An iterative graph cut optimization algorithm for a double MRF prior

Christian Wolf

An iterative graph cut optimization algorithm for a double MRF prior

Christian Wolf

Technical Report LIRIS

Laboratoire d'informatique en images et systèmes d'information - UMR 5205
Université de Lyon
INSA-Lyon, Bât. J.Verne; 20, Av. Albert Einstein
69621 Villeurbanne cedex, France
Tel.: 0033 4 72 43 63 08 Fax.: 0033 4 72 43 71 17
Email: christian.wolf@liris.cnrs.fr
Web: http://liris.cnrs.fr/christian.wolf

July 19, 2008

Abstract

In a previous publication we presented a double MRF model capable of separatly regularizing the recto and verso side of a document suffering from ink bleed through. In this paper we show that this model 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.

Keywords
Markov Random Fields, Bayesian estimation, Graph cuts, Document Image Restoration.

1 Introduction

In a previous paper [15] we presented a double Markov random field (MRF) model designed for document restoration, more specifically document ink bleedthrough removal, i.e the extraction and replacement of the verso ink traversing the paper and showing on the recto side. The original method used simulated anneling for optimization, which is theoretically known to converge to global solution but which is painfully slow in practice. In this paper we propose an efficient optimization technique based on graph cuts.

The paper is organized as follows: section 2 very briefly describes the graphical model, details on its motivation and its derivation can be found in [15]. Section 3 formulates the iterative optimization algorithm based on graph cuts. Section 4 illustrates the experiments performed to evaluate the results on a dataset of scanned documents, and section 5 finally concludes.

2 The graphical model

The model introduced in [15] consists of two hidden label fields, one for the recto side and one for the verso side, as well as a single observation field (the scanned image). We therefore consider 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 *back ground*. Formally, we have a 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^1_s, F^2_s 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^1_s and F^2_s 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 to nodes r and s if and only if $r \in N_s$ and $s \in N_r$.

The dependency graph (see figure 1) contains the following cliques types: first order and second order "intra-field" cliques in the subgraph F^1, first order

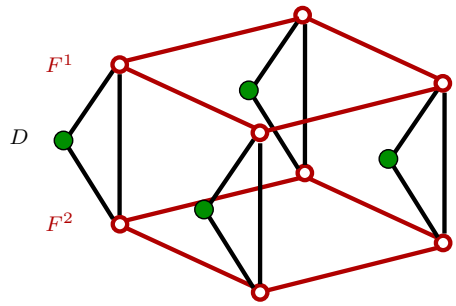


Figure 1: The dependency graph for a 2×2 pixel image. The model consists of the two label fields F^1 and F^2 ("empty" nodes) and the single observation field D (shaded nodes).

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.

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

$$P(f^1, f^2, d) = \frac{1}{Z} \exp\left\{-\left(U(f^1) + U(f^2) + U(f^1, f^2, d)\right)/T\right\} \quad (1)$$

The terms $U(f^1)$ and $U(f^2)$ correspond to two Potts models, one prior for each field:

$$\begin{aligned} U(f^1) + U(f^2) &= U(f^1, f^2) \\ &= \sum_{\{s\} \in \mathcal{C}_1} \alpha^1 f_s^1 + \sum_{\{s,s'\} \in \mathcal{C}_2} \beta_{s,s'}^1 \delta_{f_s^1, f_{s'}^1} \\ &+ \sum_{\{s\} \in \mathcal{C}_1} \alpha^2 f_s^2 + \sum_{\{s,s'\} \in \mathcal{C}_2} \beta_{s,s'}^2 \delta_{f_s^2, f_{s'}^2} \end{aligned} \quad (2)$$

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.

The term $U(f^1, f^2, d)$ corresponds to the data likelihood, which factorizes as follows:

$$P(d|f^1, f^2) = \prod_s \mathcal{N}(d_s;\ \boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s) \quad (3)$$

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 $\mu_r \neq \mu^r$ etc.):

$$\begin{aligned} \boldsymbol{\mu}_s &= \begin{cases} \boldsymbol{\mu}_r & \text{if } f_s^1 = \textit{text} \\ \boldsymbol{\mu}_v & \text{if } f_s^1 = \textit{background} \text{ and } f_s^2 = \textit{text} \\ \boldsymbol{\mu}_{bg} & \text{else} \end{cases} \\ \boldsymbol{\Sigma}_s &= \begin{cases} \boldsymbol{\Sigma}_r & \text{if } f_s^1 = \textit{text} \\ \boldsymbol{\Sigma}_v & \text{if } f_s^1 = \textit{background} \text{ and } f_s^2 = \textit{text} \\ \boldsymbol{\Sigma}_{bg} & \text{else} \end{cases} \end{aligned} \quad (4)$$

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.

