# Improving recto document side restoration with an estimation of the verso side from a single scanned page

Christian Wolf
*Laboratoire d'informatique en images et systèmes d'information - UMR 5205*
*INSA-Lyon, Bât. J.Verne; 20, Av. Albert Einstein*
*69621 Villeurbanne cedex, France*
*christian.wolf@liris.cnrs.fr*

## Abstract

*We present a new method for blind document bleed through removal based on separately restoring the recto and the verso side. The segmentation algorithm is based on separate Markov random fields (MRF) which results in a better adaptation of the prior to the content creation process (e.g. superimposing two pages), and the improvement of the estimation of the verso pixels through an estimation of the verso pixels covered by recto pixels. The labels of the initial recto and verso clusters are recognized without using any color or gray value information. The proposed method is evaluated empirically as well as through OCR improvement.*

## 1. Introduction

In this paper we concentrate on the problem of 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*, *back ground*, or eventually *recto-and-verso* (simultaneously). Techniques proposed for this problem are independent components analysis [14, 15], wavelets [17] and other frequency based approaches [10], PDEs [4], non-blind techniques requiring alignement [12, 5, 13] and model guided thresholding [2].

MRF regularization has already been used for this kind of problem [16, 3, 19]. It allows to create a statistical model adapted to the knowledge on the degradation process as well as the prior knowledge on the image contents. However, previous methods treated recto/verso separation in the same way as conventional image segmentation. In this paper we show that the performance can be improved when knowledge of the existence of two different document sides is taken into account and modeled accordingly.

This paper is organized as follows. Section 2 proposes a dependency graph for the joint probability distribution of the full set of variables — hidden recto and verso as well as observed. Section 3 outlines the parameter estimation procedure as well as the minimization of the posterior probability. Finally, section 4 presents the experimental results and section 5 concludes.

## 2. The graphical model

MRFs exploit assumptions, as for instance smoothness criteria — homogeneous areas are considered more probable then frequent label changes. This is not justified when the observed image is the result of the superposition of two or more "sources" in which case, *a priori* knowledge may be available for each of the source images, but not for the mixture of these images.

We therefore propose a prior model with two different hidden label fields : one for the recto side ($F^1$) and one for the verso side ($F^2$), resulting in a model with two hidden variables for each pixel (*recto* and *verso*) and a configuration space of two possible labels for each variable (*text* and *non-text*). The advantages of this formulation are two-fold: first, the priors regularize fields which directly correspond to the natural process "creating" the contents (e.g. hand writing letters). Second, 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, thus increasing the performance of the algorithm.

Markov Random Fields [6, 9] are non causal models on undirected graphs which treat images as stochastic

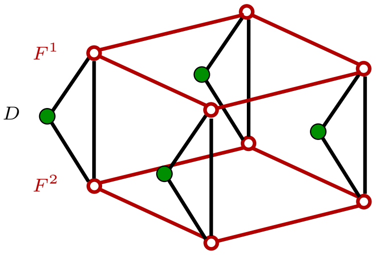


**Figure 1. The dependency graph of the model created for a 4 pixel image.**

processes and assign a probability to each realization of a set of variables $f$, where the joint probability distribution follows a Gibbs distribution [7] defined on the maxima cliques of the graph:

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

where $Z$ is a normalization constant, $T$ is a temperature factor assumed to be 1, $U(f) = \sum_{c\in\mathcal{C}} V_c(f)$ 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 $c$.

In general, MRFs model the a priori knowledge on the hidden labels and are combined with a likelihood term on the observed variables in the framework of Bayesian estimation. Given the nature of our problem, we preferred to interpret the whole set of hidden and observed variables as a MRF. 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 $E$, where $V$ 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 neighbors of $s$ are denoted as $N_s$.

The proposed dependency graph (see figure 1) contains the following cliques types: first order and second order cliques in the subgraph $F^1$, first order and second order 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:

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

The last equality indicates the Bayesian interpretation of the problem: the first factor corresponds to the prior knowledge and the second factor corresponds to the data likelihood determined by the observation/degradation model, where the prior probability is actually the product of the two probabilities of the two fields $F^1$ and $F^2$: $P(f^1, f^2) = P(f^1)P(f^2)$. 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 grayvalue 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.

