2009 10th International Conference on Document Analysis and Recognition

Document image binarisation using Markov Field Model

Thibault Lelore, Frédéric Bouchara
UMR CNRS 6168 LSIS
Southern University of Toulon-Var
BP 20132, 83957 La Garde Cedex
thibault-lelore@etu.univ-tln.fr, bouchara@univ-tln.fr

Abstract

This paper presents a new approach for the binarization of seriously degraded manuscript. We introduce a new technique based on a Markov Random Field (MRF) model of the document. Depending on the available information, the model parameters (clique potentials) are learned from training data or computed using heuristics. The observation model is estimated thanks to an expectation-maximization (EM) algorithm which extracts text and paper's features. The performance of the proposition is evaluated on several types of degraded document images where considerable background noise or variation in contrast and illumination exist.

1 Introduction

Binarization is one of the initial steps of most document image analysis and understanding systems (Initial classification, Optical character recognition, etc.). It plays a key role in document processing since its performance affects quite critically the degree of success in a subsequent character segmentation and recognition. Degradations appear frequently and may occur due to several reasons which range from the acquisition source type to environmental conditions.

Since the earlier work based on global thresholding, new methods have been proposed using local computation. Several methods of the literature compute the local threshold using statistical parameters. Thus Bernsen [2] proposed a method based on the minimal and the maximal values of a local window. Other works used the standard deviation and the mean [11, 13]. In [5], Gatos proposed a method in two main steps: the gray level of the background is first computed thanks to Sauvola's algorithm and used to binarize the image efficiently.

Recently, new methods based on a Markov Random Field (MRF) applied on degraded multimedia document have been proposed [14, 8].

In this paper we propose a new algorithm for the binarization of textual document. Our model involves not only local parameters in order to deal with non-uniform background, but also global ones which make it robust against noise. The proposed approach is based on a Bayesian framework using a MRF model of the image.

The paper is organized as follows. In the next section we present the global model of our approach. Section 2 and 3 are respectively devoted to the prior model and the description of the estimation process. Finally, in section 5, the performance of the proposed algorithm is assessed and compared with other methods previously published.

2 Model

We simply model the image as the noisy mixture of the background, noted b and the text, noted t. These processes are defined on the finite grid of sites $\mathcal{S}$ and we shall note in the sequel o_s the value of o on the site s of this grid.

Thus we have:

$$o_s = (1 - z_s).(b_s + n_s^b) + z_s.(t_s + n_s^t) \qquad (1)$$

In the previous equation, z_s is a variable such that $z_s = 1$ if s is a text pixel and $z_s = 0$ otherwise. n^b and n^t are two centered random processes which represent observational and model noises of respectively b and t.

The binarization problem is formalized through a MAP optimization, that is, the unknown field z is estimated by maximizing the conditional probability $P(z/o)$:

$$z = \arg\,.\max_z P(z/o) = \arg\,.\max_z P(o/z)P(z) \qquad (2)$$

The observations o_s are supposed to be conditionally independent given z. The likelihood $P(o/z)$ is hence given by the following relation:

978-0-7695-3725-2/09 $25.00 © 2009 IEEE
DOI 10.1109/ICDAR.2009.117

$$P(o/z) = \prod_{s \in \mathcal{S}} f_s(o_s/z_s) \quad (3)$$

The law $P(z)$ encodes our a priori knowledge of labelling z thanks to a Markov Random Fields (MRF) model [6, 3]. Due to the equivalence between MRF and the Gibbs Random Fields stated by the Hammersley-Clifford theorem, the expression of $P(z)$ is given by a Gibbs distribution:

$$P(z) = \frac{1}{W} \exp\left(-U(z)\right) \quad (4)$$

where W, the partition function, is a constant of normalization and the energy function $U(z)$ is defined as the sum of potential functions:

$$U(z) = \sum_{d \in \mathcal{D}} V_d(z) \quad (5)$$

$\mathcal{D}$ is the set of all cliques defined by a neighborhood system: a clique is a set of sites in which all pairs of sites are mutual neighbors. Note that a given potential function $V_d(z)$ depends only on the state of z at sites in the clique d.

