the document image with MRF regularization [38]. As prior model they chose an Ising model with non-zero clique types of 1, 2, 3, and 9 pixels. The observation model contains a convolution term with an unknown parameter. Donaldson and Myers apply MRF regularization with two priors to the problem of estimating a super-resolution image from several low-resolution observations [12]. A first prior measures smoothness, whereas a second prior measures a bi-modality criterion of the histogram.

In this approach we ignore degradations no. 1 (non stationarity) and 3 (stains) mentioned at the beginning of the paper and propose an approach based on a stationary model. Non homogeneous observation models will be developed in forthcoming publications. The geometry changes in curvature degradations can be treated with preprocessing, e.g. with the method developed in our team [23] or other recent work [7], [44]. The illumination changes inherent in strong curvature degradation can be tackled by non homogeneous observation models.

We formulate our method as Bayesian MAP (maximum a posteriori) estimation problem in terms of two different models:

- the *a priori* knowledge on the segmented document is captured by the prior model. In our case, the prior model consists of two MRFs, one for each side of the document.
- the observation model captures the knowledge on the document degradation process.

In a previous paper, we described a MRF model for document image segmentation [42]. The goal, however, was to learn the spatial properties of text in document images in order to improve the binarization performance. In this paper the emphasis is set to regularization. Therefore, a parametric prior model has been chosen.

The contribution of this paper is threefold:

1) Creation of a double MRF model with a single observation field.
2) Formulation of an iterative optimization algorithm based on the minimum cut/maximum flow in a graph. The proposed inference algorithm is an extension of the widely used α-expansion move algorithm [4][19].
3) A complete restoration process beginning with an algorithm for the initial recognition of recto and verso labels without using any color or gray value information and finishing with a hierarchical algorithm for the calculation of the

background gray value replacing the verso pixel.

This paper is organized as follows. Section II proposes a dependency graph for the joint probability density of the full set of variables (the hidden recto and verso variables as well as the observed variables) and derives the prior probability. Section III proposes the observation model. Section IV describes the posterior probability and formulates an iterative inference algorithm based on graph cuts. Section V outlines the estimation procedure for the prior parameters and the parameters of the observation model. Section VI illustrates the pre- and post processing steps of the method and section VII presents the experiments we performed on scanned document images in order to evaluate the performance of the proposed method. Section VIII finally concludes.

II. The prior model

MRFs capture the spatial distribution of the pixels of an image by assigning a probability (or an energy) to a given configuration, i.e. a given segmentation result. This is normally used to regularize the segmentation process, i.e. to favor certain configurations which are considered more probable. One of the most widely used assumptions is the smoothness criterion - homogeneous areas are considered more probable then frequent label changes.

This assumption is normally justified[1] considering that very often high frequencies in the image content correspond to noise and assuming that the MRF model has been adapted to the prior knowledge on the image content. However, this changes when the observed image is the result of the superposition of two or more "source" images, which is the situation dealt with in this work. In this case, *a priori* knowledge may be available for each of the source images, but not for the mixture of these images. Applying a simple regularization on the combined image may over-smooth areas which should actually contain high frequency edges due to the superposition process.

We therefore propose to create a prior model with two different label fields : one for the recto side (F^1) and one for the verso side (F^2). Instead of a segmentation problem with a configuration space of 3 or eventually 4 labels for each site (*recto*, *verso*, *background*, and eventually *recto-and-verso*), we get a segmentation problem where each pixel corresponds to two different hidden labels, one for each field, and where each label is chosen from a space of two labels: *text* and *background*. The advantages of this formulation are three-fold:

- the separation into two different label fields creates a situation where the priors regularize fields which directly correspond to the natural process "creating" the contents (e.g. hand writing letters), as opposed to the single field case, where the prior tries to regularize a field which is the result of overlapping two content fields.
- Correctly estimating verso pixels which are shadowed by recto pixels, which is only possible with two separate fields, is not just desirable in the case where the verso field is needed. More so, a correct estimation of the covered verso pixels, through the spatial interactions encoded in the MRF, helps to correctly estimate verso pixels which are *not* covered by a recto pixel, increasing the performance of the algorithm.
- As we will see in section IV, the formulation with separate labels leads to a simple optimization routine based on graph cuts.

Note, that a similar result could be achieved with a single hidden label field and by adapting the prior model such that its regularization handles different label interactions differently. In general this produces rather complicated energy functions equivalent to rather simple interactions in the respective fields. Moreover, the formulation of the inference algorithm would have been more complex.

In the following and as usual, uppercase letters denote random variables or fields of random variables and lower case letters denote realizations of values of random variables or of fields of random values. In particular, $P(F{=}f)$ will be abbreviated as $P(f)$ when it is convenient.

MRFs [15][25] are non causal models on undirected graphs which treat images as stochastic processes. A field F of random variables $F_s,\ s{=}1 \dots N$, where N is the number of variables (pixels in our case), is a MRF if and only if

$$\begin{aligned} &P(F{=}f) > 0 \quad \forall f \in \Omega \text{ and} \\ &P(F_s{=}f_s | F_r{=}f_r, r \neq s) \;=\; P(F_s{=}f_s | F_r{=}f_r, r \in N_s) \end{aligned} \tag{1}$$

where f is a configuration of the random field, Ω is the space of all possible configurations and N_s is the neighborhood of the site s. In other words, the conditional probability for a pixel depends only on the pixels of a pre-defined neighborhood around it.

On a graph, each neighborhood defines a set of cliques, where a clique is fully connected sub graph. According to the Hammersley-Cifford theorem [16] [2], the joint probability density functions of MRFs are equivalent to Gibbs distributions defined on the

[1] Label changes on the borders of regions can be ignored, dealt with by a separate line processes[15] or directly in the main process [8].

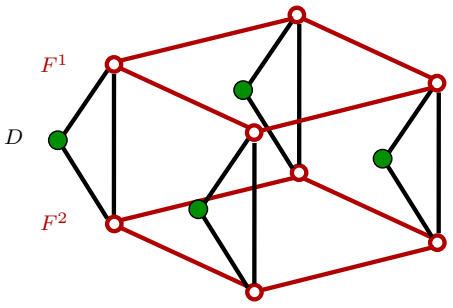


Fig. 1. The dependency graph for a 2×2 pixel image. The model consists of the two label fields F^1 and F^2 ("empty" nodes) and the single observation field D (shaded nodes).

maxima cliques, i.e. are of the form

$$P(f) = \frac{1}{Z}\exp\{-U(f)/T\} \tag{2}$$

where $Z = \sum_f e^{-U(f)/T}$ is a normalization constant, T is a temperature factor which can be assumed to be 1 for simplicity, $U(f) = \sum_{c\in\mathcal{C}} V_c(f)$ is a user defined energy function, $\mathcal{C}$ is the set of all possible cliques of the field and $V_c(f)$ is the energy potential for the realization f defined on the single clique c.

Given the nature of the problem, the three different label fields (two hidden and one observed) should be considered in a holistic way in order to precisely describe the interactions between the two fields and to define a joint probability distribution on the full set of labels. In the rest of this paper we therefore consider a full graph $\mathcal{G} = \{V, E\}$ with a set of nodes V and a set of edges E. V is partitioned into three subsets: the two fields of hidden variables F^1 and F^2 and the field of observed variables D. The three fields are indexed by the same indices corresponding to the pixels of the image, i.e. F_s^1, F_s^2 and D_s denote, respectively, the hidden recto label, the hidden verso label and the observation for the same pixel s. The hidden variables F_s^1 and F_s^2 may take values from the set $\Lambda = \{0, 1\}$, where 0 corresponds to background and 1 corresponds to text. The set of edges E defines the neighborhood on the graph, i.e. there is an edge between two nodes r and s if and only if $r \in N_s$ and $s \in N_r$.

The model described in this work is generative, i.e. it tries to explain the process of creating the observed variables from the hidden ones. Considering the relationships between the observed variables and the hidden variables, i.e. the degradation process (see section III), we assume a first-order MRF, which means that the following two conditions hold (a common assumption in the MAP-MRF framework):

1) The observations D_s are independent conditional to the hidden labels F^1 and F^2.
2) $P(D_s|F^1, F^2) = P(D_s|F_s^1, F_s^2)$

As a consequence, the dependency graph (see figure 1) contains the following clique types: first order and second order "intra-field"[2] cliques in the subgraph F^1, first order and second order "intra-field" cliques in the subgraph F^2 (we will assume the 3-node clique potentials to be zero) and finally the "inter-field" cliques between F^1, F^2 and D. For reasons which will become clear in section III, we will set the potentials for the pairwise inter-field cliques to zero, i.e. the second order cliques with one node $\in F^1$ and one node $\in D$ as well as the second order cliques with one node in $\in F^2$ and one node $\in D$. The only contributing inter-field cliques are therefore three-node cliques with one node of each respective field (F^1, F^2 and D).

The joint probability distribution of the whole graph can therefore be given as follows:

$$\begin{aligned} &P(f^1, f^2, d) \\ &= \frac{1}{Z}\exp\left\{-\Big(U(f^1) + U(f^2) + U(f^1, f^2, d)\Big)/T\right\} \end{aligned} \tag{3}$$

Splitting the partition function $\frac{1}{Z}$ into two factors $\frac{1}{Z_1}$ and $\frac{1}{Z_2}$ and uniting the cliques involving hidden labels only in a single function $U(f^1, f^2)$, which is a change of notation only, we get:

$$\begin{aligned} P(f^1, f^2, d) \;=\; &\frac{1}{Z_1}\exp\left\{-U(f^1, f^2)/T\right\}\cdot \\ &\frac{1}{Z_2}\exp\left\{-U(f^1, f^2, d)/T\right\} \end{aligned} \tag{4}$$

Using the Hammersely-Clifford theorem (1) and Bayes rule, we can interpret equation (4) as a Bayesian problem, which leads us to:

$$P(f^1, f^2, d) \;=\; P(f^1, f^2)P(d|f^1, f^2) \tag{5}$$

The first factor on the right hand side corresponds to the prior knowledge and the second factor corresponds to the data likelihood determined by the observation/degradation model. Inferring the set of hidden labels from the observed labels corresponds to a maximization of the posterior probability (see section IV).