3 The posterior probability and its maximization with graph cuts

Applying the Bayes rule, the posterior probability of the two label fields can be given as follows:

$$\begin{aligned} P(f^1, f^2|d) &= \frac{1}{P(d)} P(f^1, f^2) P(d|f^1, f^2) \\ &\propto P(f^1) P(f^2) P(d|f^1, f^2) \end{aligned} \quad (5)$$

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 it with the two Potts models and the data likelihood, we get the following energy potential function:

$$\begin{aligned} & U(f^1, f^2, d) \\ &= \sum_{\{s\} \in \mathcal{C}_1} \alpha^1 f_s^1 + \sum_{\{s,s'\} \in \mathcal{C}_2} \beta_{s,s'}^1 \delta_{f_s^1, f_{s'}^1} \\ &+ \sum_{\{s\} \in \mathcal{C}_1} \alpha^2 f_s^2 + \sum_{\{s,s'\} \in \mathcal{C}_2} \beta_{s,s'}^2 \delta f_s^2, f_{s'}^2 \\ &+ \sum_{\{s\} \in \mathcal{C}_1} \frac{1}{2} (d_s - \boldsymbol{\mu}_s)^T \boldsymbol{\Sigma}_s^{-1} (d_s - \boldsymbol{\mu}_s) \end{aligned} \quad (6)$$

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 (5) 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 [4], 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 [7]. 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 [1][2][3][5].

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:

$$\begin{aligned} & U(f^1, f^2, d) \\ &= \sum_{\{s\} \in \mathcal{C}_1} \left[\alpha^1 U_1(f_s^1) + \alpha^2 U_1(f_s^2)\right] \\ &+ \sum_{\{s,s'\} \in \mathcal{C}_2} \left[\beta_{s,s'}^1 U_2(f_s^1, f_{s'}^1) + \beta_{s,s'}^2 U_2(f_s^2, f_{s'}^2)\right] \\ &+ \sum_{\{s\} \in \mathcal{C}_1} U_2'(f_s^1, f_s^2; d_s) \end{aligned} \quad (7)$$

where $U_1(f_s) = f_s$, $U_2(f_s, f_{s'}) = \delta_{f_s, f_{s'}}$ and $U'_2(f^1_s, f^2_s; d_s) = \frac{1}{2}(d_s - \boldsymbol{\mu}_s)^T \boldsymbol{\Sigma}_s^{-1}(d_s - \boldsymbol{\mu}_s)$. We consider $U'_2(.,.;.)$ 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 (7) cannot be found using algorithms based on graph cuts. As shown by Kolmogorov et al. [5], 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:

$$E(0,0) + E(1,1) \leq E(0,1) + E(1,0) \qquad (8)$$

It can be easily seen that this is the case of the terms $U_2(.,.)$ in equation (7), but not necessarily for all terms $U'_2(.,.;.)$. According to the value of the observation d_s at site s, $U'_2(f^1_s, f^2_s; d_s)$ may be regular or not. In other words, if the observation likelihood for equal labels f^1_s and f^2_s 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. [2] for labeling problems with multiple labels ($|\Lambda| > 2$) and improved by Kolmogorov et al. [5]. In the original 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 label 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 my take 4 values (*background, recto, verso, recto-verso*) and the pairwise clique potentials are adapted accordingly. This optimization schedule may be improved by fixing only the variables whose site s is not regular, and jointly estimate the variables f^1_s and f^2_s 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 $U'_2(f^1_s, f^2_s; d_s)$ is regular or not:

$$H_s = \begin{cases} 1 & \text{if} \quad U'_2(0,0,d_s) + U'_2(1,1,d_s) \leq \\ & \quad\; U'_2(0,1,d_s) + U'_2(1,0,d_s) \\ 0 & \text{else} \end{cases} \qquad (9)$$