To apply the rule given by eq. (2), we need the expressions of the likelihood $P(o/z)$ and the potential functions $V_d(z)$. In the parametric case, $P(o/z)$ is supposed to be an analytic function described by a vector parameter that we shall note Ψ in the sequel. In the proposed algorithm, these two kinds of parameters (i.e. Ψ and the set of potential functions) are estimated thanks to two different processes: online in the case of Ψ and off-line for the potential functions. In the next section, we describe the prior model of our algorithm. The estimation of the likelihood parameters achieved thanks to the EM algorithm, is described in the next section.

3 Prior model

The prior model represents the contextual information introduced by the Markov model. This information is defined by the set of potential functions $V_d(z)$ associated with each clique configuration.

When a training set is available, these potential functions are usually estimated thanks to a statistic treatment with a classical formula:

$$Vd(z) = -log(P(Conf)) \quad (6)$$

The absolute probability of a clique labeling $P(Conf)$ can be estimated from the frequency of its occurrence in the training images or, if the configuration is not found, by using the probability of close configurations[10].

However, this method has two main drawbacks: it does not add new information and it need a minimal training set, which is, for some kind of documents such as old manuscripts, often unavailable. To estimate V_d without learning, we proposed a different approach based uniquely on simple heuristic rules. Such a method has been previously applied by Messelodi et al. [9] and Kim et al. [7] for the case of text detection.

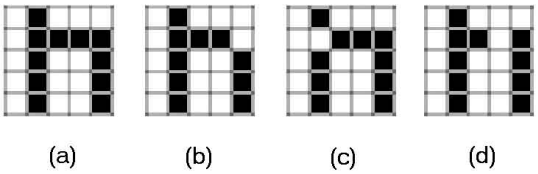

Figure 1. Example of Cliques. (a) $V_d^2(z) = 1$**, (b)** $V_d^2(z) = 2$**, (c)** $V_d^2(z) = 3$**, (d)** $V_d^2(z) = 4$

We define the potential function V_d as the sum of two different terms $V_d^1(z)$ and $V_d^2(z)$.

The term V_d^1 is introduced to reduce the effect of noise. The purpose of the second term V_d^2 is to improve the character connectivity.

Remove the noise. We have identified two types of noise: isolated pixels and stamps. The function $V_d^1(z)$ is defined in order to penalize configurations corresponding to important number of text pixel and isolated pixels. This definition is efficient to generate an important gap between the energies of the two configurations ($z_s =$ text or $z_s =$ background) when the number of text pixels is too low or too high.

Improve the character connectivity. The definition of V_d^2 is based on the number of 8 and 4 connected components (respectively noted $NbCC_4$ and $NbCC_8$ in the sequel) which are good indicators of the character connectivity. We define $V_d^2(z)$ with the following formula:

$$V_d^2(z) = NbCC_8.NbCC_4 \quad (7)$$

As an illustration of the behavior of this function, we give in figure 1 some examples of V_d^2 for different kinds of clique.

From these definitions, we shall consider two different cases:

When a training set is available the potential $V_d(z)$ will be computed by using a linear combination of the values respectively given by the previous rules and the learning (usually set to 0.5-0.5). The learning potential associated with the non encountered configurations will be simply set to a constant value.

In the non supervised cases, the potential V_d will be computed by using only the heuristic rules.

4 Parameter estimation using the EM algorithm

The definition of the observational model as given by equ. 1 makes it similar to an independent mixture model. Usually, in such a case, the estimation of the underlying laws are achieved thanks to the EM algorithm [4] which estimates a vector parameter Ψ by maximizing at each iteration q the quantity given by:

$$E\left[\log\left(P(o,z/\Psi^{(q)})\right)\right] \quad (8)$$

This algorithm is divided in two parts: the computation of the previous expression (the expectation step E), and its maximization with respect to Ψ (the maximization step M). In our case, Ψ is composed with two kinds of parameters: a local parameter defined by the two hidden processes t and b and a global parameter θ which depends on the model of the noise.