Let us now direct our attention to the prior probability $P(f^1, f^2)$. In the derivation above we saw that the prior was composed of cliques involving hidden variables only, and that there are no cliques containing variables from both fields F^1 and F^2, which can also directly be seen in the dependency graph: the

[2]The reader may have noticed that we frequently denote the subsets of sites F^1, F^2 and D as "fields" and will excuse the slight ambiguity with the "full" Markov random field which consists of all three fields.

two hidden label fields F^1 and F^2 do not share any common nodes nor edges. Therefore,

$$\begin{aligned} P(f^1, f^2) &= \frac{1}{Z} \exp\left\{-U(f^1, f^2)/T\right\} \\ &= \frac{1}{Z} \exp\left\{-\left(U(f^1) + U(f^2)\right)/T\right\} \\ &= \frac{1}{Z_{1'}} \exp\left\{-U(f^1)/T\right\} \cdot \\ &\quad \frac{1}{Z_{2'}} \exp\left\{-U(f^2)/T\right\} \\ &= P(f^1)P(f^2) \end{aligned} \tag{6}$$

We can see that the prior probability is actually the product of the two probabilities of the two fields F^1 and F^2. In other words, the writing on the recto is independent of the writing on the verso page, which makes sense since the two different pages do not necessarily influence each other — they may even have been created by different authors. However, this independence only concerns the situation where no observation has been made. In the presence of observations (the scanned image), the two hidden fields are not independent anymore due to the cliques involving pairs of hidden variables and one observed variable. Intuitively speaking this can be illustrated by the following example: if the observation of a given pixel suggests that at least one of the document sides contains text on this spot (e.g. the gray value is rather low for a white document with dark text), then the knowledge that the recto label is background will increase the probability that the verso pixel will be text.

For a single hidden field, we adopted the widely used Potts model:

$$U(f) = \sum_{\{s\} \in \mathcal{C}_1} \alpha f_s + \sum_{\{s,s'\} \in \mathcal{C}_2} \beta_{s,s'} \delta_{f_s, f_{s'}} \tag{7}$$

where $\mathcal{C}_1$ is the set of single site cliques, $\mathcal{C}_2$ is the set of pair site cliques and δ is the Kronecker delta defined as $\delta_{i,j}$=1 if i=j and 0 else. We chose a stationary and anisotropic model, therefore the single site parameter α depends on the label f_s of the corresponding site s whereas the pair site parameters $\beta_{s,s'}$ depend on the direction of the clique (horizontal or vertical).

Combining (6) and (7), we can see that the whole prior energy defined on both hidden fields is given as the sum of two Potts models:

$$\begin{aligned} & U(f^1, f^2) \\ &= \sum_{\{s\} \in \mathcal{C}_1} \alpha^1 f_s^1 + \sum_{\{s,s'\} \in \mathcal{C}_2} \beta^1_{s,s'} \delta_{f^1_s, f^1_{s'}} \\ &+ \sum_{\{s\} \in \mathcal{C}_1} \alpha^2 f_s^2 + \sum_{\{s,s'\} \in \mathcal{C}_2} \beta^2_{s,s'} \delta_{f^2_s, f^2_{s'}} \end{aligned} \tag{8}$$

Note that only the intra-field cliques from the sets $\mathcal{C}_1$ and $\mathcal{C}_2$ are used in the prior model, the clique potentials from the set $\mathcal{C}_3$ are part of the observation model and will be defined in the next section.

This choice results in a prior parameter vector θ_p which consists of 6 parameters (3 for the recto field and 3 for the verso field):

$$\theta_p = [\alpha^1, \beta^1_h, \beta^1_v \, \alpha^2, \beta^2_h, \beta^2_v \,]^T \tag{9}$$

where superscripts denote the chosen field (1 corresponds to the recto field whereas 2 corresponds to the verso field) and subscripts indicate the direction of the pairwise clique (h denotes horizontal and v denotes vertical).

III. The observation model

We propose a degradation model which allows for several variations and degradations:

- There is no assumption whatsoever on the specific gray level or color of the recto and verso text. However, we assume constant or Gaussian variability of the colors of recto text, verso text and background.
- Eventual linear attenuation of the verso colors by the bleed-through process.
- 100% opaque ink, i.e. that in the observation field a recto text pixel totally covers the corresponding verso pixel, whereas a recto background pixel does not.
- Additional Gaussian noise from the scanning device.

The Gaussian assumption may seem to be an oversimplification of the complex process involved in the degradation of historic documents which very often have been stored for centuries in not optimal conditions. However, the choice is motivated by several reasons : the simplicity of the Gaussian function makes the mathematical formulation of the model easy and very often the oversimplifications of the observation model are compensated by the regularizing effect of the prior.

Degradation models designed for document images do exist and are widely used in the document image community. Unfortunately most of them have been developed for the evaluation of document analysis algorithms and therefore have been designed as binary operations, e.g. a series of morphological operations [1] [45]. In [17], Kanungo et al. propose a degradation model which takes into a account the page bending process as well as the perspective distortion and the illumination change which results from it. These formulations are hard, up to impossible to use in a probabilistic estimation framework.

The assumption of 100% opaque ink could theoretically lead to problems in a situation where the verso

strokes are darker than the recto strokes, requiring that a lighter recto stroke splits a darker verso strokes into two parts. However, this situation is somewhat unlikely due to the fact that the verso strokes are normally brightened by the ink bleeding process. In our large archive of manuscripts and early printed documents, we did not a find a single occurrence of a situation violating this assumption.

The given assumption has not been chosen to simplify the problem. Partial transparent ink would allow us to use a mixture model and therefore infer information on the value of a verso pixel even if the recto pixel is "text". In this case, the mixture model can sometimes be inverted using ICA or related methods, or it can be included into the likelihood term of our proposed model. This does not question neither the framework itself nor the inference algorithm.

The assumptions described above can be expressed as follows:

$$D = \phi(F^1, F^2, \mu^r, \Sigma^r, \mu^v, \Sigma^v, \mu^{bg}, \Sigma^{bg}) + \mathcal{N}(0, \Sigma^N) \tag{10}$$

where $\mathcal{N}$ is the normal law, μ^r, μ^v, μ^{bg} are, respectively, the mean of the recto, verso and background and the covariances Σ^* are defined similarly. Σ^N denotes the covariance of the noise process and ϕ is given as follows:

$$\phi(\ldots) = \begin{cases} \mathcal{N}(\mu^r, \Sigma^r) & \text{if } F^1 = 1 \\ \boldsymbol{A}(\mathcal{N}(\mu^v, \Sigma^v)) - b & \text{if } F^1 = 0 \wedge F^2 = 1 \\ \mathcal{N}(\mu^{bg}, \Sigma^{bg}) & \text{else} \end{cases} \tag{11}$$

The matrix $\boldsymbol{A}$ and the vector b correspond to a possible dampening of the colors by the bleed-through process. For completely transparent paper, A corresponds to the identity matrix and b to the null vector.

As can be seen from the form of the combining function ϕ, the degraded process D consists of three different classes, each following a normal law. The zero mean Gaussian noise of the scanning device as well as the dampening of the verso color are not explicitly modeled, as their effect can be integrated into the parameters of the normal laws of the degraded image. Note that the conditional independence assumptions stated on page 4 are justified by the given model.

As a consequence, the likelihood factorizes as follows:

$$P(d|f^1, f^2) \quad = \quad \prod_s \mathcal{N}(d_s;\ \boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s) \tag{12}$$

where $\boldsymbol{\mu}_s$ is the mean for class f_s (*in the degraded image*) and $\boldsymbol{\Sigma}_s$ is the covariance matrix for class f_s given as follows (note that generally $\mu_r \neq \mu^r$ etc.):

$$\begin{aligned} \boldsymbol{\mu}_s &= \begin{cases} \boldsymbol{\mu}_r & \text{if } f_s^1 = \textit{text} \\ \boldsymbol{\mu}_v & \text{if } f_s^1 = \textit{background} \text{ and } f_s^2 = \textit{text} \\ \boldsymbol{\mu}_{bg} & \text{else} \end{cases} \\ \boldsymbol{\Sigma}_s &= \begin{cases} \boldsymbol{\Sigma}_r & \text{if } f_s^1 = \textit{text} \\ \boldsymbol{\Sigma}_v & \text{if } f_s^1 = \textit{background} \text{ and } f_s^2 = \textit{text} \\ \boldsymbol{\Sigma}_{bg} & \text{else} \end{cases} \end{aligned} \tag{13}$$

where $\boldsymbol{\mu}_r, \boldsymbol{\mu}_v$ and $\boldsymbol{\mu}_{bg}$ are, respectively, and *in the degraded image*, the means for the recto class, the verso class and the background class, and the covariances are denoted equivalently.

