Maximum Entropy Models for Skin Detection*

Bruno Jedynak
Laboratoire de
Mathématiques Appliquées, USTL
Bât M2, Cité scientifique
59655 Villeneuve d'Ascq, France
Bruno.Jedynak@univ-lille1.fr

Huicheng Zheng, Mohamed Daoudi and Didier Barret
MIIRE Group
ENIC-Telecom Lille 1 / INT
Rue G. Marconi, Cité scientifique
59655 Villeneuve d'Ascq, France
(Zheng,Daoudi)@enic.fr

Abstract

We consider a sequence of three models for skin detection built from a large collection of labelled images. Each model is a maximum entropy model with respect to constraints concerning marginal distributions. Our models are nested. The first model, called the baseline model is well known from practitioners. Pixels are considered as independent. Performance, measured by the ROC curve on the Compaq Database is impressive for such a simple model. However, single image examination reveals very irregular results. The second model is a Hidden Markov Model which includes constraints that force smoothness of the solution. The ROC curve obtained shows better performance than the baseline model. Finally, color gradient is included. Thanks to Bethe tree approximation, we obtain a simple analytical expression for the coefficients of the associated maximum entropy model. Performance, compared with previous model is once more improved.

1. Introduction

Skin detection consists in detecting human skin pixels from an image. The system output is a binary image defined on the same pixel grid as the input image.

Skin detection plays an important role in various applications such as face detection [13], searching and filtering image content on the web [15][14]. Research has been performed on the detection of human skin pixels in color images and on the discrimination between skin pixels and "non-skin" pixels by use of various statistical color models. Some researchers have used skin color models such as Gaussian , Gaussian mixture or histograms [12][10]. In most experiments, skin pixels are acquired from a limited number of people under a limited range of lighting conditions.

Unfortunately, the illumination conditions are often unknown in an arbitrary image, so the variation in skin colors is much less constrained in practice. This is particularly true for web images captured under a wide variety of conditions. However, given a large collection of labeled training pixels including all human skin (Caucasians, Africans, Asians) we can still model the distribution of skin and non-skin colors in the color space. Recently Jones and Rehg [9] proposed techniques for skin color detection by estimating the distribution of skin and non-skin color in the color space using labeled training data. The comparison of histogram models and Gaussian mixture density models estimated with EM algorithm was analyzed for the standard 24-bit RGB color space. The histogram models were found to be slightly superior to Gaussian mixture models in terms of skin pixel classification performance for this color space.

A skin detection system is never perfect and different users use different criteria for evaluation. General appearance of the skin-zones detected, or other global criteria might be important for further processing. For quantitative evaluation, we will use false positives and detection rates. False positive rate is the proportion of non-skin pixels classified as skin and detection rate is the proportion of skin pixels classified as skin. The user might wish to combine these two indicators his own way depending on the kind of error he is more willing to afford. Hence we propose a system where the output is not binary but a floating number between zero and one, the larger the value, the larger the belief for a skin pixel. The user can then apply a threshold to obtain a binary image. Error rates for all possible thresholding are summarized in the Receiver Operating Characteristic (ROC) curve.

We have in our hands the Compaq Database [9]. It is a catalog of almost twenty thousand images. Each of them is manually segmented such that the skin pixels are labelled. Our goal in this paper is to explore different ways in which this set of data can be used to perform skin detection on new images. We will use the well known Markov Random Field approach [16] [3], as well as Maximum Entropy Modeling [7] [4] chapter 11, referred to as MaxEnt.

The rest of this paper is organized as follows: After set-

*This work was partially supported by Internet Action Plan Contract Number POESIA-2117/27572

ting up the notations in section 2, section 3 will present a very simple and crude model that we will refer to as the baseline model. In section 4, we present a hidden Markov Random Field model that takes into account the spatial regularity of skin and non-skin regions. A novel method for parameter estimation will be explored. In section 5, we will examine models that take into account joint distributions between nearby pixels in skin regions as well as in non-skin regions. Finally, in Section 6 we will present concluding remarks.

2 Notations

Let's fix the notations. The set of pixels of an image is S. The color of a pixel $s \in S$ is x_s. It is a 3 dimensional vector, each component being coded on one octet. We notate $C = \{0, \ldots, 255\}^3$. The "skinness" of a pixel s, is y_s with $y_s = 1$ if s is a skin pixel and $y_s = 0$ if not. We will also use the term "label" in place of skinness. The color image, which is the vector of color pixels, is x and the binary image made up of the y_s's is notated y.