Using equations (2-5) we can write:

$$\begin{aligned} &E\left[\log\left(P(o,z/\Psi^{(q)})\right)\right] = \\ &\sum_{s\in\mathcal{S}}\sum_{z_s\in\{0,1\}} P(z_s/o)log\left(f_s(o_s/z_s)\right) + \\ &\sum_{c} E[V_d] - \log W \end{aligned} \quad (9)$$

Step (M) is hence equivalent to the maximization of the function Q defined by:

$$Q(\Psi/\Psi^{(q)}) = \sum_{s\in\mathcal{S}}\sum_{z_s\in\{0,1\}} P(z_s/o)log\left(f_s(o_s/z_s)\right) \quad (10)$$

Following Qian and Titterington [12], we propose to compute an approximation of this expression in two steps:

- *Step (1)*: In this step, an ICM algorithm [3] is carried out to compute equation (2). Each site s is visited lexicographically and updated by applying the rule:

$$z_s^{(q)} = \arg.\min_{z_s}\left[-ln(f_s(o_s/z_s)) + \sum_{d\ni z} V_d(z)\right] \quad (11)$$

  where (q) denotes the iteration.

- *Step (2)*: The probability $P(z_s/o)$ is computed by using the approximation of the pseudo likelihood proposed by Besag [3]:

$$\begin{aligned} P(z_s=1/o)^{(q)} &= P(z_s=1/o, z_{\partial_s}^{(q)}) \quad (12)\\ &= \frac{\pi_{ts}^{(q)} f_s(o_s/z_s=1;\theta^{(q)},t_s^{(q)})}{P(o_s/\theta^{(q)},b_s^{(q)})} \end{aligned}$$

where:

$$\begin{aligned} &P(o_s/\theta^{(q)},b_s^{(q)}) = \pi_{ts}^{(q)} f_s(o_s/z_s=1;\theta^{(q)},t_s^{(q)}) \\ &+\pi_{bs}^{(q)} f_s(o_s/z_s=0;\theta^{(q)},b_s^{(q)}) \end{aligned} \quad (13)$$

and:

$$\begin{aligned} &\pi_{ts}^{(q)} = \\ &\frac{\exp\left(\sum_{d\in s} V_d(z_{z_s=1}^{(q)})\right)}{\exp\left(\sum_{d\in s} V_d(z_{z_s=1}^{(q)})\right) + \exp\left(\sum_{d\in s} V_d(z_{z_s=0}^{(q)})\right)} \\ &\pi_{bs}^{(q)} = 1 - \pi_{ts}^{(q)} \end{aligned} \quad (14)$$

The probability $P(z_s = 0/o)^{(q)}$ is defined similarly and we shall note, in the sequel, $E_{si}^{(q)} = P(z_s = i/o)^{(q)}$ with $i \in \{0,1\}$.

Without lost of generality, we will now describe the proposed algorithm in the case of a Gaussian noise model. The vector parameter θ is defined by the covariance matrices, Σ_b and Σ_t, of the processes n^b and n^t. The functions b and t can be viewed as the expectations of the non centered processes defined respectively by $(b+n^b)$ and $(t+n^t)$.

During the Maximization (M) iteration, θ, t and b are computed. The last two parameters are estimated locally for each site. Thus, the estimation of the j^{th} component of these two vectors is given by:

$$t_s^{(q)}(j) = \frac{\sum_{N(s)} E_{s1}^{(q)} o_s(j)}{\sum_{N(s)} E_{s1}^{(q)}} \quad (15)$$

$$b_s^{(q)}(j) = \frac{\sum_{N(s)} E_{s0}^{(q)} o_s(j)}{\sum_{N(s)} E_{s0}^{(q)}} \quad (16)$$

where $N(s)$ is a square neighborhood of s.

In the regions which do not contain text pixels, such a local estimation may generate artifacts. To prevent from this effect, we detect the case where no text is present by applying the following rule:

$$\|t_s^{(q)} - b_s^{(q)}\|^2 < \frac{\|\text{diag}\Sigma_t + \text{diag}\Sigma_b\|}{16\sqrt{2}} \quad (17)$$

Where diagΣ is the vector composed with the diagonal elements of matrix Σ. When the previous condition is met, the mixture model is reduced to a single mode (the background).