IV. The posterior probability and its maximization with graph cuts

Applying Bayes rule to equation (5) and combining the result with equation (6), we get the posterior probability of the two label fields:

$$\begin{aligned} P(f^1, f^2|d) &= \frac{1}{P(d)} P(f^1, f^2) P(d|f^1, f^2) \\ &\propto P(f^1) P(f^2) P(d|f^1, f^2) \end{aligned} \tag{14}$$

As usual, we can ignore the factor $\frac{1}{P(d)}$ not depending on the hidden variables and maximize the joint probability, or minimize its energy. Combining (7), (12) and (14) we get the following energy potential function:

$$\begin{aligned} & U(f^1, f^2, d) \\ &= \sum_{\{s\} \in \mathcal{C}_1} \alpha^1 f_s^1 + \sum_{\{s,s'\} \in \mathcal{C}_2} \beta^1_{s,s'} \delta_{f_s^1, f_{s'}^1} \\ &+ \sum_{\{s\} \in \mathcal{C}_1} \alpha^2 f_s^2 + \sum_{\{s,s'\} \in \mathcal{C}_2} \beta^2_{s,s'} \delta f_s^2, f_{s'}^2 \\ &+ \sum_{\{s\} \in \mathcal{C}_1} \frac{1}{2} (d_s - \boldsymbol{\mu}_s)^T \boldsymbol{\Sigma}_s^{-1} (d_s - \boldsymbol{\mu}_s) \end{aligned} \tag{15}$$

where and μ_s and Σ_s are the sufficient statistics for the observation model given by the labels f_s and $f_{s'}$. To estimate the binary images, equation (14) must be maximized. Unfortunately, the function is not convex and standard gradient descent methods will most likely return a non global solution. Simulated Annealing has been proven to return the global optimum under certain conditions [15], but is painfully slow in practice. Loopy belief propagation is another option, giving an approximative solution by iteratively applying Pearl's belief propagation algorithm originally designed for belief networks [29]. In this work we will take advantage of the nature of the dependency graph (binary labels and cliques with not more than 2 hidden labels) in order to derive an optimization algorithm based on the calculation of the minimum cut/maximum flow in a graph [4][5][10][19].

For convenience we will rewrite the energy function for the whole graph in terms of unary functions U_1 and two types of binary functions U_2 and U_2' as follows:

$$\begin{aligned} & U(f^1, f^2, d) \\ = & \sum_{\{s\} \in \mathcal{C}_1} \left[\alpha^1 U_1(f_s^1) + \alpha^2 U_1(f_s^2) \right] \\ + & \sum_{\{s,s'\} \in \mathcal{C}_2} \left[\beta_{s,s'}^1 U_2(f_s^1, f_{s'}^1) + \beta_{s,s'}^2 U_2(f_s^2, f_{s'}^2) \right] \\ + & \sum_{\{s\} \in \mathcal{C}_1} U_2'(f_s^1, f_s^2; d_s) \end{aligned} \tag{16}$$

where $U_1(f_s) = f_s$, $U_2(f_s, f_{s'}) = \delta_{f_s, f_{s'}}$ and $U_2'(f_s^1, f_s^2; d_s) = \frac{1}{2}(d_s - \boldsymbol{\mu}_s)^T \boldsymbol{\Sigma}_s^{-1}(d_s - \boldsymbol{\mu}_s)$. We consider $U_2'(.,.;.)$ as a binary function since we do not maximize over the third argument, which is an observed variable.

Although the problem involves two possible labels for each hidden variable ($|\Lambda| = 2$), the exact solution for equation (16) cannot be found using algorithms based on graph cuts. As shown by Kolmogorov et al. [19], a function of binary variables composed of unary terms and binary terms is *graph-representable*, i.e. it can be minimized with algorithms based on the calculation of the maximum flow in a graph, if and only if each binary term $E(.,.)$ is *regular*, i.e. it satisfies the following equation:

$$E(0,0) + E(1,1) \leq E(0,1) + E(1,0) \tag{17}$$

It can easily be seen that this is the case of the terms $U_2(.,.)$ in equation (16), but not necessarily for all terms $U_2'(.,.;.)$. According to the value of the observation d_s at site s, $U_2'(f_s^1, f_s^2; d_s)$ may be regular or not. In other words, only if the observation likelihood for equal labels f_s^1 and f_s^2 is higher than the observation likelihood for different labels, then the term is regular for site s.

We therefore propose an adaptation and extension of the iterative α-expansion move algorithm proposed by Boykov et al. [5] for labeling problems with multiple labels ($|\Lambda| > 2$) and improved by Kolmogorov et al. [19]. In the original iterative formulation for multi label problems, each subproblem is a binary problem where each hidden variable may take two *virtual* labels: x_s and α, where x_s is the original (current) label, and α is a new label, whose value is changed at each iteration.

In our case, the iteratively solved binary labeled and regular subproblems arise fixing the hidden labels of one of the two fields F^1 and F^2 and estimating the labels of the other one. Completely fixing a whole set of variables corresponds to running an α-expansion move algorithm on a single field dependency graph where each single hidden variable f_s may take 4 values (*background, recto, verso, recto-verso*) and the pairwise clique potentials are adapted accordingly.

Fig. 2: The inference algorithm iteratively optimizing two different binary subproblems. H is a matrix storing for each pixel whether its likelihood term is regular or not. $U^{\mapsto 2}(f^1, f^2, d, H)$ and $U^{\mapsto 1}(f^1, f^2, d, H)$ correspond to the posterior energy with different hidden variables clamped.

Input: d (a realization of the observed field)
Output: f^1, f^2 (estimated label fields)

$F^1, F^2 \leftarrow$ Initialize the label fields (e.g. with k-means)
$H \leftarrow$ Determine the regular sites s
repeat
- Fix f_s^1 for $H_s = 0$, optimally estimate f_s^2 for all s and f^1 for $H_s = 1$ maximizing $U^{\mapsto 2}(f^1, f^2, d, H)$
- Fix f_s^2 for $H_s = 0$, optimally estimate f_s^1 for all s and f^2 for $H_s = 1$ maximizing $U^{\mapsto 1}(f^1, f^2, d, H)$

until *Convergence*

This optimization schedule may be improved by fixing only the variables whose sites s are not regular, and jointly estimating the variables f_s^1 and f_s^2 for the regular sites s. For convenience we introduce a binary matrix H indicating for each site s whether it is regular or not, i.e. whether the associated function $U_2'(f_s^1, f_s^2; d_s)$ is regular or not:

$$H_s = \begin{cases} 1 & \text{if} \quad U_2'(0,0,d_s) + U_2'(1,1,d_s) \leq \\ & \quad\ U_2'(0,1,d_s) + U_2'(1,0,d_s) \\ 0 & \text{else} \end{cases} \tag{18}$$

Figure 2 outlines the inference algorithm, which iteratively calculates the exact solution of two different binary subproblems, maximizing, respectively, $U^{\mapsto 2}(f^1, f^2, d, H)$ and $U^{\mapsto 1}(f^1, f^2, d, H)$. These two energy functions are actually equivalent, however, the set of fixed variables and the set of estimated variables being different, they lead to two different cut graphs. In order to show the derivation of the cut graphs, we will rewrite the two functions by reordering some terms. Without loss of generality, in the rest of this section we describe $U^{\mapsto 2}(f^1, f^2, d, H)$, i.e. the subproblem where a subset of the variables in F^1 is fixed, whereas the variables of F^2 and the complementary subset of variables in F^1 are estimated. The function $U^{\mapsto 1}(f^1, f^2, d, H)$ corresponding to the complementary subproblem can be derived in similar way

After separating terms according to the contents of H, the corresponding energy function can be given as follows:

$$
\begin{array}{l|ll}
 & U^{\mapsto 2}(f^1, f^2, d, H) & \\
1 & = & \sum_{\{s\}\in\mathcal{C}_1:H_s=0} \alpha^1 U_1(f_s^1) \\
2 & + & \sum_{\{s\}\in\mathcal{C}_1:H_s=1} \alpha^1 U_1(f_s^1) \\
3 & + & \sum_{\{s\}\in\mathcal{C}_1} \alpha^2 U_1(f_s^2) \\
4 & + & \sum_{\{s,s'\}\in\mathcal{C}_2:H_s=0\wedge H_{s'}=0} \beta_{s,s'}^1 U_2(f_s^1, f_{s'}^1) \\
5 & + & \sum_{\{s,s'\}\in\mathcal{C}_2:H_s=1\wedge H_{s'}=1} \beta_{s,s'}^1 U_2(f_s^1, f_{s'}^1) \\
6 & + & \sum_{\{s,s'\}\in\mathcal{C}_2:H_s\neq H_{s'}} \beta_{s,s'}^1 U_2(f_s^1, f_{s'}^1) \\
7 & + & \sum_{\{s,s'\}\in\mathcal{C}_2} \beta_{s,s'}^2 U_2(f_s^2, f_{s'}^2) \\
8 & + & \sum_{\{s\}\in\mathcal{C}_1:H_s=0} U_2'(f_s^1, f_s^2; d_s) \\
9 & + & \sum_{\{s\}\in\mathcal{C}_1:H_s=1} U_2'(f_s^1, f_s^2; d_s)
\end{array}
\tag{19}
$$

Written in this notation, The energy functions can be directly translated into a cut graph using the method introduced by Kolmogorov et al. [19]. The cut graph then contains, besides the terminal nodes $source$ and $sink$, one node for each variable F_s^2 as well as one node for each variable F_s^1 satisfying $H_s = 1$. Each unary term is translated into a t-edge, and each binary term is translated into an n-edge as well as two t-edges.