Figure 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^1_s for all s and f^2_s for $H_s = 0$, estimate optimal f^2 for all s and f^2_s for $H_s = 1$ maximizing $U^{\mapsto 2}(f^1, f^2, d, H)$
- Fix f^2_s for all s and f^1_s for $H_s = 0$, estimate optimal f^1 for all s and f^1_s for $H_s = 1$ maximizing $U^{\mapsto 1}(f^1, f^2, d, H)$

until *Convergence*

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 graph, we will rewrite the two functions by reordering some terms. W.l.o.g., 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 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:

$$
\begin{array}{l|ll}
 & U^{\mapsto 2}(f^1, f^2, d, H) & \\
1 & = & \sum_{\{s\}\in\mathcal{C}_1 : H_s = 0} \alpha^1 U_1(f_s^1) \\
2 & + & \sum_{\{s\}\in\mathcal{C}_1 : H_s = 1} \alpha^1 U_1(f_s^1) \\
3 & + & \sum_{\{s\}\in\mathcal{C}_1} \alpha^2 U_1(f_s^2) \\
4 & + & \sum_{\{s,s'\}\in\mathcal{C}_2 : H_s = 0 \wedge H_{s'} = 0} \beta^1_{s,s'} U_2(f_s^1, f_{s'}^1) \\
5 & + & \sum_{\{s,s'\}\in\mathcal{C}_2 : H_s = 1 \wedge H_{s'} = 1} \beta^1_{s,s'} U_2(f_s^1, f_{s'}^1) \\
6 & + & \sum_{\{s,s'\}\in\mathcal{C}_2 : H_s \neq H_{s'}} \beta^1_{s,s'} U_2(f_s^1, f_{s'}^1) \\
7 & + & \sum_{\{s,s'\}\in\mathcal{C}_2} \beta^2_{s,s'} U_2(f_s^2, f_{s'}^2) \\
8 & + & \sum_{\{s\}\in\mathcal{C}_1 : H_s = 0} U_2'(f_s^1, f_s^2; d_s) \\
9 & + & \sum_{\{s\}\in\mathcal{C}_1 : H_s = 1} U_2'(f_s^1, f_s^2; d_s)
\end{array}
\tag{10}
$$

Written in this notation, The energy functions can be directly translated into a cut graph using the method introduced by Kolmogorov et al. [5]. 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 (10) 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 (7)), but one of the two arguments is fixed in equation (10) 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 1 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.

4 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[1] for our experiences, mainly because it is easily scriptable. We performed some selected experiments with the product of the market leader, Abby Finereader 8[2], which performs slightly better without changing the ranking of the restoration methods.

We compared the proposed method with several competing methods. One group of algorithms purely exploits the fact that, according to the hypothesis that set recto pixels completely cover verso pixels, without taking into account interactions between neighboring pixels. Examples are the k-

[1] http://code.google.com/p/tesseract-ocr
[2] http://finereader.abbyy.com

n-edges for node pairs:	Weight	Line in eq. (10)
$F_s^2, F_{s'}^2 : (s,s') \in \mathcal{C}_2$	$-\beta^2_{s,s'}$	7
$F_s^1, F_{s'}^1 : (s,s') \in \mathcal{C}_2 \wedge H_s = 1 \wedge H_{s'} = 1$	$-\beta^1_{s,s'}$	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. (10)
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
$F_s^1 : H_s = 1$	$\sum_{s':H_{s'}=0 \wedge f^1_{s'}=1} \beta^1_{s,s'} - \sum_{s':H_{s'}=0 \wedge f^1_{s'}=0} \beta^1_{s,s'}$	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 1: The edges added to the cut graph for the proposed inference algorithm: each edge corresponds to a term in eq. (10). 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.

means clustering algorithm with k=3 clusters (followed by our restoration algorithm replacing verso pixels, explained in [15], as well as two thresholding algorithms. We chose two methods which represent the state of the art in adaptive thresholding: Niblack's algorithm [6] which performed best in a widely cited evaluation paper [13] as well as an improvement of Niblack's algorithm by Sauvola et al. [8]. 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 1, 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 [9][10][11][12]. 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 [9] and which we described in section 1 is applicable. The second method, introduced in [12] 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 itself.

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 7 and 8.

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 [5] and combined with the same parameter estimation and pre- and post-processing as our proposed method.

