Binarization of Low Quality Text using a Markov Random Field Model

Christian Wolf[1]

[1] Laboratoire Reconnaissance de Formes et Vision
INSA de Lyon, Bât. J.Verne 20, Av. Albert Einstein
Villeurbanne, 69621 cedex, France
wolf@rfv.insa-lyon.fr

David Doermann[2]

[2] Laboratory for Language and Media Processing
Institute for Advanced Computer Studies
Univ. of Maryland, College Park, MD 20742-3275
doermann@umiacs.umd.edu

Abstract

Binarization techniques have been developed in the document analysis community for over 30 years and many algorithms have been used successfully. On the other hand, document analysis tasks are more and more frequently being applied to multimedia documents such as video sequences. Due to low resolution and lossy compression, the binarization of text included in the frames is a non trivial task. Existing techniques work without a model of the spatial relationships in the image, which makes them less powerful. We introduce a new technique based on a Markov Random Field (MRF) model of the document. The model parameters (clique potentials) are learned from training data and the binary image is estimated in a Bayesian framework. The performance is evaluated using commercial OCR software.

1. Introduction

MRFs and Gibbs Distributions have been used successfully for image restoration [2]. We believe, that MRF models are very well suited for modeling text properties, but the models found in the literature are not rich and do not exploit enough of the properties of text characters.

Thouin et al. use 3×1 and 1×3 cliques and energy functions stimulating horizontal and vertical homogeneous strokes to restore bimodal text images from low resolution samples [10]. Cui and Huang use 2×2 cliques with similar energy functions to extract binarized car license plate images from gray scale sequences [1]. Both approaches use a simple model, where the MRF is used to model the prior for an MAP estimator which tends to remove simple noise and augments small straight strokes.

Of course, given all the different possible types of text in different fonts, sizes, styles, orientations, colors etc., it is difficult or near impossible to create an exact model of text which fits all possible text observations. In part, this is the reason for the simplicity of the existing text models. However, the fonts of most of the low quality text in video broadcasts are very simple and without serifs so they remain readable when rendered at low resolution.

This paper is outlined as follows: Section 2 gives a short introduction into MRFs and Gibbs Distributions. Section 3 presents the observation model and Section 4 the prior model. Section 5 explains the annealing process used to minimize the energy function. In Section 6 we describe the experiments we performed to evaluate the system and give some results. Section 7 finishes with concluding remarks.

2. MRFs and Gibbs Distributions

MRF models treat images as stochastic processes. A field X of random variables X_{s_1}, X_{s_2}, ... X_{s_N} is a MRF iff $P(X{=}z) > 0 \quad \forall z \in \Omega$ and $P(X_s{=}x_s|X_r{=}x_r, r \neq s) = P(X_s{=}x_s|X_r{=}x_r, r \in g_s)$ where z is a configuration of the random field, Ω is the space of all possible configurations and g_s is the neighborhood of the pixel X_s. In other words, the conditional probability for a pixel of the image depends only on the pixels of a pre-defined neighborhood around this pixel. This neighborhood is defined in terms of cliques, where a clique is the set of pixels which are neighbors of the given pixel. It has been proven that the joint probability density functions of MRFs are equivalent to Gibbs distributions, i.e. are of the form

$$\pi(z) = \frac{1}{Z} e^{-U(z)/T}$$

where Z is a normalization constant, T is a temperature factor which can be assumed to be 1 for simplicity, $U(z) = \sum_{c \in C} V_c(z)$ is an energy function, C is the set of all possible cliques of the field and V_c is the energy potential defined on a single clique. Hence, the model of the spatial relationship in the image can be defined in terms of energy potentials V_c of the clique labelings.

In this paper the MRF models the prior $P(z)$ in a Bayesian maximum a posteriori (MAP) estimator which de-

termines the binary estimate from a degraded observation:

$$\hat{z} = \arg\max_z P(z|f) = \arg\max_z P(z)P(f|z) \quad (1)$$

The likelihood $P(f|z)$ depends on the observation and the noise process. The prior and likelihood models are described in the next two sections.