The terms in lines 1 and 4 of equation (19) do not depend on estimated variables and therefore can be omitted during the minimization. The terms in lines 2 and 3 contain standard unary functions and will be represented by t-edges. The terms in lines 5 and 7 contain standard binary functions (pairwise cliques of the Potts model) and will be represented by n-edges. The terms in line 6 are binary functions (also pairwise cliques of the Potts model) in the full original expression (equation (16)), but one of the two arguments is fixed in equation (19) describing the sub problem. They can therefore be represented as t-edges in the cut graph. Similarly, the terms in line 8 are non-regular pairwise functions of the observation model, which can be represented as t-edges. The terms in line 9, finally, correspond to the regular pairwise function of the observation model, which can be represented as n-edges.

Table I gives a full description of the different edges of the cut graph and their weights. Figure 3 shows an example of a dependency graph for a toy problem, a 3×1 image, and two different cut graphs. Figure 3b shows the cut graph for the α-expansion move like algorithm, i.e. all sites s are considered as non-regular. The cut graph is shown for the case where the complete set of variables F^1 is fixed whereas the complete set of variables F^2 is estimated.

Figure 3c shows the extended algorithm, where the middle and the right site are considered regular, whereas the left site is considered non-regular. For the middle and the right site, the variables F_s^1 and F_s^2 are jointly estimated, whereas for the left site only F_s^2 is estimated whereas F_s^1 is fixed.

V. Parameter estimation

Since realizations of the label fields F^1 and F^2 are not available, the parameters of the prior model and the observation model must be estimated from the observed data or from intermediate estimations of the label fields. In this work we chose to estimate the parameters in a unsupervised manner, i.e. different parameters are estimated specifically from and for each input image. To this end, we create initial label fields with the k-means method and median filter them before applying the supervised estimation technique described below. Alternatives to this unsupervised approach would be, for instance, iterated conditional estimation [6] or the mean field theory [43].

The parameters of the observation model are estimated using the classical maximum likelihood estimators, i.e. the empirical means and covariances.

A. The MRF hyper-parameters

For the supervised estimation of the MRF parameters we use least squares estimation, which was first proposed by Derin et al. [9]. For a single MRF the estimation procedure may be described as follows.

The potential function for a single site s may be given as

$$U(f_s, f_{N_s}, \theta_p) = \theta_p^T N(f_s, f_{N_s}) \tag{20}$$

where N_s are the intra-field neighbors of s: $N_s = \{f_{we}, f_{ea}, f_{no}, f_{so}\}$, θ_p is the prior parameter vector and $N(f_s, f_{N_s})$ can be derived from (7) as follows:

$$
\begin{aligned}
N(f_s, f_{N_s}) = [\quad & \delta_{f_s,1}, \\
& \delta_{f_s,f_{we}} + \delta_{f_s,f_{ea}} \\
& \delta_{f_s,f_{no}} + \delta_{f_s,f_{so}} \quad]^T
\end{aligned}
\tag{21}
$$

From (20) and the basic definition of conditional probabilities on MRFs:

$$P(f_s|N_s) = \frac{e^{-U(f_s,f_{N_s},\theta_p)}}{\sum_{f_s\in\mathcal{L}} e^{-U(f_s,f_{N_s},\theta_p)}} \tag{22}$$

n-edges for node pairs:	Weight	Line in eq. (19)
$F_s^2, F_{s'}^2 : (s,s') \in \mathcal{C}_2$	$-\beta_{s,s'}^2$	7
$F_s^1, F_{s'}^1 : (s,s') \in \mathcal{C}_2 \wedge H_s = 1 \wedge H_{s'} = 1$	$-\beta_{s,s'}^1$	5
$F_s^1, F_s^2 : H_s = 1$	$U_2'(0,1,d_s) + U_2'(1,0,d_s) - U_2'(0,0,d_s) - U_2'(1,1,d_s)$	9

(a)

t-edges (to source if weight> 0) for nodes:	Weight	Line in eq. (19)
F_s^2	α^2	3
$F_s^2 : H_s = 0$	$U_2'(f_s^1,1,d_s) - U_2'(f_s^1,0,d_s)$	8
$F_s^1 : H_s = 1$	α^1	2
$F_s^1 : H_s = 1$	$\sum_{s':H_{s'}=0 \wedge f_{s'}^1=1} \beta_{s,s'}^1 - \sum_{s':H_{s'}=0 \wedge f_{s'}^1=0} \beta_{s,s'}^1$	6
$F_s^1 : H_s = 1$	$U_2'(1,0,d_s) - U_2'(0,0,d_s)$	9
$F_s^2 : H_s = 1$	$U_2'(1,1,d_s) - U_2'(1,0,d_s)$	9

(b)

TABLE I

THE EDGES ADDED TO THE CUT GRAPH FOR THE PROPOSED INFERENCE ALGORITHM: EACH EDGE CORRESPONDS TO A TERM IN EQ. (19). EACH T-EDGE IS CONNECTED TO THE SOURCE IF THE WEIGHT IS POSITIVE, OR CONNECTED TO THE SINK IF THE WEIGHT IS NEGATIVE, IN WHICH CASE THE ABSOLUTE VALUE OF THE WEIGHT IS USED. MULTIPLE EDGES BETWEEN SAME NODES (TAKING INTO ACCOUNT THE ORIENTATION) ARE REPLACED BY A SINGLE EDGE, ITS WEIGHT BEING THE SUM OF THE INDIVIDUAL WEIGHTS.

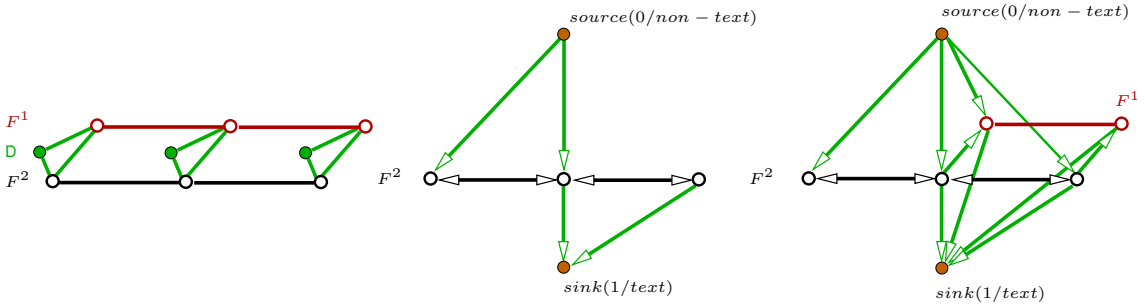


Fig. 3. (a) The dependency graph of a simple model containing three pixels in a single row; (b) the cut graph for an α-expansion move like inference algorithm: inference of the verso pixels; (c) the cut graph for the proposed inference algorithm: joint inference of the verso pixels and of a subset of the recto pixels. In this example, the potential functions related to the observation model are regular for the middle and for the right pixel ($H_s = 1$), but not for the left one ($H_s = 0$).

the following relationship can be derived [9]:

$$\theta_p^T [N(f_s', f_{N_s}) - N(f_s, f_{N_s})] = \ln \left(\frac{P(f_s, f_{N_s})}{P(f_s', f_{N_s})} \right) \quad (23)$$

where f_s' is a label different of f_s. The RHS of (23) can be estimated using histogram techniques [9], counting the number of occurrences of the clique labellings in the label field. Considering all possible combinations of f_s, f_s' and f_{N_s}, (23) represents an over determined system of linear equations which can be rewritten in matrix form as follows:

$$\boldsymbol{N}\boldsymbol{\theta}_p = \boldsymbol{p} \quad (24)$$

where $\boldsymbol{N}$ is a $M \times 6$ matrix, M being the number of data points, i.e. the number of different combinations of label pairs f_s and f_s' having the same neighborhood labels f_{N_s}. The rows of $\boldsymbol{N}$ contain the transposed vectors $[N(f_s', f_{N_s}) - N(f_s, f_{N_s})]^T$. The rows of the vector $\boldsymbol{p}$ contain the corresponding values from the RHS of (23). The system (24) can be solved using standard least squares techniques, as for instance the pseudo inverse.

For practical purposes, note that labeling pairs with one or both of the probabilities $P(f_s, f_{N_s})$ and $P(f'_s, f_{N_s})$ equal to zero cannot be used. Furthermore, Derin et al. suggest to discard equations with low labeling counts in order to make the estimation more robust.

Adapting the estimation procedure for a double MRF is straight forward. We estimate the parameters on the recto field only, since this field is more stable — all its labels are directly related to the observation field. We get the parameters of the verso field from the assumption that, statistically speaking, the verso field is a flipped version of the recto field, which does not change the parameters.

VI. Pre- and postprocessing

A. Initialisation of the label fields

An initial estimation of the two label fields f^1 and f^2 is needed for the iterative algorithm described in the previous section. A natural choice is to apply a segmentation technique without regularization, e.g. a k-means segmentation, in order to classify the pixels into three clusters. However, we need to determine for each cluster whether it is background, recto or verso. For most images that could be done using gray level information only, background being the lightest cluster and recto being the darkest one. In order to make this choice more robust, we developed a cluster labeling method which does not use the gray level of the pixels. Instead, it is based on the following two assumptions:

Assumption 1: Most space on the document page is occupied by background.

Assumption 2: The ink is 100% opaque and therefore a recto text pixel completely covers a verso pixel. The first assumption is used to determine the background cluster as the one having most pixels, which is rather straightforward and very efficient. The second assumption is used to determine which one of the two remaining cluster labels — henceforth denoted label a and label b — is the recto label. The basic idea is the following: since recto pixels cover verso pixels, connected components in the (unobservable) verso label field are often cut into several connected components in the observation field when they interact with connected components from the (unobservable) recto label field.