Figures 4 to 6 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 4b) is noisy as opposed to the MRF results, the double MRF (Figure 4d) improves the regularity of the single MRF (Figure 4c). The OCR 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 5a) produces spurious components, especially in areas with few text, and Sauvola (Figure 5b) tends to cutting characters into

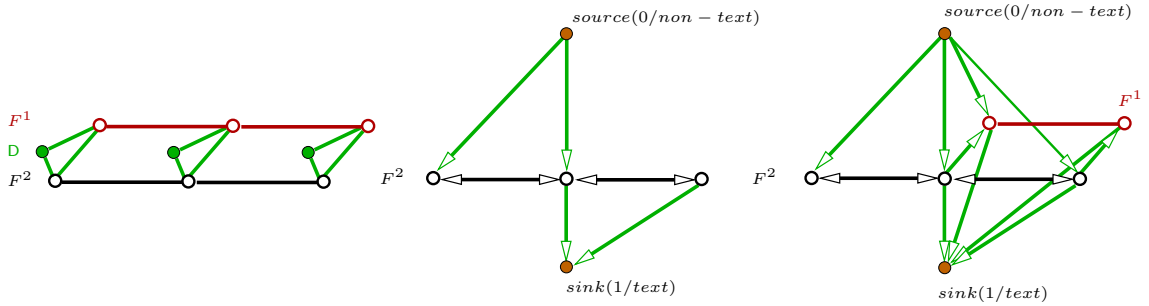


Figure 3: (a) The dependency graph of a simple model containing three pixels in a single row; (b) the cut graph for an α-expansion move like inference algorithm: inference of the verso pixels; (c) the cut graph for the proposed inference algorithm: joint inference of the verso pixels and of a subset of the recto pixels. In this exemple, 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 ($H_s = 0$).

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 [6] (segm. only)		-	-	-	-
	Sauvola et al. [8] (segm. only)		78.75	66.78	45,363	100
	K-Means (k=3)		79.82	68.43	42,675	100
Source-sep.	Tonazzini et al. [12] - src #1	‡	41.00	30.05	*74,819*	66
	Tonazzini et al. [12] - src #2	†	-	-	-	-
	Tonazzini et al. [9] - src #1	†	-	-	-	-
	Tonazzini et al. [9] - src #2	†	-	-	-	-
	Tonazzini et al. [9] - 3 sources	‡	50.52	33.90	*101,280*	89
MRF	Single MRF & α-exp. move [5]		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).

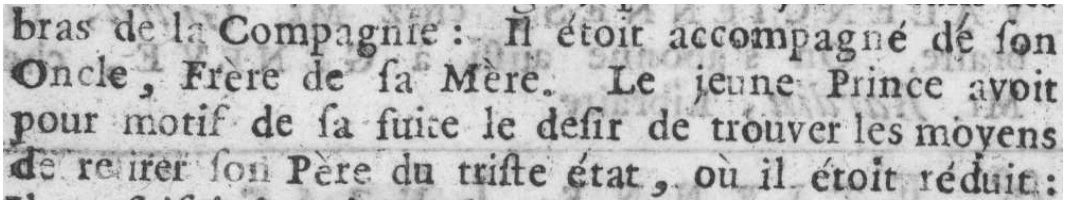
Table 2: OCR results on a database of 104 scanned document images: non-restored input images and different restoration methods.

several parts due to its assumptions on the gray-value distribution in the image.

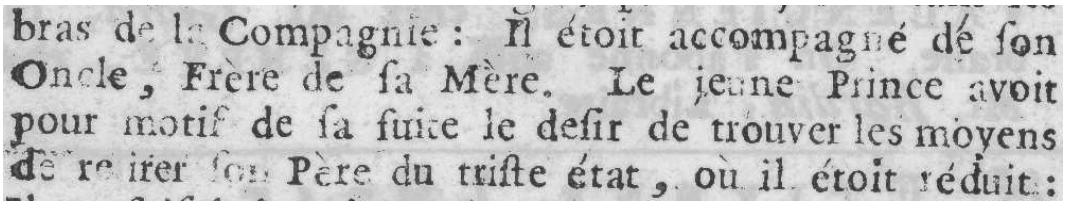
Figures 5c and 5d show the first two source components of the non-blind source separation method [12] applied to the color components red and green of the color input image. All source separation results are shown without the post processing recommended by the authors (see below).

The second, blind method [9], shown in Figures 6a-c, delievers 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 6d 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 [9]. 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.