θ constitutes the global parameter of our model and is hence classically estimated using the whole image. Each (i, j) component of these two covariance matrices is given by:

$$\Sigma_t^{(q)}(i,j) = \frac{\sum_s E_{s1}^{(q)} \left(o_s(i) - t_s(i)^{(q)}\right)\left(o_s(j) - t_s(j)^{(q)}\right)}{\sum_s E_{s1}^{(q)}} \quad (18)$$

$$\Sigma_b^{(q)}(i,j) = \frac{\sum_s E_{s0}^{(q)} \left(o_s(i) - b_s(i)^{(q)}\right)\left(o_s(j) - b_s(j)^{(q)}\right)}{\sum_s E_{s0}^{(q)}} \quad (19)$$

The initialization of parameter Ψ is achieved by using the classical K-Means.

5 Experimental results

In order to assess the proposed algorithm, we have achieved several tests on handwritten and typographic documents. We have compared the performance of our algorithm with three well-known binarization techniques: Sauvola et al. adaptive thresholding [13], Wolf's MRF model [14] and Gatos's local thresholding [5].

As stated above, we have considered two kind of test: in the first case, the prior model is computed by using only the heuristic rules (for manuscript documents), and in the second case, the prior is obtained by combining the learning parameters with the heuristic rules. This solution is applied in the case of typographic documents.

5.1 Manuscript documents

To illustrate the performance of the approach, we have used different manuscript documents. Based on visual criteria, the proposed binarization gives the best results on heavily degraded documents. Our solution is less sensitive to background noise while being more sensitive to the text details. In order to properly assess the performance of the different algorithms, we have chosen to show details of the binarization of three manuscripts (fig.2).

Notice how the proposed algorithm can cope with low contrast and non uniform background to get the true contour of the text.This is especially visible in the first sample of fig. 2 where both center of "**e**" are open and thin characters like "**p**" are connected. The second sample shows the good behavior of the method when gray values of text are high.

5.2 Typographic documents

The next experiment considers the case of the binarization using a learning stage applied on typographic documents. We used 2 synthetic documents of approximately

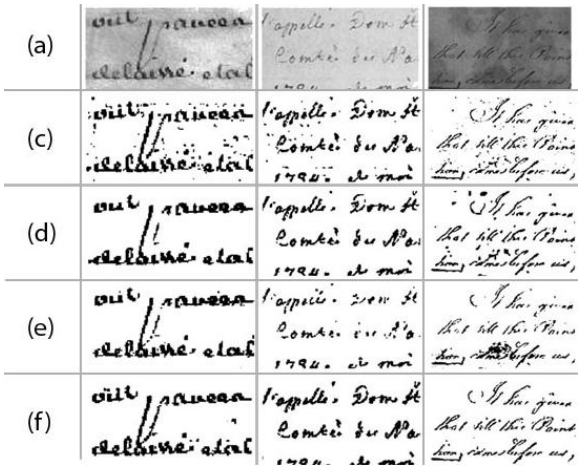

Figure 2. Binarization of old documents. (a) Original image, (b) Sauvola et al. method [13], (c) Wolf et al. method [14], (d) Gatos et al. method [5], (e) Proposed method

1000 words (produced with LaTeX) as training images. In order to perform experiments on documents with the same damage than the old manuscripts, we consider different typographic documents artificially degraded: we down sampled the gray scale input images, added some noise, and drew shadows and light.

Following Gatos *et al.*, we have evaluated the performances of the different algorithms by comparing the results obtained with the well-known OCR engine ABBYY FineReader 9.0 [1]. In figure 3 we give an example of results obtained with the five algorithms (FineReader, Gatos, Sauvola, Wolf and the proposed method) on a sample of the first document. In the case of this example, our algorithm clearly outperforms the other methods.

The results obtained on four documents are compared by using the Levenshtein distance between the correct text (ground truth) and the resulting text(tab. 1).