Since we do not have the unobservable label fields — which would make the task trivial — we use histogram statistics on the initial segmentation to exploit this fact. We search for all the places in the image where the two labels interact, i.e. where there are two neighboring pixels, one having label a and the other having label b. The recto label is obtained counting the different connected components touching each label and choosing the label having a minimum number of different components.

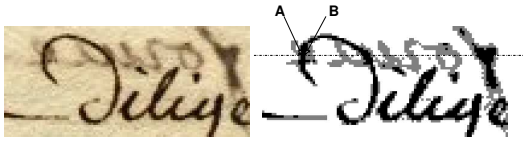


Fig. 4. (a) an input image; (b) the result of the initial k-means clustering. As described by assumption 2, the word in the upper part of the image, which belongs to the verso side of the document, has been cut into several connected components by a letter written on the recto side.

As an example, consider the two transitions indicated by the two points A and B in figure 4. For these two transitions, 2 connected components are counted for the gray-ish label, while only one connected component is counted for the darker label.

B. Restoration

The principle of the restoration algorithm is simple: replace the color or gray value of the pixels classified as verso by the color or gray value of the background. Directly using the mean of the background class will produce visible artifacts due to the noise in the image. A better solution is to use the mean of the neighboring pixels classified as background.

Searching these pixels might be laborious in cases where we need to fill larger areas of verso pixels. We therefore resort to a hierarchical pyramidal structure for the calculation of the replacement values. The pyramid is characterized by a 2×2 reduction function and a receptive field of 3×3 children for each parent site. Each sites holds the mean value of the children's grayvalues as well as the number of children which have been labeled as background. In order to replace a verso pixel, we traverse the pyramid from bottom to the top and stop at a level where enough pixels have been found contributing to a meaningful background value.

VII. Experimental results

Evaluating document restoration algorithms is a non trivial task since ground truth is very hard to come by. Short of manually classifying each pixel in a scanned image, the only way to get reliable ground truth data on pixel level is to test the algorithm on synthetic data. These tests, on the other hand, may not be realistic enough to capture all the subtleties of a real environment. To evaluate our algorithm we therefore decided to test its ability to improve the performance

of an OCR algorithm when applied to real scanned documents.

We chose a dataset consisting of 104 pages of low quality printed text from the 18^{th} century, the *Gazettes de Leyde*. This journal in French language was printed from 1679 to 1798 in the Netherlands in order to escape the censorship in France at the 18^{th} century and relates news of the world. The Gazettes are currently used by several research projects in social and political sciences, some of which are currently collaborating with our team in the framework of digitization projects.

From an image processing point the view, the data situates itself between the difficulty of manuscripts and the regularity of printed documents. The images of sizes around 1030×1550 pixels are of very low quality compared to modern printed text. Recognition is possible, although the performance on the non-restored images is not very high. We chose the open source OCR software "Tesseract" published by Google, mainly because it is easily scriptable[3], but we also performed some selected experiments with the product of the market leader, Abby Finereader 8[4], which performs slightly better without changing the ranking of the restoration methods.

As mentioned in section V, the parameters of the method have been estimated in an unsupervised manner, i.e. for each image we estimate specific parameters.

We compared the proposed method with several competing methods. One group of algorithms purely exploits the fact that, according to the hypothesis stated in section III, set recto pixels completely cover verso pixels, without taking into account interactions between neighboring pixels. Examples are the k-means clustering algorithm with k=3 clusters (followed by our restoration algorithm replacing verso pixels, explained in section VI-B), as well as two thresholding algorithms. We chose two methods which represent the state of the art in adaptive thresholding: Niblack's algorithm [27] which performed best in a widely cited evaluation paper [39] as well as an improvement of Niblack's algorithm by Sauvola et al. [30]. Since a restoration is not straightforward from a binary output, we directly fed the binary images to the OCR in the case of the two thresholding algorithms.

As mentioned in section I, statistical source separation is one of the most active areas in bleed-through removal with several works published by Tonazzini et al. on this subject [34][35][36][37]. We therefore decided to compare the proposed method with two of them: since the scans of the *Gazettes de Leyde* are in color, the color model introduced in [34] and which we described in section I is applicable. The second method, introduced in [37] and based on orthogonalization, is non-blind and therefore requires the presence of the verso side of the image. In a personal communication sent for the experiments in this paper, Prof. Tonazzini recommended the use of two different planes of the color image as recto and verso observations, which we did in our experiments. The source codes have kindly been provided by the author, Prof. Tonazzini herself.

[3]http://code.google.com/p/tesseract-ocr

[4]http://finereader.abbyy.com

Color space	Distance	Recall	Prec.	Cost
RGB	Euclidean	78.91	68.23	42,835
Grayvalue	Euclidean	**79.82**	68.43	42,675
L*u*v*	Euclidean	78.30	68.50	41,800
L*a*b*	Euclidean	78.57	**69.43**	**40,375**
L*a*b*	CIE94 [26]	78.50	68.95	41,142

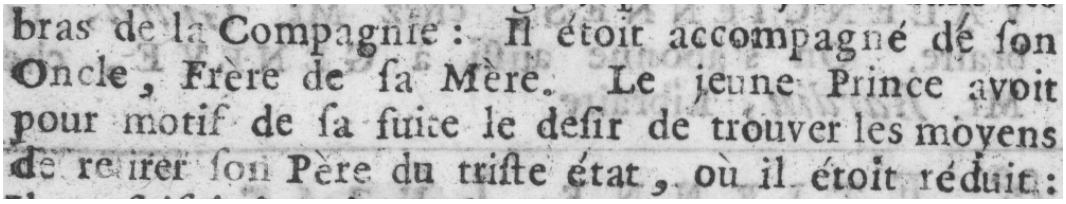
TABLE III
OCR RESULTS FOR THE K-MEANS METHOD USING SEVERAL DIFFERENT COLOR SPACES.

The tested source separation methods are not automatic, they need user interaction in order to chose the correct output source plane. While the number of the correct recto plane may be different between different images, tests showed that for all 104 images of the *Gazettes de Leyd*, the order of the source images was the same. The source planes resembling the most to the assumed recto plane where, for both methods, source #1 and source #2, which we both included into the experiments. This was not the case for other images, as for instance the manuscripts shown in figures 8 and 9.

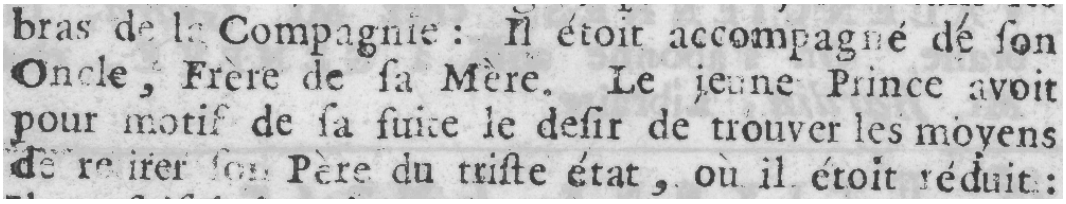
The last method compared to the proposed algorithm is a standard single MRF with a Potts model and three labels (*recto*, *verso* and *background*), optimized using Kolmogorov et al's version of the α-expansion move algorithm [19] and combined with the same parameter estimation and pre- and post-processing as our proposed method.

The k-means method has been tested with different color spaces: grayscale, RGB, L*a*b*, L*u*v*, using the Euclidean distance for each space. Additionally, the CIE94 color metric [26] has been tested for the L*a*b* space (see table III). The best results have been obtained with the L*a*b* and the Euclidean distance.

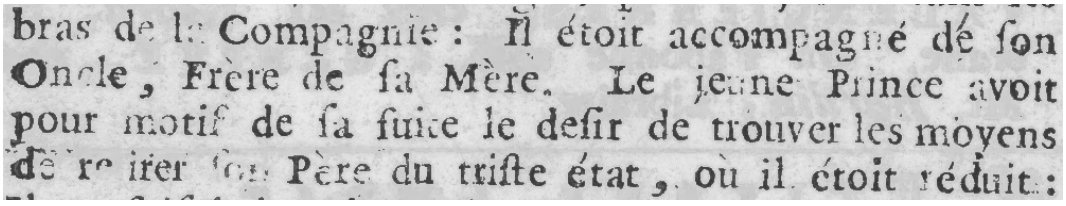
Figures 5 to 7 illustrate the OCR results on a small image taken from the *Gazettes* dataset. As we can see, being based on segmentation, the results for k-means and the two MRF methods are similar. The k-means result (Figure 5b) is noisy as opposed to the MRF results, the double MRF (Figure 5d) improves the regularity of the single MRF (Figure 5c). The OCR



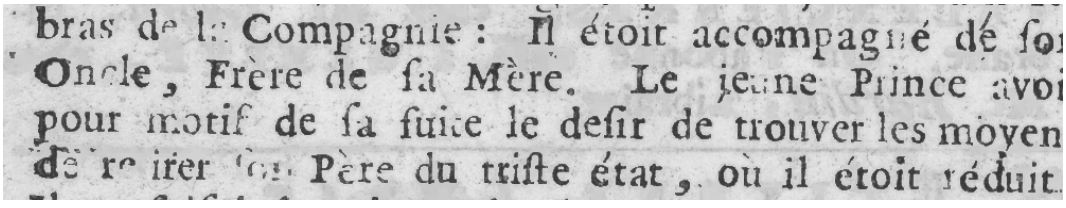
```
bras ‘d5:glgrComprzgnîc‘: |·Y· ;]L\ët4ôi;n àccoâniàagnë Lié ibn _
Uncle } ]F1‘èré‘dc"i‘a""MÈrèQ_ Lei j,cxn,c> Pxjnèci ziypir .
four motif dc fa fqizc lc déiïr de/trouver ies mbicns i
d’é"rë;ifèi‘ Qin Përç dg triiïcfgitar,. 0ùlî1.é;0it Yë(gliE1_’·
```



```
bras de lx =>, Compsignielz   émir accompagné de ibn _
Oncle ,’· Frère dela Mere, Le j,eune· Prince avoit .
pour motif de la faire le delîr de trouver les moyens
 irer fh;] Père du t.tiPce état,. où. il, étoit réduit;
```