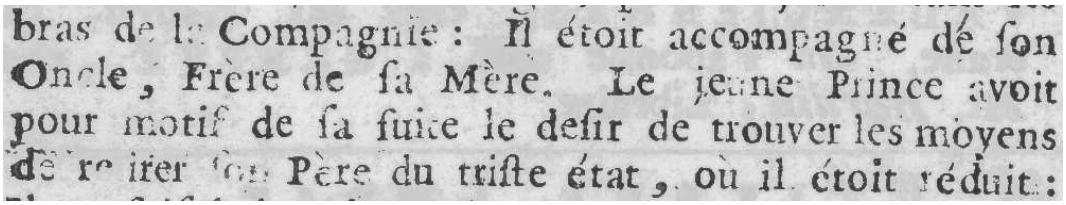
In order to evaluate the amount of recognition improvement of the restoration method, we manually created groundtruth for the 104 images, and calculated the Levenstein edit distance between two strings [14], which finds the optimal transformation from on string into another with elementary operations (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 2 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 indepen-



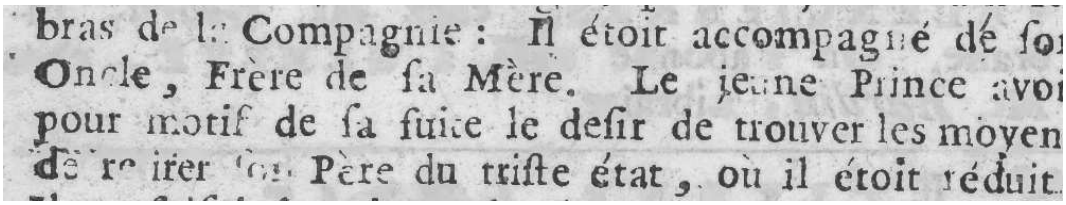
```
bras ‘d5:glgrComprzgnîc‘: |·Y· ;]L\ët4ôi;n àccoâniàagnë Lié ibn _
Uncle } ]F1‘èré‘dc"i‘a""MÈrèQ_ Lei j,cxn,c> Pxjnèci ziypir .
four motif dc fa fqizc lc déiïr de/trouver ies mbicns i
d’é"rë;ifèi‘ Qin Përç dg triiïcfgitar,. Oùlî1.é;Oit Yë(gliE1_’·
```



```
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;
```

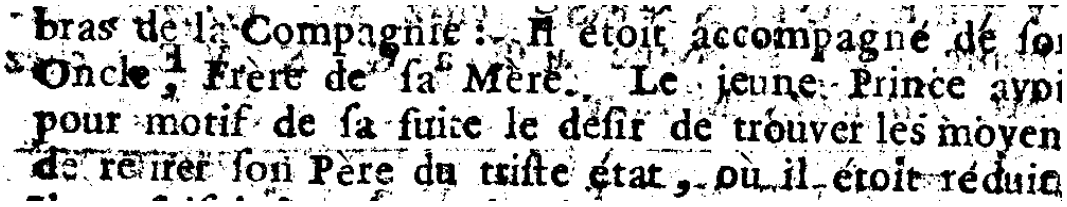


```
bras de l:îCOmP;rg1rrE: Il étoit accompagné dé (on _
Oncle ,’ Frère de fu Mere. Le ierzne Prince avoit .
pour motif de la fuite le delîr de trouver les mbvens
’dë‘r·· iter *}>;> Père du ttiftc état,. où il. étoit réduitr;
```

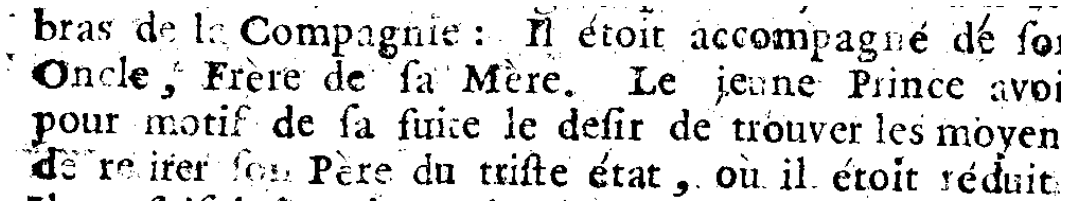


```
’ bras de l:1Comp;1gni·:: Il étoit accompagne dé Ton _
‘ Oncle, Frère de fa Mere. Le j.e;:ne· Prince avoit
pour motif dc fa fuite le defir de trouver les moyens
‘àië’r<<· iter ‘\}>>:· Père du triûe état,. ou il étoit réduit.:
```