Original image	doc 1	doc 2	doc 3	doc 4
Abbyy FineReader	1679	649	197	3908
Gatos et al. [5]	1117	803	206	2412
Sauvola et al. [13]	611	125	69	80
Wolf et al. [14]	685	79	73	555
Our proposition	464	45	57	42

Table 1. Levenshtein distance from the ground truth

These results show bad behavior of Sauvola's and Wolf's algorithms when documents have high range of illumination. Indeed, due to the global parameter k, these algorithms

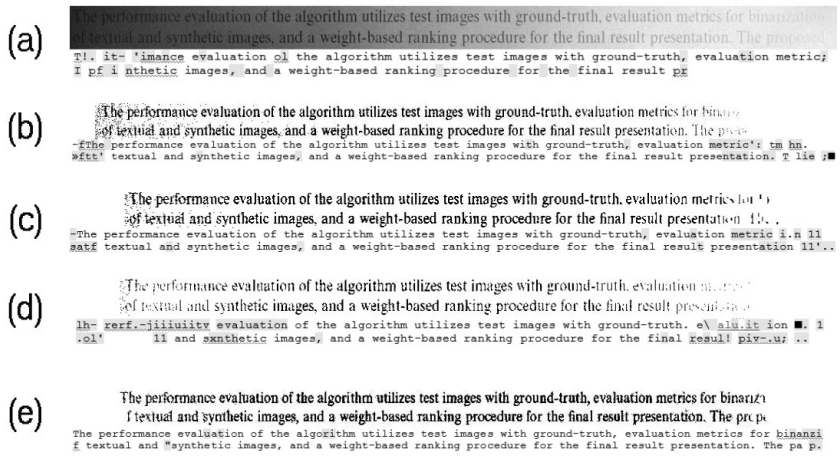

Figure 3. Example document image portion and its corresponding OCR results.(a) ABBYY's method, (b) Sauvola et al. method [13], (c) Wolf et al. method [14], (d) Gatos et al. method [5], (e) Proposed method

cannot deal with low contrast and simultaneously dark regions and bright regions on the document. The global threshold of the Otsu's method can explain the poor results of the character recognition. The Gatos's method which is based on the Sauvola's algorithm also presents poor results.

6 Conclusion

In this paper, we introduced a new technique based on a Markov Random Field (MRF) model. The improvement made by our solution is twofold: the likelihood model is more effective to detect text than other local thresholding methods, and the Markov Random Field adds space information to coerce the shape of the letters. Indeed, thanks to the prior model, the binarization is refined during the minimization. We tested the algorithm on many images, and the results show an especially good adaptation into different defect types such as illumination, noise and down-sampling. After extensive experiments, our method shows a greater sensitivity to thin line than other techniques, while being less sensitive to noise on numerous degraded document images using visual criteria as well as quantitative measures.

References

[1] F. 9.0. Abbyy finereader 9.0. `http://www.abbyy.com/`, 2008.

[2] J. Bernsen. Dynamic thresholding of grey-level images. In *Proc. Eighth Int 'l Conf. on Pattern Recognition*, pages 1251–1255, 1986.

[3] J. Besag. On the statistical analysis of dirty pictures. *Royal Statistical Society, Series B (Methodological)*, 48:259–302, 1986.

[4] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Royal Statistical Society, Series B (Methodological)*, 39:1–38, 1977.

[5] B. Gatos, I. Pratikakis, and S. J. Perantonis. Adaptive degraded document image binarization. *Pattern Recognition*, 39(3):317–327, 2006.

[6] 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):721741, 1984.

[7] H. Kim. Efficient automatic text location method and content-based indexing and structuring of video database. *Journal of Visual Communication and Image Representation*, 7(4):336–344, 1996.

[8] J. G. Kuk, N. I. Cho, and K. M. Lee. Map-mrf approach for binarization of degraded document image. *ICIP08*, pages 2612–2615, 2008.

[9] S. Messelodi and C. Modena. Automatic identification and skew estimation of text lines in real scene images. *Pattern Recognition*, 32(5):791–810, 1999.

[10] D. Milun and D. Sher. Improving sampled probability distributions for markov random fields. *Pattern Recognition Letters*, 14(10):781788, 1993.

[11] W. Niblack. *An introduction to digital image processing*. Prentice Hall (July 1986), 1986.

[12] W. Qian and D. M. Titterington. Estimation of parameters in hidden markov models. *Phil. Trans. R. Soc. Lond. A*, 337:407–428, 1991.

[13] J. Sauvola and M. Pietikinen. Adaptive document image binarization. *Pattern Recognition*, 33(2):225 – 236, 2000.

[14] C. Wolf and D. Doermann. Binarization of low quality text using a markov random field model. *Proceedings of the 16 th International Conference on Pattern Recognition (ICPR'02)*, 3:30160, 2002.