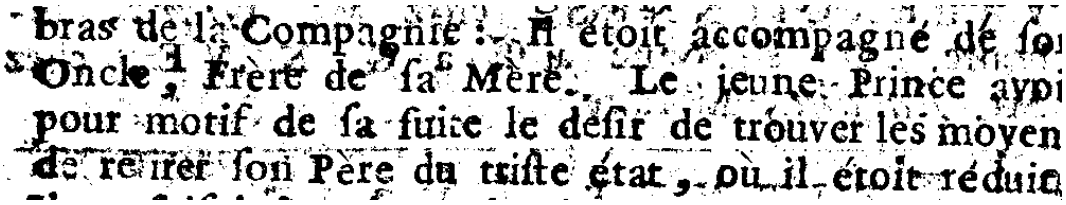


```
bras de l:îC0mP;rg1rrE: Il étoit accompagné dé (on _
Oncle ,’ Frère de fu Mere. Le ierzne Prince avoit .
pour motif de la fuite le delîr de trouver les mbvens
’dë‘r·· iter ⋆}>;> Père du ttiftc état,. où il. étoit réduitr;
```

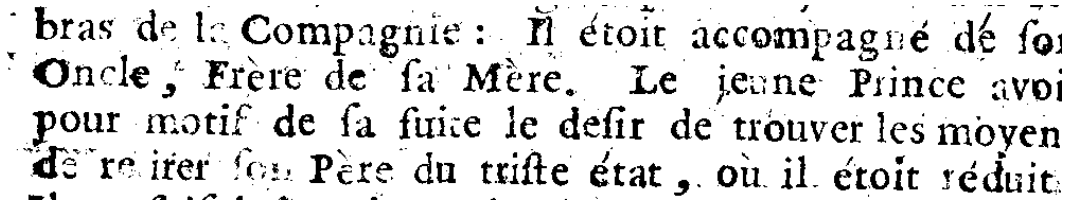


```
’ bras de l:1Comp;1gni·:: Il étoit accompagne dé Ton _
‘ Oncle, Frère de fa Mere. Le j.e;:ne· Prince avoit
pour motif dc fa fuite le defir de trouver les moyens
‘àië’r<<· iter ‘\}>>:· Père du triûe état,. ou il étoit réduit.:
```

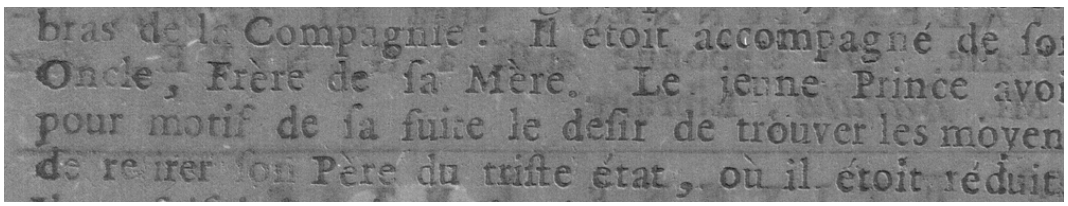
Fig. 5. Small extracts of the OCR results obtained on scanned document images: (a) input image (no restoration); (b) k-means segmentation + restoration; (c) single MRF segmentation [19] + restoration; (d) double MRF (proposed method).



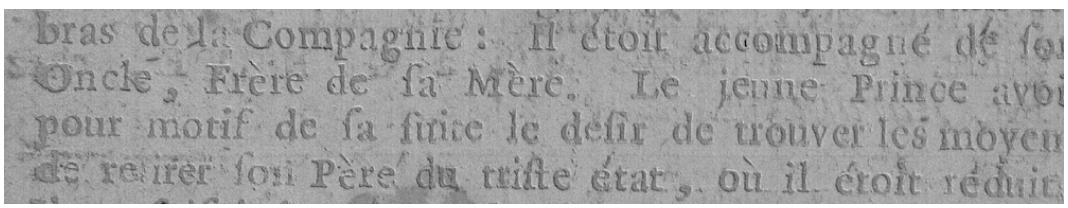
```
swmkg È \  >>· ggëfg   g. ;˜*·">>.<<ë'f >>'î?\P**$%? ,+Y·É·9Y1 . ˜
  #r ˜?'?œ o 3; M F ëw Y·o˜;,JFoo·Ei ï2fmcE ;1yx>u. î
J'o},îY "'¿f','if1EÃ?. , $*;foY1ë<;1o<< 4ëfî1L3i9"crëovA<àf lëë mb chai
,  foy lfërâ dg tsiûç?§ta,1;,-Apî1˜î1îé;oî1˜w;é<<Ãï<<ig;g;"
```



```
'braslde 1:1 Compzigîiies Il étoit accompagné de ibn _
'Once_,'· Frères dc" a"Mëre. Le ienne Prince avoit .
pour motif de fa fuite le delit de trouvetles moyens
îdëresiter fb;] Père du triûe état,. oùi il, étoit i'éduit.i:_
```

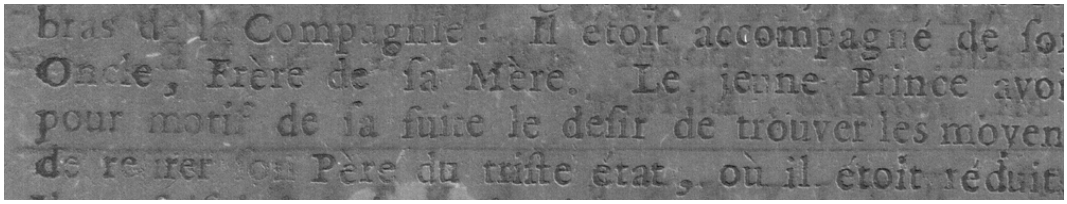


```
Not available
```

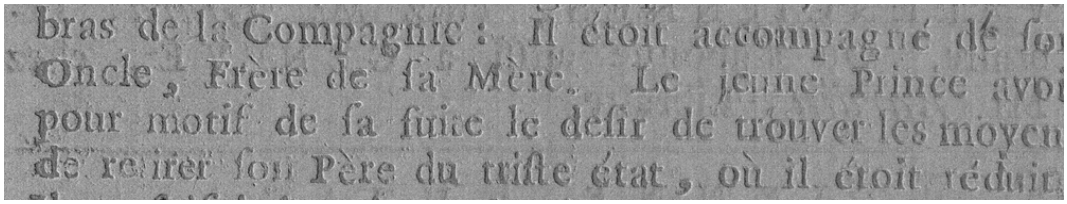


```
Not available
```

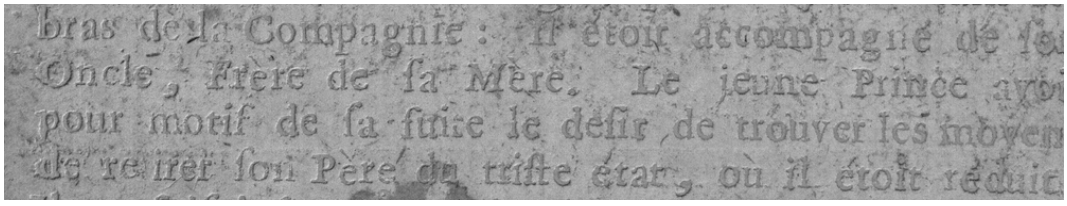
Fig. 6. Small extracts of the OCR results obtained on scanned document images: (a) segmentation with Niblack [27]; (b) segmentation with Sauvola et al. [30]; (c) Tonazzini et al. [37] source #1, Tonazzini et al. [37] source #2



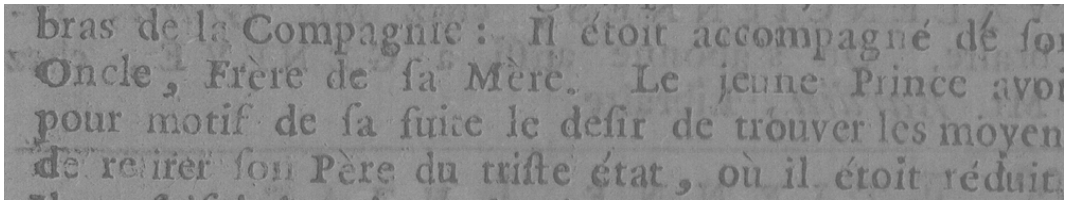
Not available



Not available



Not available



```
î1;;fCq mp:1g1iîc': · ·:  %1"§_ ëtQ§c _ilCCOâiH‘É$1>>gIllë Aidé fon _
Uncle Q Qlîrèrë <lc"(a'·î‘M‘Èxë; _, V L6 j,Eu11c· Prinèc qybi: M
ây qB;A;Ag§çPf>> dc_ fa fmcc Qç gicfîx dc'Érbp,vc1··1cSmbycmw
dë. rx:·>>xrçr    Par; dg t1;1ftç% grat. ,. pu, 11. étroit 1'éduim ,
```

Fig. 7. Small extracts of the OCR results obtained on scanned document images: (a) Tonazzini et al. [34] source #1 (b) Tonazzini et al. [34] source #2 (c) Tonazzini et al. [34] source #3 (d) Tonazzini et al. [34] all three sources combined.