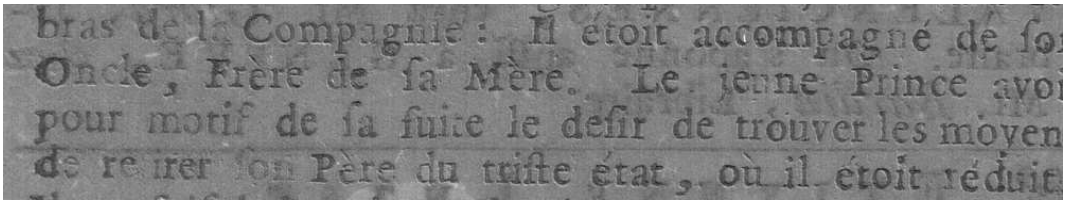
Figure 4: Small extracts of the OCR results obtained on scanned document images: (a) input image (no restoration); (b) k-means segmentation + restoration; (c) single MRF segmentation [5] + restoration; (d) double MRF (proposed method).



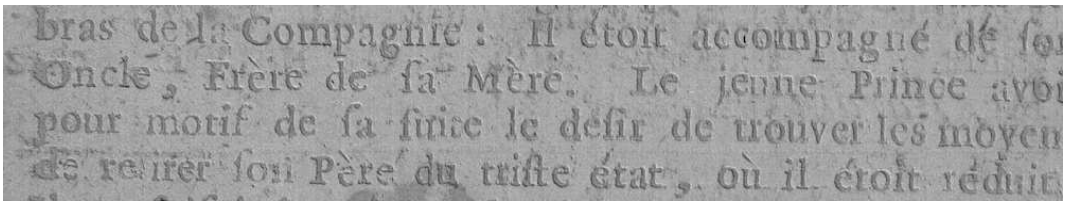

```
swmkg È \  >>· ggëfg   g. ;˜*·">>.<<ë'f >>'î?\P**$%? ,+Y·É·9Y1 . ˜
  #r ˜?'?œ o 3; M F ëw Y·o˜;,JFoo·Ei ï2fmcE ;1yx>u. î
J'o},îY "'¿f','if1EÃ?. , £*;foY1ë<;1o<< 4ëfî1L3i9"crëovA<àf lëë mb chai
,  foy lfërâ dg tsiûç?§ta,1;,-Apî1˜î1îé;oî1˜w;é<<Ãï<<ig;g;"
```




```
'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:_
```