3. The observation model

The likelihood or conditional density is the probability of the observation given an estimate. It depends on the observation model assumed for the process. Most approaches use simple models, e.g. Gaussian noise with zero mean and variance σ_n^2 and text and background gray values of 0 and 255 respectively, which leads to the probability density function (p.d.f.)

$$P(f|z) = (2\pi\sigma_n^2)^{-N/2} \exp\left\{\frac{||z-f||^2}{2\sigma_n^2}\right\}$$

where f is the observed gray scale image and z is the estimated binary image. However, in real life this is rarely the case. Modeling the image degradation this way results in fixed thresholding with threshold 127.5 if the prior distribution is set to uniform — a method which is not acceptable.

Given a reliable estimate of the text gray value(s), background gray value(s) and the variance of the noise process, a model assuming Gaussian noise results in the p.d.f.

$$P(f|z) = (2\pi\sigma_n^2)^{-N/2} \exp\left\{\frac{||c(z)-f||^2}{2\sigma_n^2}\right\} \quad (2)$$

where $c(z)$ is the function delivering the text gray value for text pixels and background gray value for background pixels. Unfortunately, estimating the text and background color (or gray value) is a difficult task and far from trivial. We performed experiments with an MRF based binarization scheme using the observation model (2) together with a uniform prior. Due to inaccurate estimates, the results are not as good as the ones obtained by existing binarization techniques such as Sauvola et al's method [9]. The goal of our approach is to improve the classic techniques by using spatial relationships defined in the prior p.d.f. Therefore, as a desired property, we want the new technique to give at least the same results as these existing techniques, if the prior distribution is set to uniform.

Niblack's binarization method [5] has been used successfully and has performed as one of the best in a well known survey [7]. Niblack's algorithm calculates a threshold surface by gliding a rectangular window across the image. The threshold T for the center pixel of the window is computed using the mean m and the variance s of the gray values in the window: $T = m + k \cdot s$, where k is a constant set to -0.2. The results are not very sensitive to the window size as long as the window covers at least 1-2 characters. However, the drawback is noise created in areas which do not contain any text (due to the fact that a threshold is created in these cases as well). Sauvola et al. [9] solve this problem by adding a hypothesis on the gray values of text and background pixels, which results in the following formula for the threshold:

$$T = m \cdot (1 - k \cdot (1 - \frac{s}{R}))$$

where R is the dynamics of the standard deviation fixed to 128 and the parameter k is set to 0.5. We incorporate this method into the observation model for the new technique as follows:

$$P(f|z) = (2\pi\sigma_n^2)^{-N/2} \exp\left\{\frac{||z-f+T-127.5||^2}{2\sigma_n^2}\right\} \quad (3)$$

Fixed thresholding is replaced by thresholding using the threshold surface computed by Sauvola's algorithm. The last parameter to estimate is σ_n^2, the variance of the noise process. If the prior is uniform, then a change of the variance does not effect the result. However, if a non uniform prior is taken into account, then the noise variance should be estimated as closely as possible since it controls the weight between the prior model and the observation model.

We use local statistics of the gray levels of the image to estimate the noise variance. A window is shifted across the image and the gray level histogram of the pixels in the window is computed. Then the histogram is separated into two classes by calculating Otsu's optimal threshold [6] (which is equivalent to maximizing the intraclass variance). Finally, the standard deviation of the noise is estimated as $\sigma_n = w\ \sigma_{ic}$, where σ_{ic} is the intraclass standard deviation and w is a weighting factor we fixed to 0.5.

4. The prior distribution

Our prior model is defined on a large neighborhood (4×4 pixel cliques), which makes it more powerful. On the other hand, simple tabularization of the clique potentials is not possible anymore, since we would need to specify the energy potentials for a large number of clique labelings ($2^B = 65535$, where B=16 is the number of pixels in the clique). We decided on a different approach, learning the clique potentials from training data. The absolute probability of a clique labeling θ_i can be estimated from the frequency of its occurrence in the training images. The probability can be converted into a clique potential as follows:

$$V_c(\theta_i) = -\frac{1}{B} ln(P(\theta_i))$$

Not all of the theoretically possible clique labelings are found in the training images, so the question arises how to

find the potentials for the missing cliques. One solution has been proposed by Milun and Sher [4] as an application of the work of Hancock and Kittler [3]. The probability distribution of the clique labelings $\theta_i \in \Theta$ is smoothed using the following function:

$$P'(\theta_i) = \sum_{\theta_j} P(\theta_j) p^{d(i,j)} (1-p)^{u(i,j)}$$

where $d(i,j)$ is the number of bits which differ between the clique labelings θ_i and θ_j, and $u(i,j)$ are the number of bits which are the same. p is the smoothing coefficient, higher values denote more smoothing.