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

$$\begin{aligned}
&U(f^1) + U(f^2) = U(f^1, f^2)\\
&= \sum_{\{s\}\in\mathcal{C}_1} \alpha^1 f^1_s + \sum_{\{s,s'\}\in\mathcal{C}_2} \beta^1_{s,s'}\delta_{f^1_s, f^1_{s'}}\\
&+ \sum_{\{s\}\in\mathcal{C}_1} \alpha^2 f^2_s + \sum_{\{s,s'\}\in\mathcal{C}_2} \beta^2_{s,s'}\delta_{f^2_s, f^2_{s'}}
\end{aligned} \tag{3}$$

where $\mathcal{C}_1$ is the set of single site cliques, $\mathcal{C}_2$ is the set of pair site cliques and $\delta$ 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:

$$P(d|f^1, f^2) \quad = \quad \prod_s \mathcal{N}(d_s;\, \boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s) \tag{4}$$

where $\boldsymbol{\mu}_s$ is the mean for class $f_s$ and $\boldsymbol{\Sigma}_s$ is the covariance matrix for class $f_s$ given as follows:

$$\begin{aligned}
\boldsymbol{\mu}_s &= \begin{cases} \boldsymbol{\mu}_r & \text{if } f^1_s = \textit{text} \\ \boldsymbol{\mu}_v & \text{if } f^1_s = \textit{background} \text{ and } f^2_s = \textit{text} \\ \boldsymbol{\mu}_{bg} & \text{else} \end{cases}\\
\boldsymbol{\Sigma}_s &= \begin{cases} \boldsymbol{\Sigma}_r & \text{if } f^1_s = \textit{text} \\ \boldsymbol{\Sigma}_v & \text{if } f^1_s = \textit{background} \text{ and } f^2_s = \textit{text} \\ \boldsymbol{\Sigma}_{bg} & \text{else} \end{cases}
\end{aligned} \tag{5}$$

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. Parameter estimation and energy minimization

In this work we chose to estimate the parameters in a supervised manner on the median filtered k-means initializations of the label field with least squares estimation, which was first proposed by Derin et al. [1]. 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. The parameters of the verso field are directly calculated from the parameters of the recto field based on the assumption that, statistically speaking, the verso field is a flipped version of the recto field. The parameters of the observation model are estimated using the classical maximum likelihood estimators (the empirical means and covariances).

In order to minimize the posterior energy we use simulated annealing [6] and preliminary results are given for graphcuts [8]. More details on the latter algorithm for double MRF priors are given in [18].

## 4. Experimental results

To evaluate our algorithm we decided to perform tests on synthetic data with ground truth as well as real data. In all cases we performed the experiments on gray scale images. The label meanings of the initalized hidden fields are recognized without using any information from the histogram of the images. Instead, statistics on label changes of neighboring pixels assuming that connected components of the recto label tend to cover other components.

**Experiments on synthetic images** We created synthetic images according to the degradation model described in section 2. For each synthetic image, two perfect images have been superimposed and Gaussian noise with different variances has been added. Table 1a shows the results of our algorithm as well as set of other algorithms : a simple k-means algorithm, a MRF segmentation with a single label field and a double MRF segmentation. We see that the double MRF method outperforms the other methods.

**Experiments on scanned documents - restoration** The method has also been tested on real document images, a small image of the printed part is shown in figure 2, the dataset also contains manuscripts.As can be seen, the MRF regularization is capable of removing many artifacts present in the k-means segmented image. The double MRF method further decreases the number of artifacts. There is no ground truth for these kind of images, so we presented a set of images to 16 different people (of course after randomly shuffling the result images) and let them rank the 3 result images by perceived quality. The results of these $N$=64 tests is shown in Table 1b. Our method has been ranked first 33 times against 18 times (k-means) and 13 (single-MRF).

| Noiselevel | $\sigma$=10 | $\sigma$=15 | $\sigma$=20 |
|---|---|---|---|
| K-Means | 0.25 | 1.40 | 3.56 |
| Single-MRF | 0.03 | 0.23 | 0.73 |
| Double-MRF | **0.01** | **0.08** | **0.31** |

(a)

| | complete | | | no single-MRF | |
|---|---|---|---|---|---|
| Method | $1^{st}$ | $2^{nd}$ | $3^{rd}$ | $1^{st}$ | $2^{nd}$ |
| K-Means | 18 | 10 | 36 | 21 | 43 |
| Single-MRF | 13 | 39 | 12 | - | - |
| Double-MRF | **33** | 15 | 16 | **43** | 21 |
| Total | 64 | 64 | 64 | 64 | 64 |

(b)

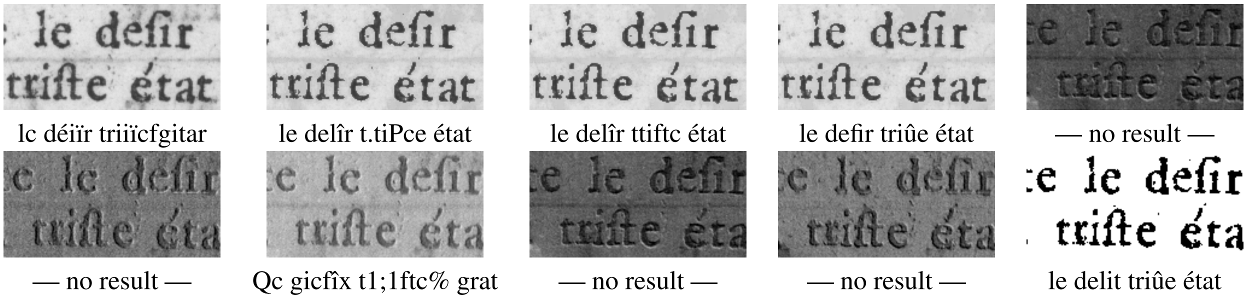
**Table 1. (a) classification error (in %) on synthetic images with different noise levels (b) results of the empirical tests. MRF results obtained with sim. annealing.**

