FAMILIES OF MARKOV MODELS FOR DOCUMENT IMAGE SEGMENTATION

Christian Wolf

Université de Lyon, CNRS
INSA-Lyon, LIRIS, UMR5205, F-69621, France
christian.wolf@liris.cnrs.fr

ABSTRACT

In this paper we compare several directed and undirected graphical models for different image segmentation problems in the domain of document image processing and analysis. We show that adapting the structure of the model to specific sitations at hand, for instance character restoration, recto/verso separation and segmenting high resolution character images, can significantly improve segmentation performance. We propose inference algorithms for the different models and we test them on different data sets (manuscripts and printed text of different qualities).

1. INTRODUCTION

The segmentation of printed or hand written document images is an important step for several applications, either as a post processing step for OCR, text/graphics separation, physical and logical page segmentation, or as a goal in itself, for instance in document restoration tasks. Bayesian estimation and probabilistic graphical models are a powerful tool which allow to model the statistical distributions of observed and hidden values and to take into account the statistical relationships between them. Although there is a wide variety in model families and types, in the domain of document segmentation the most frequently used models are flat Markov random fields (MRF) [1, 2, 3, 4, 5].

In this paper we make the case of adapting the model structure to the type of document at hand, to the imaging conditions, the specific application etc. We show that choosing the model structure according to very simple criteria, e.g. the scanning resolution, may significantly boost the classification results.

In the following we will present graphical models defined on directed or undirected graphs, where each node corresponds to a random variable which may be hidden or observed. Hidden variables are denoted as F_i and take values from an alphabet $\Lambda = \{1, 2, \dots C\}$. Observed variables are denoted as D_i, and are related to the grayvalue or color information of a pixel or a group of pixels. Depending on the type of the model, the indices i are related to spatial position, scale etc.

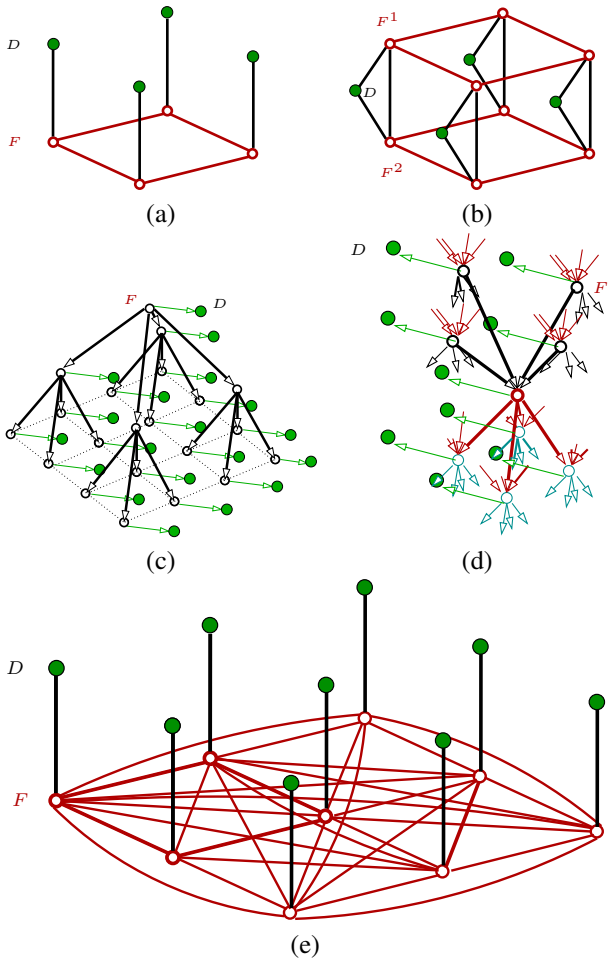


Fig. 1. Different graphical models: flat MRF (a), double layer MRF (b), quad tree (c) cube (d) flat MRF with large 3×3 cliques (e)

In the classical case of a flat MRF defined on the image pixels, the observed variables correpond to grayvalues or colors, and the hidden variables correspond to classification results for a single pixel. Figure 1a shows the dependency graph for 4-pixel MRF with second order cliques on horizontal and vertical neighborhood relationships: ob-

served nodes are shown as shaded, hidden nodes are empty. The maximum *a posteriori* (MAP) estimator will return realizations f_i of the hidden variables F_i maximizing the joint probability $p(f, d)$ of the hidden and observed variables. The prior probability $p(f)$ is defined through the energy potentials of the cliques involving hidden labels.