A typical training document image contains far more background pixels than text pixels (a typical value is about 1% of text pixels). This leads to very low energy potentials for the cliques containing many background pixels. We therefore normalized the clique potentials by a term which only depends on the histogram of the clique. Let θ_a be a clique labeling which contains z text pixels and n background pixels. In an image randomly generated from a stationary but biased source, which generates text pixels with probability $P(X{=}0) = \alpha$ and background pixels with probability $P(X{=}255) = \beta$, the probability to find θ_a on a location is $P_i(\theta_a) = \alpha^z \beta^n$. Dividing the measured clique probability by this factor we normalize the energy distribution by increasing the energy of cliques with many background pixels and decreasing the energy of cliques with many text pixels (assuming that $\alpha < \beta$, as it is the case for document images). The final energy potential is calculated as follows:

$$V_c(\theta_i) = -\frac{1}{B} ln(\frac{1}{Z} \frac{P'(\theta_i)}{P_i(\theta_i)})$$

where $Z = \sum_{\theta_j} \frac{P'(\theta_j)}{P_i(\theta_j)}$ is a normalization constant.

5. Optimization

To estimate the binary image, equation (1) 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 [2].

During the annealing process, for each pixel the energy potential is calculated before and after flipping it. The decision whether to flip the pixel or not is based on the value $q = e^{-\Delta/T}$, where Δ is the difference of the energy potentials before and after the pixel change. If $q > 1$ then the change is favorable and accepted. If $q < 1$ then it is accepted with probability q. This prevents the algorithm from staying in a local minimum by sometimes performing changes which increase the energy function. The probability q depends on a global "temperature" factor T, which is slowly lowered during the iterative process making unfavorable changes less and less likely. For the cooling schedule we used the suggestions in [8] (page 356), where the temperature T is set to $T(k) = T(1) \cdot c^{k-1}$ where c is a constant controlling the speed of the cooling process and k denotes the current iteration. The start temperature must be sufficiently high to switch to energetically very unfavorable states with a certain probability. We set $T(1) = 2\max(\Delta)$ which allows a change from the clique with lowest temperature (i.e. very favorable) to the clique with the highest temperature with probability 0.61.

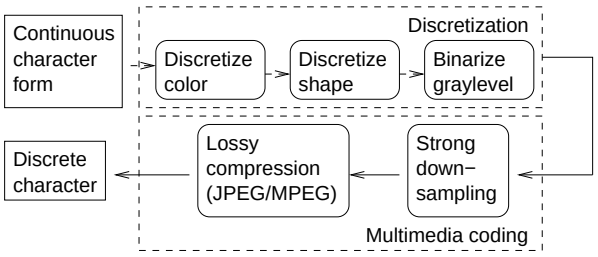

Figure 1. The degradation model

6. Experimental Results

To evaluate the performance of our algorithm we binarized images and passed them to the commercial OCR program Finereader 5.0. The documents were taken from the Pink Panther database [12] and the University of Washington document image collection [11]. We used 13 synthetic documents (produced with LaTeX) as training images and 5 grayscale documents as test documents.

To be able to perform experiments in realistic conditions we degraded the images. Several theoretical document image degradation models are known in the document image processing community. However, since we concentrated on low quality text (e.g. taken from multimedia documents), we ignored the degradation following from the discretization of the continuous characters, because these effects are normally overshadowed by the stronger degradations following from the multimedia coding (see Figure 1). To simulate these degradations we down sampled the gray scale input images by a factor of two and compressed them with the JPEG algorithm with a quality factor of 75%.

The maximum difference Δ we calculated from the clique energy function was 2.5 so we set the start temperature of the annealing algorithm to $T(1){=}2\Delta{=}5$. About 10% of all possible clique labelings were found in the training set, the probability for the others was set to 0 if Hancock/Kittler (HK) smoothing was applied and to $\frac{1}{2}\min(P(\theta_i))$ otherwise. We ran the annealing procedure with 400 iterations and a cooling factor of $c{=}0.97$.