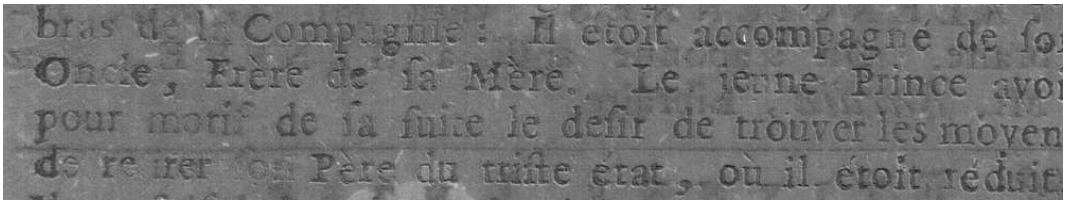


Not available

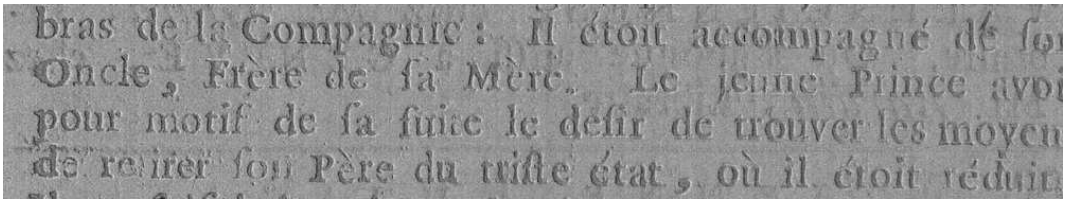


Not available

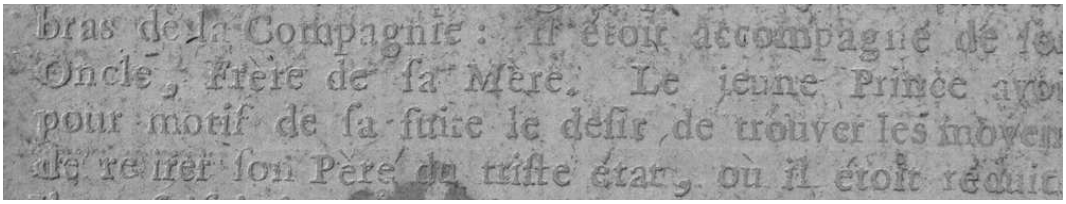
Figure 5: Small extracts of the OCR results obtained on scanned document images: (a) segmentation with Niblack [6]; (b) segmentation with Sauvola et al. [8]; (c) Tonazzini et al. [12] source #1, Tonazzini et al. [12] source #2