The paper is outlined as follows: sections 2, 3 and 4 present types of graphical structures which extend the possibilities of MRFs in order to more precisely model specific situations or applications, namely recto/verso separation, scale behavior, and character restoration, respectively. Section 5 deals with inference algorithms for the presented models and section 6 compares the models in terms of restoration power. Section 7 finally concludes.

2. SEPARATELY MODELING RECTO AND VERSO

In segmentation and restoration applications, the Potts model is frequently used to favor the segmentation of homogeneous regions. It defines the energy potentials of cliques involving hidden labels only as follows:

$$U_{Potts}(f) = \sum_{\{s\}\in\mathcal{C}_1} \alpha f_s + \sum_{\{s,s'\}\in\mathcal{C}_2} \beta_{s,s'}\delta_{f_s,f_{s'}} \qquad (1)$$

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 model is parametrized through α and β_x, where x denotes a direction index (horizontal, vertical).

In the case of bleedthrough removal, where information from the verso side is showing through and must removed, each hidden label may take values in (*recto*, *verso* and *background*). In this case, the hypothesis of homogeneous regions is not justified anymore, since the unknown and undegraded source information is composed of two sources images, and *a priori* knowledge may be available for each of the source images, but not for the mixture of these images.

The situation can be modeled with two different hidden fields, F^1 and F^2, connected through a single observed field D, and where the hidden labels may take two different values (*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.

The dependency graph of the model is shown in Figure 1b. It is composed of two hidden fields with pairwise cliques connected through cliques containing two hidden variables and an observed variable. Assuming that the 2-node cliques involving pairs $\in F^1 \times F^2$ have zero potential, a property which can be derived from the application, we can easily see from the way the cliques connect F^1, F^2 and D, that the following holds (see also [6]):

$$U(f^1, f^2, d) = U_{Potts}(f^1) + U_{Potts}(f^2) + U_{Obs}(f^1, f^2, d) \qquad (2)$$

where U_{Obs} is an observation model which factorizes over pixels, i.e. which decomposes into cliques containing hidden and observed variables related to same pixel only.

Transforming the clique energy potentials back to probabilities using the Hammersley-Clifford theorem [7], we can see that the prior probability $P(f^1, f^2)$ is actually the product of the two probabilities of the two fields $P(f^1)$ and $P(f^2)$. 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 gray value is rather low for a white document with dark text), then the knowledge that the recto label is background will increase the probability that the verso pixel will be text.

Assuming 100% opaque ink, Gaussian color variations and eventual Gaussian noise of the scanning device (see also [6]) as well as the usual conditional independence assumption, the expression $U_{Obs}(f^1, f^2, d)$ of the observation model factorizes as follows:

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

The parameters of the Gaussian distributions (means and covariances) depend on the labels of each site s, resulting in three different distributions — the distributions for $f_s^1{=}1, f_s^2{=}0$ and $f_s^1{=}1, f_s^2{=}1$ are the same.

3. MODELING SCALE BEHAVIOR

Hierarchical models introduce a scale dependent component into the segmentation algorithm, which allows the algorithm to better adapt itself to the image characteristics. The nodes of the graph are partitioned into different scales, where lower scale levels correspond to finer versions of the image and higher scale levels correspond to coarser versions of the image. Examples are stacks of flat MRFs [8], pyramidal graph structures [9] and the scale causal multi-grid

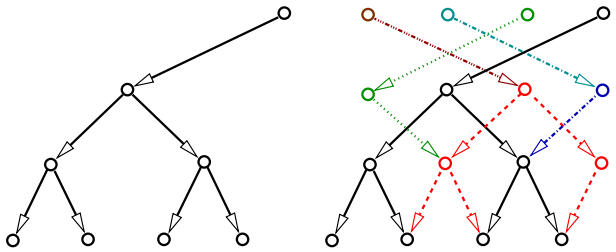


Fig. 2. A one dimensional representation of the extension of the quad tree — shown as a dyadic tree (a) — to the cube (b). In both figures the observed nodes have been omitted.