The surprising performance of the k-means algorithm is due to the fact, that it tends to keep more recto pixels than the single MRF one. To test the statistical significance of this result, let us assume the following Null hypothesis $H_0$*: "the method double-MRF is as efficient as the other two methods"*.

We can conclude from the data that the method is not less efficient, it suffices therefore to reject $H_0$. Our test statistics will be $U$ = the number of times the method *double-MRF* is ranked first. Assuming $H_0$, the probability for a method to be ranked first for an image is $\pi=\frac{1}{3}$, $U$ is therefore distributed $\mathcal{B}(64, \frac{1}{3})$ (Binomial). The probability of getting the value $U$=33 or a more extreme value is 0.00197. Given a standard one-sided significance level of $\alpha$=0.025, $H_0$ must be rejected.

To prove that the method is better than the second ranked one, we look at the results after ignoring the method ranked as third, also shown in Table 1b, and creating a new $H_0$*: "The method double-MRF is as efficient as the method K-means"*. $U$ is distributed $\mathcal{B}(64, 0.5)$, the probability for $U >= 43$ is 0.004073. The *double-MRF* is thus more efficient than *K-Means*.

**Experiments on scanned documents - OCR** Google's OCR Tesseract has been applied to low quality documents and tested against several competing algorithms (shown in figure 2). The double MRF performs best with a character recall and precision of 83.23% and 74.85% against the closest competitor, the single MRF, achieving 81.99% and 72.12%. More



lc déiïr triiïcfgitar

le delîr t.tiPce état

le delîr ttiftc état

le defir triûe état

— no result —

— no result —

Qc gicfîx t1;1ftc% grat

— no result —

— no result —

le delit triûe état

**Figure 2. Restoration and OCR results on real data, from left to right, top to bottom: input image, k-means, single MRF + $\alpha$-exp. move [8], double MRF, 3× Tonazzini et. al [14] (plane #1, plane #2, all 3 planes), 2× Tonazzini et al. [15] (plane #1, plane #2), Sauvola et al. [11]. MRF results obtained with graphcuts optimization.**

details on the experimental setup is given in [18].

## 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. The method has been validated empirically as well as through the improvement of OCR results.

## References

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

[2] H.-S. Don. A noise attribute thresholding method for document image binarization. *I.J. on Doc. Anal. and Rec.*, 4(2):131–138, 2000.

[3] K. Donaldson and G. Myers. Bayesian super-resolution of text in video with a text-specific bimodal prior. *I.J. on Doc. Anal. and Rec.*, 7(2-3):159–167, 2005.

[4] F. Drira, F. Lebourgeois, and H. Emptoz. OCR accuracy improvement through a PDE-based approach. In *I.C. on Doc. Anal. and Rec.*, volume 2, pages 1068–1072, 2007.

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

[6] S. Geman and D. Geman. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Tr. on PAMI*, 6(6):721–741, 11 1984.

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

[8] V. Kolmogorov and R. Zabih. What energy functions can be minimized via graph cuts? *IEEE Tr. on PAMI*, 26(2):147–159, 2004.

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

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

[11] 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.

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

[13] C. Tan, R. Cao, and P. Shen. Restoration of archival documents using a wavelet technique. *IEEE Tr. on PAMI*, 24(10):1399–1404, 2002.

[14] A. Tonazzini and L. Bedini. Independent component analysis for document restoration. *I.J. on Doc. Anal. and Rec.*, 7(1):17–27, 2004.

[15] A. Tonazzini, E. Salerno, and L. Bedini. Fast correction of bleed-through distortion in grayscale documents by a blind source separation technique. *I.J. on Doc. Anal. and Rec.*, 10(1):17–25, 2007.

[16] A. Tonazzini, S. Vezzosi, and L. Bedini. Analysis and recognition of highly degraded printed characters. *I.J. on Doc. Anal. and Rec.*, 6(4):236–247, 2003.

[17] Q. Wang, T. Xia, C. Tan, and L. Li. Directional wavelet approach to remove document image interference. In *I.C. on Doc. Anal. and Rec.*, pages 736–740, 2003.

[18] C. Wolf. An iterative graph cut optimization algorithm for a double mrf prior. Technical Report LIRIS RR-2008-017, Laboratoire d'Informatique en Images et Systèmes d'Information, INSA de Lyon, France, 2008.

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