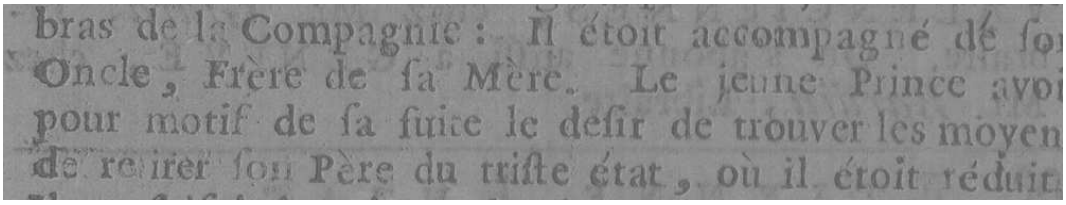
Not available



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 ,
```

Figure 6: Small extracts of the OCR results obtained on scanned document images: (a) Tonazzini et al. [9] source #1 (b) Tonazzini et al. [9] source #2 (c) Tonazzini et al. [9] source #3 (d) Tonazzini et al. [9] all three sources combined.

Figure 7: Restoration results on real data. From left to right, top to bottom: input image, k-means, single MRF & α-exp. move [5], double MRF (proposed method), Tonazzini et al. [12] source #1, Tonazzini et al. [12] source #2, Tonazzini et al. [9] source #1, Tonazzini et al. [9] source #2 (source #3 not displayed).

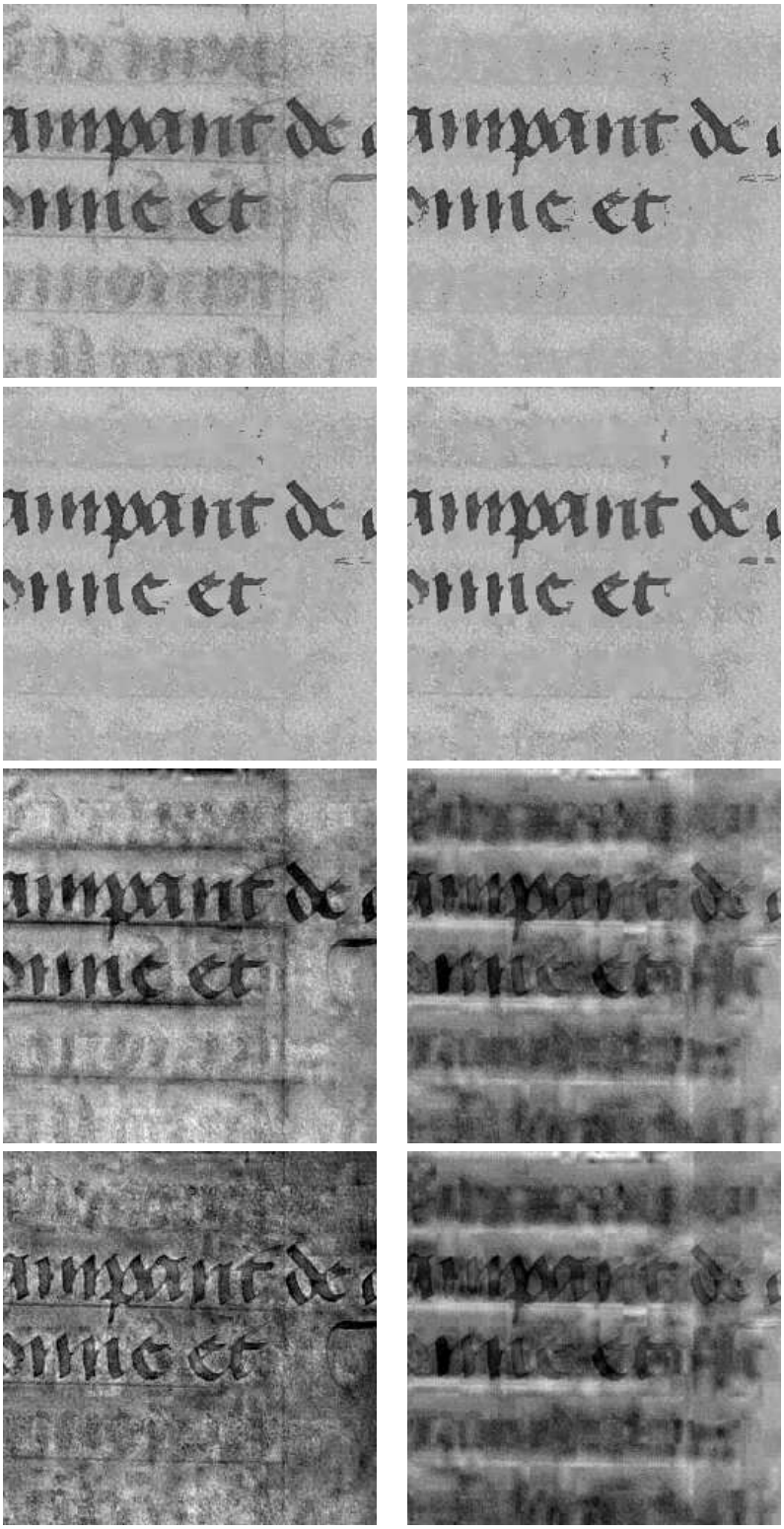


Figure 8: Restoration results on real data. From left to right, top to bottom: input image, k-means, single MRF & α-exp. move [5], double MRF (proposed method), Tonazzini et al. [12] source #2, Tonazzini et al. [12] source #1, Tonazzini et al. [9] source #3, Tonazzini et al. [9] source #1 (source #2 not displayed).

dant 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 [12] and all three planes of the second method [9], 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 7 and 8 show restoration results on two different manuscript images. The source separation methods remove more of the verso text in Figure 7, 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 8 reveils similar strengths and weaknesses, typical to the two types of approaches: the regularized segmentation approaches create crisp images but show localized artificacts, whereas the artificats created by the source separation methods are more spread out across the image and seem to touch more of the low frequency components.

Type	Method	1026 ×1587	2436 ×3320	
Context-free	Niblack [6]	0.6	2.9	†
	Sauvola et al. [8]	0.6	2.9	†
	K-Means (k=3)	1.9	10.5	†
Source-separation	Tonazzini et al. [9]	36.9	134.6	‡
	Tonazzini et al. [12]	17.0	74.5	‡
MRF	Single MRF [5]	7.2	34.9	†
	Double MRF	7.4	36.4	†
† Code in C++				
‡ Code in Matlab/GNU Octave				

Table 3: Execution times in seconds for various methods. The replacement of the background pixels is included if the method is based on segmentation.

4.1 Computational complexity

The computational complexity of the proposed method is dominated by the inference part with minimum cut/maximum flow 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 [1] which has been optimized for typical graph structures encountered in computer vision and whose running time is nearly linear in running time in practice [2]. Table 3 gives effective run times measured on a 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.

5 Conclusion and Outlook

In this paper 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.

6 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.

References

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

[2] 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, November 2001.

[3] 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.

[4] 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.

[5] 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.

[6] W. Niblack. *An Introduction to Digital Image Processing*, pages 115–116. Englewood Cliffs, N.J.: Prentice Hall, 1986.

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

[8] 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.

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

[10] 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.

[11] 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.

[12] 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.

[13] 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.

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

[15] C. Wolf. Document ink bleed-through removal with two hidden Markov random fields and a single observation field. Technical Report LIRIS RR-2006-019, Laboratoire d'Informatique en Images et Systèmes d'Information, INSA de Lyon, France, 2006.