[10]. The computational complexity issues encountered by these models have been tackled by Bouman and Shapiro who where the first to propose a causal hierarchical model for image segmentation [11] (refined by Laferte *et al.* [12]). A quad tree models the spatial interactions between the leaf pixel sites through their interactions with neighbors in scale, as shown in figure 1c. The main problem of the quad tree structure is the non stationarity it induces into the random process of the leaf sites, which results in "blocky" artifacts in the segmented image — a serious problem in document image processing.

We proposed an extension of the quad tree to a cube structure [13] shown in figure 2, where for easier representation the one dimensional case — a dyadic tree — is shown. Figure 2a shows the standard tree. First, a second dyadic tree is added to the graph, which adds a common parent to all neighboring leaf sites which did not yet share a common parent. In the full two dimensional case, three new quad trees are added. The problem is solved for the first level, where the number of parents increased to 4 (for the full 2D model). We repeat the process for each level. New trees connect sites of the original quad tree, but also sites of the trees added at the lower levels. The final result can be seen in figure 2b. Note, that the final graph is not a pyramid anymore, since each level contains the same number of nodes. In general, each node has 4 parents (2 in the 1D representation) except border nodes.

The resulting full Markov cube including observed nodes is a belief network, a small example part of the 2D case is shown in Figure 1d. The model is parametrized through three probability distributions: the discrete prior distribution of the top level labels $p(f)$, the transition probabilities $p(f_s|f_{s^-})$, where the sites s^- are the parents of site s, and the likelihood of the observed nodes given the corresponding hidden nodes $p(d_s|f_s)$. For the inference algorithm, observations at different cube levels are needed. The only available observations are at the base level, but the higher levels can be calculated recursively, e.g. through a mean filter.

4. MODELING CHARACTER SHAPES

In some applications the document images are of very low resolution, for instance when we deal with screen text extracted from video sequences or captured with cell phones. In this case we encounter characters having widths and heights of a few pixels only, therefore the classical goal of favoring homogeneous regions is not applicable anymore. On the other hand, it might be desirable to learn the character shapes in order to restore degraded characters [14].

This can be achieved by defining a prior model on binary labels and a very large neighborhood, for instance 4×4 pixel cliques. A dependency graph of a smaller version on 3×3 pixel cliques is shown in Figure 1e. Characterizing the shapes of the different characters rules out simple parametric potentials as the Potts model (1), and simple tabularization of the clique potentials is not possible, since we would need to specify the energy potentials for a large number of clique labellings ($2^B = 65535$, where B=16 is the clique size). A different approach is to learn the clique potentials from training data. The absolute probability of a clique labeling θ_i composed of B binary values can be estimated from the frequency of its occurrence in the training images. The probability can be converted into a clique potential as follows:

$$U(\theta_i) = -\frac{1}{B} ln(P(\theta_i)) \tag{4}$$

Not all of the theoretically possible clique labellings are found in the training images, so the question arises how to find the potentials for the missing cliques. One solution is Hancock-Kittler smoothing proposed by Milun and Sher [15]. The probability distribution of the clique labelings is smoothed using the following function:

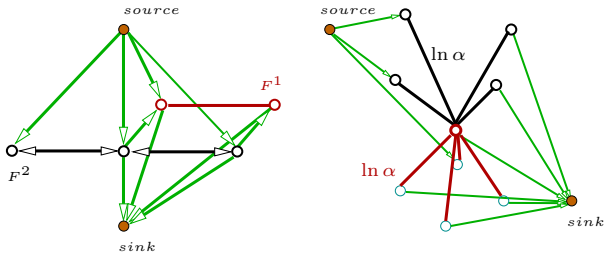
$$P'(\theta_i) = \sum_{\theta_j} P(\theta_j) p^{H(i,j)} (1-p)^{B-H(i,j)}$$

where $H(i,j)$ is the Hamming distance between θ_i and θ_j. p is the smoothing coefficient, higher values denote more smoothing.