Figures 2a and 2b illustrate an example of a character repaired using the spatial information in the prior. Figure

(a)  (b)

θ_b	θ_a	$V_c(\theta_b)$	$V_c(\theta_a)$	θ_b	θ_a	$V_c(\theta_b)$	$V_c(\theta_a)$
		1.05	0.95			1.82	1.38
		1.48	1.15			1.85	1.30
		2.00	1.36			2.14	1.40
		1.80	1.79			1.77	1.52
		1.87	1.16			1.84	1.57
		1.72	1.32			1.66	1.42
		2.00	1.28			2.08	1.57
		1.89	1.50			1.93	1.69
					$\sum$	28.91	22.38

(c)

Figure 2. Example binarized with Sauvola's method (a) with the MRF method (b) The clique labelings of the repaired pixel(c)

2c shows the clique labelings of the repaired pixel before (θ_b) and after (θ_a) the pixel has been flipped. Note, that all 16 cliques favor the change of the pixel.

We performed experiments with HK smoothing parameters p=0 (no smoothing) and p=0.001, and with or without normalization of the clique probabilities by the factor $P_i(\theta_i)$. We could not confirm the results of Milun and Sher, which reported improvements of the results if HK smoothing was applied. On the other hand, application of the normalization step improved the quality of the results.

The OCR experiments have been conducted with two classes of images: Documents binarized with Sauvola's method and documents binarized with the MRF method (normalization of the clique potentials, no HK smoothing). As can be seen from Figure 3, the MRF prior is capable of repairing some damage in the characters. The recognition rates are 79.0% for Sauvola's method and 82.0% for the MRF method. Table 1 gives the results for each document.

7. Conclusion and Outlook

In this paper we presented a binarization technique based on a probabilistic MRF model. Although the model is very powerful, the results of the method do not improve existing techniques significantly. This is partly due to the sensitivity of the method to model parameters (the noise variance) and to the energy function. We have shown that learning the clique parameters from training data does not result in an energy function invariant to changes in the text properties. We nevertheless believe that given the nature of existing binarization techniques further research in probabilistic

(a) (b)

Figure 3. An example binarized with Sauvola's method (a) with the MRF method (b)

Document	1	2	3	4	5	Total
Sauvola	77.1	39.8	77.1	99.0	98.7	79.0
MRF	81.0	40.5	87.3	99.3	98.8	82.0

Table 1. The OCR results

modeling of text is necessary in order to create a deeper understanding of the properties of text. Our future work will explore new ways to define the energy potentials of the model.

References

[1] Y. Cui and Q. Huang. Character Extraction of License Plates from Video. In *Proceedings of the 1997 IEEE conf. on Computer Vision and Pattern Recognition*, pages 502–507, 1997.

[2] 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.

[3] E. Hancock and J. Kittler. A Label Error Process for Discrete Relaxation. In I. C. Society, editor, *Proc. of the ICPR*, volume 1, pages 523–528, 1990.

[4] D. Milun and D. Sher. Improving Sampled Probabilty Distributions for Markov Random Fields. *Pattern Recognition Letters*, 14(10):781–788, 10 1993.

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

[6] N.Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man and Cybernetics*, 9(1):62–66, 1979.

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

[8] R. Duda and P. Hart and D. Stork. *Pattern Classification, 2nd Edition*. Wiley, New York, NY, Nov. 2000.

[9] 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.

[10] P. Thouin, Y. Du, and C.-I. Chang. Low Resolution Expansion of Gray Scale Text Images using Gibbs- Markov Random Field Model. In *2001 Symp. on Document Image Understanding Techn., Columbia, MD*, pages 41–47, 4 2001.

[11] Univ. of Washington. Document image database collection. Intelligent Systems Laboratory, Dept. of Electrical Engineering. 352500 U. of Washington, Seattle, WA 98195.

[12] B. Yanikoglu and L. Vincent. Pink Panther: A complete environment for ground-truthing and benchmarking document page segmentation. *Pattern Recognition*, 31(20):1191–1204, 1998.