Fig. 8. Restoration results on manuscripts. From left to right, top to bottom: input image, k-means, single MRF & α-exp. move [19], double MRF (proposed method), Tonazzini et al. [37] source #1, Tonazzini et al. [37] source #2, Tonazzini et al. [34] source #1, Tonazzini et al. [34] source #2 (source #3 not displayed).

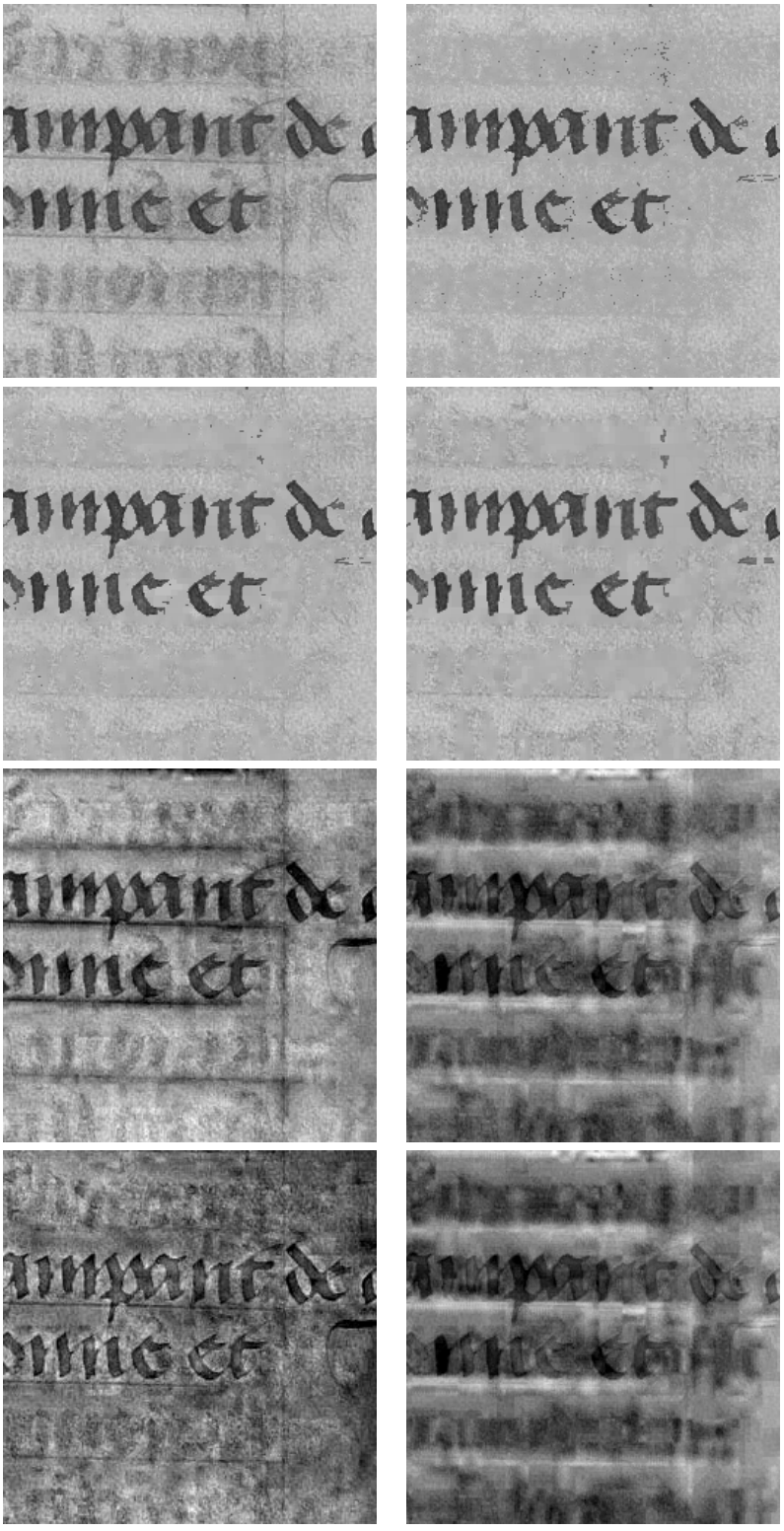


Fig. 9. Restoration results on manuscripts. From left to right, top to bottom: input image, k-means, single MRF & α-exp. move [19], double MRF (proposed method), Tonazzini et al. [37] source #2, Tonazzini et al. [37] source #1, Tonazzini et al. [34] source #3, Tonazzini et al. [34] source #1 (source #2 not displayed).

Method-type	Method		Recall (in %)	Prec. (in %)	Cost (abs.)	Size of dataset (in %)
—	No restoration		65.65	49.91	76,752	100
Context-free	Niblack [27] (segm. only)		-	-	-	-
	Sauvola et al. [30] (segm. only)		78.75	66.78	45,363	100
	K-Means (k=3)		78.57	69.43	40,375	100
Source-sep.	Tonazzini et al. [37] - src #1	‡	41.00	30.05	*74,819*	66
	Tonazzini et al. [37] - src #2	†	-	-	-	-
	Tonazzini et al. [34] - src #1	†	-	-	-	-
	Tonazzini et al. [34] - src #2	†	-	-	-	-
	Tonazzini et al. [34] - 3 sources	‡	50.52	33.90	*101,280*	89
MRF	Single MRF & α-exp. move [19]		81.99	72.12	36,744	100
	Double MRF (proposed method)		**83.23**	**74.85**	**32,537**	100

†Not available: lack of OCR performance makes a correct evaluation impossible
‡results obtained with a subset of the images only (absolute cost is not comparable).

TABLE II
OCR RESULTS ON A DATABASE OF 104 SCANNED DOCUMENT IMAGES: NON-RESTORED INPUT IMAGES AND DIFFERENT RESTORATION METHODS.

output is a little bit cleaner for the double MRF case.

The results of Niblack's algorithm and Sauvola et al.'s algorithm show the typical weaknesses of these approaches: Niblack (Figure 6a) produces spurious components, especially in areas with few text, and Sauvola (Figure 6b) tends to cutting characters into several parts due to its assumptions on the grayvalue distribution in the image.

Figures 6c and 6d show the first two source components of the non-blind source separation method [37] applied to the color components red and green of the color input image. All source separation results are shown without the post processing recommended by the authors (see below).

The second, blind method [34], shown in Figures 7a-c, delivers similar results: although we can identify a source component which does not include the verso text, the response itself is quite noisy and faint. Post-processing the image slightly improves the latter but tends to increase the noise. Figure 7d shows an image which corresponds to a grayscale conversion of a color image composed of the three different source components obtained with the color based method [34]. Although this result was not intended, as the verso component is still part of the image, the result seems to be better than the ones consisting of a single source component only. Surprisingly, this result is the only one which produces at least limited OCR output, whereas the other images do not produce anything meaningful.

In order to evaluate the amount of recognition improvement of the restoration method, we manually created groundtruth for the 104 images, and calculated the Levenstein edit distance between two strings [40], which finds the optimal transformation from on string into another with elementary operations (insertion, deletion, substitution) minimizing the global cost of these operations. Additionally, we calculated character recall and character precision derived from the transformation operation of this distance. Table II compares the measures for the different methods described above, as well as the recognition performance on not restored images. Note, that precision and recall are independant of the dataset size, whereas the total transformation cost is not.

We can see that all methods based on identifying the verso component (k-means and the two MRF methods, including the proposed one) are capable of significantly improving the recognition results compared to no restoration at all. Not surprisingly, regularizing the segmentation with *a priori* knowledge boosts the performance. Separating the regularization of the recto and verso side further improves recognition, gaining 1.2 percent points in recall compared to the single MRF and 2.7 percentage points in precision. Totally, compared to no restoration at all, the proposed method improves recognition at about 17 percentage points in terms of recall and around 25 percentage points in terms of precision.

Recognition on the results of Niblack's method produces only gibberish, probably because of the small

ghost objects it creates. Sauvola et al.'s method overcomes this problem and the recognition performance almost attains the quality of the three class segmentation of performed by the k-means algorithm.

Surprisingly, the recognition performance on the results of the two source separation results was very disappointing. We performed recognition experiments for both planes of the first method [37] and all three planes of the second method [34], respecting the author's recommendations to darken the images after applying the inverted mixture matrix. In a personal communication for the experiments in this paper, Prof. Tonazzini recommended subtracting the K component of the CMYK color decomposition. However, we obtained better results with a histogram stretch instead of the proposed method.

Unfortunately, the recognition performance on these results was not good enough to include it in the table. Most of the output was blank or gibberish, making an evaluation impossible. We managed to get some statistics on the first source plane of the first method, as well as on output images combining all three source planes of the second method. However, this was only possible when a subset of the dataset was removed. Even then, the results where not competitive.

Figures 8 and 9 show restoration results on two different manuscript images. The source separation methods remove more of the verso text in Figure 8, but unfortunately the contrast is very low and they are significantly disturbed by the JPEG artifacts in the input image. The performance shown in figure 9 reveils similar strengths and weaknesses, typical to the two types of approaches: the regularized segmentation approaches create crisp images but show localized artifacts, whereas the artifacts created by the source separation methods are more spread out across the image and seem to touch more of the low frequency components.