5. INFERENCE

A part from the quad tree, the discussed dependency graphs contain cycles (not taking into account the edge direction in the case of directed graphs). In the general case, minimizing the corresponding energy functions is therefore NP-hard [16]. The functions are generally not convex so standard gradient descent methods will most likely return a bad local minimum. Simulated Annealing has been proven to return the global optimum under certain conditions [17] but is painfully slow in practice.

For some specific classes energy functions, minimization with graph cuts is a very fast technique to get the global



(a) (b)

Fig. 3. Cut graphs used for energy minimization of two different models: double layer MRF (a) Markov cube (b)

minimum, or at least a very good local minimum [16]. A very easy case concerns binary labels and pairwise potentials U_2 satisfying the following submodularity constraint:

$$U_2(0,0) + U_2(1,1) \leq U_2(0,1) + U_2(1,0) \qquad (5)$$

Applied to the double MRF model (section 2), it can easily be seen that this is the case of the terms corresponding to the Potts model (1), but not necessarily for all terms corresponding to the observation model. According to the value of the observation d_s at site s, the corresponding term may be submodular or not.

The problem can be solved approximately by iteratively fixing the hidden labels of one of the two fields F^1 and F^2 and estimating the labels of the other one. More precisely, not all sites of one of the fields are fixed, but only the variables whose sites s are not regular. The variables F^1_s and F^2_s for the regular sites s are estimated jointly. Figure 3a shows an example cut graph constructed for a one line three pixel image with two submodular sites [6].

The best choice of inference technique for the Markov cube (section 3) may depend on the form of the function parameterizing the transition probabilities $p(f_s|f_{s^-})$. In the general case, loopy belief propagation gives an approximative solution [13]. However, for a large sub class with interesting properties, excellent graph cut solutions can be found, for instance for a regularizing term based on the number of parent labels which are equal to the child label:

$$p(f_s|f_{s^-}) = \frac{1}{Z}\alpha_l^{\xi(f_s,f_{s^-})} \qquad (6)$$

where α_l is a parameter depending on the level l, $\xi(f_s, f_{s^-})$ is the number of labels in f_{s^-} equal to f_s and Z is a normalization constant. The so defined transition probabilities favor homogeneous regions, which corresponds to the objective of an image segmentation algorithm. We can decompose this expression into a sum of binary terms:

$$\ln p(f_s|f_{s^-}) = \sum_{s' \in s^-} [(\ln \alpha)\, \delta_{f_s,f_{s'}}] - Z \qquad (7)$$

(a) (b)

Fig. 5. A small part of the segmentation results obtained on high resolution images: (a) MRF (b) Markovcube.

where $\delta_{a,b}$ is the Kronecker delta. Since each binary term is submodular, a global solution can be obtained using graph cuts. Figure 3b shows an example cut graph constructed for the dependency graph of Figure 1d.

Since the large majority of the energy terms for the large clique MRF (section 4) are in general not submodular, graph cut optimization is not applicable in this case. Similarly, loopy belief propagation will be too costly since it is exponential in clique size. A standard technique in these situations is simulated annealing [17].

6. EXPERIMENTAL RESULTS

To evaluate the different models and inference algorithms we tested them on real applications. The double layer MRF and the Markovcube have been tested on medium and high resolution images of low quality printed manuscripts from the 18^{th} century. The objective was restoring the images degraded with ink bleedthrough, i.e. removing the verso component from the recto scan, which makes it a three class segmentation problem. The medium resolution dataset contained 104 images and the high resolution dataset contained 9 images. The difficulties of the two datasets in terms of image contents were different, the results are therefore not comparable across datasets. We tested the methods' abilities to improve the performance of an OCR algorithm and compared them to several widely cited algorithms: k-means clustering, a flat Markov random field (MRF) with graph cuts optimization [16], as well as two well known methods[1] based on source separation [18, 19].

Figure 4 shows parts of the restoration results on medium resolution images together with OCR results. The restored images are obtained by replacing all pixels classified as *verso* by the means of the grayvalues of the surrounding pixels classified as *background*. Figure 5 shows parts of the segmentation results on high resolution images.

We manually created groundtruth and calculated the recall and precision measures on character level, which are

[1]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.

Fig. 4. Restoration and OCR results obtained on medium res. images. SP denotes the source plane for ICA based methods.

	104 med. res. images		9 high res. images		5 low res. images
Method	Recall	Precision	Recall	Precision	Accuracy
No restoration	65.65	49.91	$\ll$	$\ll$	–
Niblack [21] (segm. only)	$\ll$	$\ll$	–	–	–
Sauvola et al. [20] (segm. only)	78.75	66.78	–	–	79.0
K-Means (k=3)	78.57	69.43	61.23	51.74	–
Tonazzini et al. [19]	‡ 41.00	‡ 30.05	–	–	–
Tonazzini et al. [18]	$\ll$	$\ll$	–	–	–
Tonazzini et al. [18] - all 3 sources	‡ 50.52	‡ 33.90	13.13	25.43	–
MRF - Potts & α-exp. move [16]	81.99	72.12	69.10	58.42	–
MRF 4×4, simulated annealing	–	–	–	–	82.0
Double layer MRF	83.23	74.85	75.76	68.08	–
Markovcube & graph cut	–	–	69.34	61.19	–

$\ll$ Low quality result : correct evaluation impossible
‡ Low quality result: evaluation of a subset of the images only

Table 1. Evaluation of OCR improvement by different restoration methods when applied to scanned document images. Three different datasets have been used with different difficulties. The results are not comparable across datasets.

given in table 1. Surprisingly, the recognition performance on the results of the two source separation results was very disappointing. Unfortunately, the recognition performance on these results was not good enough for the whole datasets, we give results only on partial datasets only.

The methods based on graphical models outperform all other methods. Not surprisingly, the double layer MRF performs best since it has been designed for recto/verso separation. However, we can also note that the (single layer) Markovcube outperforms the single layer MRF on the high resolution images. On image segmentation problems not related to recto/verso separation the Markovcube is therefore able to improve the performance of flat models, which can be explained by its ability to take into account the image characteristics at multiple scales. A double layer Markov cube could further boost performance.

The large clique MRF has been tested on a binarization task and low resolution images, a configuration which it has been designed for. We compare two different techniques: adaptive document binarization by Sauvola et al. [20] as well as the large clique MRF with a Gaussian observation model corrected by Sauvola et al.'s algorithm [14]. As table 1 shows, the learned clique potentials significantly improve OCR performance. Figure 6 shows a zoom of a single character of the dataset binarized with both methods.

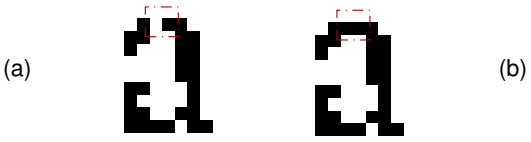


Fig. 6. Low resolution example binarized with Sauvola's method (a) with the MRF method (b)

7. CONCLUSION

We compared several different graphical models on document image segmentation tasks. As pointed out, the models perform best on the images they have been adapted for. Applied to the 3 class recto/verso separation problem, the double layer MRF performs best since it takes into account the specific situation of two independent sources images. The Markov cube is able to outperform the flat MRF model on high resolution images, although on the recto/verso problem it does not perform better than the double layer MRF. A double Markov cube might further improve results on high resolution recto/verso separation problems. The large clique MRF with non parametric learned energy potentials finally is able to restore characters in very low resolution images. Perspectives are extensions to discriminative models and/or to pairwise or triplet Markov models.