A. Computational complexity

The computational complexity of the proposed method is dominated by the inference part based on the minimum cut/maximum flow in a graph whose complexity is bounded by $O(|\mathcal{E}| * f)$, where $|\mathcal{E}|$ is the number of edges in the graph and f is the maximum flow. We use the graph cut implementation by Boykov and Kolmogorov [4] which has been optimized for typical graph structures encountered in computer vision and whose running time is nearly linear in running time in practice [5]. Table IV gives effective run times measured on a laptop computer equipped with an Intel Core 2 processor running at 2.5Ghz and 4GB of RAM (only one core was used). The running time of the proposed method is comparable to the running time of a single MRF with graph cut optimization and quite competitive given its restoration performance.

VIII. Conclusion and Outlook

We presented a method to separate the verso side from the recto side of a single scan of document images. The novelty of the method is the separation of the MRF prior into two different label fields, each of which regularizes one of the two sides of the document. This separation allows to estimate the verso pixels of the document which are covered by the recto pixels, which, again through the MRF prior, improves the estimation of the verso pixels not covered by recto pixels, thus increasing the performance of the regularization. We showed that this formulation leads to an efficient algorithm based on graph cuts.

The performance of the method has been evaluated on scanned document images from the 18^{th} century, showing that the restoration is able to improve the recognition performance of an OCR significantly, compared to non restored images but also compared to competing methods.

Involved in several digitization projects around the world, our team is currently looking into the following perspectives of this work:

- Creation of a homogeneous (adaptive) observation model, which increases the performance on larger images. This model needs to take into account several text colors, as well as other kinds of degradation (see section I).
- Creation of a hierarchical Markov model in the lines of [3][18][21] which is able to take into account larger neighborhood structures.
- Creation of a discriminative model, as for instance a CRF [22][20] adapted to the nature of the problem, allowing us to model dependencies between the observations.

IX. Acknowledgements

We thank Anna Tonazzini for providing us with the source code of the two source separation methods and her kind help in setting up the corresponding experiments as well as for the interesting discussions. We thank Madiha Nadri for the help in groundtruthing the document image dataset.

References

[1] H.S. Baird. Document image defect models and their uses. In *Proceedings of the International Conference on Document Analysis and Recognition*, pages 62–67, 1993.

Type	Method	1026×1587	2436×3320	
Context-free	Niblack [27]	0.6	2.9	†
	Sauvola et al. [30]	0.6	2.9	†
	K-Means (k=3)	1.9	10.5	†
Source-separation	Tonazzini et al. [34]	36.9	134.6	‡
	Tonazzini et al. [37]	17.0	74.5	‡
MRF	Single MRF [19]	7.2	34.9	†
	Double MRF	7.4	36.4	†
† Code in C++ ‡ Code in Matlab/GNU Octave				

TABLE IV

EXECUTION TIMES IN SECONDS FOR VARIOUS METHODS. THE REPLACEMENT OF THE BACKGROUND PIXELS IS INCLUDED IF THE METHOD IS BASED ON SEGMENTATION.

[2] J. Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society, Series B*, 36(2):192–236, 1974.

[3] C.A. Bouman and M. Shapiro. A Multiscale Random Field Model for Bayesian Image Segmentation. *IEEE Transactions on Image Processing*, 3(2):162–177, 3 1994.

[4] Y. Boykov and V. Kolmogorov. An experimental comparison of min-cut/max- flow algorithms for energy minimization in vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(9):1124–1137, 2004.

[5] Yuri Boykov, Olga Veksler, and Ramin Zabih. Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(11):1222–1239, 2001.

[6] B. Braathen and W. Pieczynski. Global and Local Methods of Unsupervised Bayesian Segmentaiton of Images. *Machine Graphics and Vision*, 2(1):39–52, 1993.

[7] M. Brown and W. Seales. Image restoration of arbitrarily warped documents. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(10):1295–1306, 2004.

[8] P. Charbonnier, L. Blanc-Féraud, G. Aubert, and M. Barlaud. Deterministic edge-preserving regularization in computated imaging. *IEEE Transactions on Imagea Processing*, 6(2):298–311, 1997.

[9] H. Derin and H. Elliott. Modeling and Segmentation of Noisy and Textured Images Using Gibbs Random Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 9(1):39–55, 1987.

[10] D.M.Greig, B.T. Porteous, and A.H. Seheult. Exact maximum a posteriori estimation for binary images. *Journal of the Royal Statistical Society B*, 51(2):271–279, 1989.

[11] H.-S. Don. A noise attribute thresholding method for document image binarization. *International Journal on Document Analysis and Recognition*, 4(2):131–138, 2000.

[12] K. Donaldson and G.K. Myers. Bayesian super-resolution of text in video with a text-specific bimodal prior. *International Journal on Document Analysis and Recognition*, 7(2-3):159–167, 2005.

[13] F. Drira, F. LeBourgeois, and H. Emptoz. Restoring ink bleed-through degraded document images using a recursive unsupervised classification technique. In *Proceedings of the 7th Workshop on Document Analysis Systems*, pages 38–49, 2006.

[14] E. Dubois and A. Pathak. Reduction of bleed-through in scanned manuscript documents. In *Proceedings of the Image Processing, Image Quality, Image Capture Systems Conference*, pages 177–180, 2001.

[15] S. Geman and D. Geman. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–741, 11 1984.

[16] J.M. Hammersley and P. Clifford. Markov fields on finite graphs and lattices. unpublished manuscript, 1968.

[17] T. Kanungo and R.M. Haralick an I. Philips. Global and local document degradation models. In *Proceedings of the International Conference on Document Analysis and Recognition*, pages 730–734, 1993.

[18] Z. Kato, M. Berthod, and J. Zerubia. A hierarchical Markov random field model and multitemperature annealing for parallel image classification. *Graphical Models and Image Processing*, 58(1):18–37, 1996.

[19] V. Kolmogorov and R. Zabih. What energy functions can be minimized via graph cuts? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(2):147–159, 2004.

[20] S. Kumar and M. Hebert. Discriminative random fields. *International Journal of Computer Vision*, 68(2):179–201, 2006.

[21] J.-M. Laferte, P. Perez, and F. Heitz. Discrete Markov image modelling and inference on the quad tree. *IEEE Transactions on Image Processing*, 9(3):390–404, 2000.

[22] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling data. In *International Conference on Machine Learning*, 2001.

[23] F. Lebourgeois, E. Trinh, B. Allier, V. Eglin, and H. Emptoz. Document images analysis solutions for digital libraries. *Proceedings of the first international workshop on document images analysis solutions for digital libraries*, 2004.

[24] Y. Leydier, F. LeBourgeois, and H. Emptoz. Serialized Unsupervised Classifier for Adaptative Color Image Segmentation: Application to Digitized Ancient Manuscripts. In *International Conference on Pattern Recogniton*, pages 494–497, 2004.

[25] S.Z. Li. *Markov Random Field Modeling in Image Analysis*. Springer Verlag, 2001.

[26] M. Melgosa. Testing cielab-based color-difference formulas. *Color Research & Application*, 25(1):49–55, 2000.

[27] W. Niblack. *An Introduction to Digital Image Processing*, pages 115–116. Prentice Hall, 1986.

[28] H. Nishida and T. Suzuki. Correcting show-through effects on document images by multiscale analysis. In *Proceedings of the International Conference on Pattern Recognition*, volume 3, pages 65–68, 2002.

[29] J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufman, San Mateo, 1988.

[30] J. Sauvola, T. Seppänen, S. Haapakoski, and M. Pietikäinen. Adaptive Document Binarization. In *International Conference on Document Analysis and Recognition*, volume 1, pages 147–152, 1997.

[31] M.I. Sezan and A.M. Tekalp. Survey of recent developments in digital image restoration. *Optical Engineering*, 29(5):393–404, 1990.

[32] G. Sharma. Show-through cancellation in scans of duplex printed documents. *IEEE Transactions on Image Processing*, 10(5):736–754, 2001.

[33] C.L. Tan, R. Cao, and P. Shen. Restoration of archival documents using a wavelet technique. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(10):1399–1404, 2002.

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

[35] A. Tonazzini, L. Bedini, and E. Salerno. A markov model for blind image separation by a mean-field em algorithm. *IEEE Transactions on Image Processing*, 15(2):473–482, 2006.

[36] A. Tonazzini and I. Gerace. Bayesian MRF-based blind source separation of convolutive mixtures of images. In *Proceedings of the 13th european signal processing conference*, 2005.

[37] 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.

[38] A. Tonazzini, S. Vezzosi, and L. Bedini. Analysis and recognition of highly degraded printed characters. *International Journal on Document Analysis and Recognition*, 6(4):236–247, 2003.

[39] O.D. Trier and A.K. Jain. Goal-Directed Evaluation of Binarization Methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(12):1191–1201, 1995.

[40] R.A. Wagner and M.J.Fisher. The string to string correction problem. *Journal of Assoc. Comp. Mach.*, 21(1):168–173, 1974.

[41] Q. Wang, T. Xia, C.L. Tan, and L. Li. Directional wavelet approach to remove document image interference. In *International Conference on Document Analysis and Recognition*, pages 736–740, 2003.

[42] C. Wolf and D. Doermann. Binarization of Low Quality Text using a Markov Random Field Model. In *Proceedings of the International Conference on Pattern Recognition*, volume 3, pages 160–163, 2002.

[43] J. Zhang. The mean field theory in em procedures for Markov random fields. *IEEE Transactions on Image Processing*, 40(10):2570–2583, 1992.

[44] L. Zhang, Y. Zhang, and C.L. Tan. An improved physically-based method for geometrical restoration of distorted document images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(4):728–734, 2008.

[45] Q. Zheng and T. Kanungo. Morphological degradation models and their use in document image restoration. In *Proceedings of the International Conference on Image Processing*, volume 1, pages 193–196, 2001.