8. REFERENCES

[1] A. Tonazzini, S. Vezzosi, and L. Bedini, "Analysis and recognition of highly degraded printed characters," *I.J. on Doc. Anal. and Rec.*, vol. 6, no. 4, pp. 236–247, 2003.

[2] K. Donaldson and G.K. Myers, "Bayesian super-resolution of text in video with a text-specific bimodal prior," *I.J. on Doc. Anal. and Rec.*, vol. 7, no. 2-3, pp. 159–167, 2005.

[3] 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*, 4 2001, pp. 41–47.

[4] Y. Cui and Q. Huang, "Character Extraction of License Plates from Video," in *Proc. of the Conf. on Comp. Vision and Pattern Rec.*, 1997, pp. 502–507.

[5] A. Tonazzini, L. Bedini, and E. Salerno, "A markov model for blind image separation by a mean-field em algorithm," *IEEE Tr. on Image Proc.*, vol. 15, no. 2, pp. 473–482, 2006.

[6] C. Wolf, "Document ink bleed-through removal with two hidden markov random fields and a single observation field," in *IEEE Tr. on PAMI*, (to appear).

[7] J. Besag, "Spatial interaction and the statistical analysis of lattice systems," *J. of the Roy. Stat. Soc., Series B*, vol. 36, no. 2, pp. 192–236, 1974.

[8] M.G. Bello, "A combined Markov random field and wave-packet transform-based approach for image segmentation," *IEEE tr. on image proc.*, vol. 3, no. 6, pp. 834–846, 1994.

[9] Z. Kato, M. Berthod, and J. Zerubia, "A hierarchical Markov random field model and multitemperature annealing for parallel image classification," *Graphical Models and Image Proc.*, vol. 58, no. 1, pp. 18–37, 1996.

[10] M. Mignotte, C. Collet, P. Perez, and P. Bouthemy, "Sonar image segmentation using an unsupervised hierarchical mrf model," *IEEE Tr. on Image Proc.*, vol. 9, no. 7, pp. 1216–1231, 2000.

[11] C.A. Bouman and M. Shapiro, "A Multiscale Random Field Model for Bayesian Image Segmentation," vol. 3, no. 2, pp. 162–177, 3 1994.

[12] J.-M. Laferte, P. Perez, and F. Heitz, "Discrete Markov image modelling and inference on the quad tree," *IEEE Tr. on Image Proc.*, vol. 9, no. 3, pp. 390–404, 2000.

[13] C. Wolf and G. Gavin, "Inference and parameter estimation on hierarchical belief networks for image segmentation," Tech. Rep. RR-LIRIS-2008-21, LIRIS Laboratory, 2008, Minor revision at "Neurocomputing".

[14] 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*, 2002, vol. 3, pp. 160–163.

[15] D. Milun and D. Sher, "Improving Sampled Probabilty Distributions for Markov Random Fields," vol. 14, no. 10, pp. 781–788, 1993.

[16] V. Kolmogorov and R. Zabih, "What energy functions can be minimized via graph cuts?," *IEEE Tr. on PAMI*, vol. 26, no. 2, pp. 147–159, 2004.

[17] S. Geman and D. Geman, "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images," vol. 6, no. 6, pp. 721–741, 11 1984.

[18] A. Tonazzini and L. Bedini, "Independent component analysis for document restoration," *I.J. on Doc. Anal. and Rec.*, vol. 7, no. 1, pp. 17–27, 2004.

[19] 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.*, vol. 10, no. 1, pp. 17–25, 2007.

[20] J. Sauvola, T. Seppänen, S. Haapakoski, and M. Pietikäinen, "Adaptive Document Binarization," in *International Conference on Document Analysis and Recognition*, 1997, vol. 1, pp. 147–152.

[21] W. Niblack, *An Introduction to Digital Image Proc.*, pp. 115–116, Prentice Hall, 1986.