Let's assume for a moment that we knew the joint probability distribution $p(x, y)$ of the vector (x, y), then Bayesian analysis tells us that, whatever cost function the user might think of, all that is needed is the posterior distribution $p(y|x)$.

From the user's point of view, the useful information is contained in the one pixel marginal of the posterior, that is, for each pixel, the quantity $p(y_s = 1|x)$, quantifying the belief for skinness at pixel s. In practice the model $p(x, y)$ is unknown. Instead, we have the Compaq Database. It is a collection of samples

$$\{(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})\}$$

where for each $1 \leq i \leq n$, $x^{(i)}$ is a color image and $y^{(i)}$ is the associated binary skinness image. We assume that the samples are independent of each other with distribution $p(x, y)$. The collection of samples is referred later as the training data. Probabilities are estimated by using classical empirical estimators and are denoted with the letter q.

In what follows, we build models for the probability distribution of the skinness image given the color image using maximum entropy modeling.

3 Baseline Model

3.1 Defining the model

First, we build a model that respects the one pixel marginal observed in the Compaq Database. That is, for each image x, consider the set of probability distributions over binary images defined on the same grid as x that verify:

$$\mathcal{C}_0(x) : \forall s \in S, \forall x_s \in C, \forall y_s \in \{0, 1\}, p(y_s|x_s) = q(y_s|x_s)$$

In this expression, the quantity on the right side of the equal sign doesn't depend on the particular location s. It is the proportion of pixels with label y_s, among the ones with color x_s in the training data. For each x, The MaxEnt solution under $\mathcal{C}_0(x)$, using Lagrange multipliers is the independent model:

$$p(y|x) = \prod_{s \in S} q(y_s|x_s)$$

We call this model the baseline model. It is the most commonly used model in the literature [12][10].

3.2 Experiments

Each term of the product on the right side can then be computed using probabilities estimated on the training data as follows using Bayes formula:

$$q(y_s|x_s) = \frac{1}{q(x_s)} q(x_s|y_s) q(y_s) \quad (1)$$

with

$$q(x_s) = \sum_{y_s=0}^{1} q(x_s|y_s) q(y_s)$$

Evaluation of the quantities in (1) is based on two 3-dimension histograms, $q(x_s|y_s = 1)$ and $q(x_s|y_s = 0)$ describing the one pixel color skin regions and non-skin regions respectively. Several authors have tried to get a parametric expression for these histograms as a mixture of Gaussian distribution [9] [13]. Our experience is that the Compaq Database is large enough so that crude histograms made with one color value per bin do not over-fit. The ROC curve for this model is presented in figure 2. Experiments for this model, as well as for the other ones were made using the following protocol. The Compaq database contains about 18,696 photographs. It was split into two almost equal parts randomly. The first part, containing nearly 2 billion pixels was used as training data while the other one, the test set, was let aside for ROC curve computation. Figure 3, Top Left is one of the test images. This is a color image. Top right is a grey level image. The grey-level is proportional to the quantity $p(y_s = 1|x)$ evaluated with the Baseline model. Many skin pixels are not detected. Figure 2 show ROC curves computed from 100 images (around 10 millions pixels), randomly extracted from the test set. The Baseline model (with crosses) permit to detect more than 80% of the skin pixels with less than 10% of false positive rate.

4 Hidden Markov Model

4.1 Defining the model

The baseline model is certainly too loose and one might hope to get better detection results by constraining it to a model that takes into account the fact that skin zones are not purely random but are made of large regions with regular shapes. Hence, we fix the marginals of y for all the neighboring pixels couples. We use 4-neighbors system for simplicity in all that follows. For 2 neighboring pixels s and t, the expected proportion of times that we observe $(y_s = a, y_t = b)$ should be $q(a, b)$ for $a = 0, 1$ and $b = 0, 1$, the corresponding quantities measured on the training set. We assume that the model is isotropic, aggregating the cases where s and t are in vertical position to the cases where s and t are in horizontal position. We also assume that the prior model is symmetric, that is $p(y_s, y_t) = p(y_t, y_s)$. Hence let us define the following constraints:

$$\mathcal{D} : \forall s \in S, \forall t \in \mathcal{V}(s),$$

$p(y_s = 0, y_t = 0) = q(0, 0)$ and $p(y_s = 1, y_t = 1) = q(1, 1)$

where $\mathcal{V}(s)$ are the 4 neighbors of s. For each image x, the MaxEnt model under

$$\mathcal{C}_1(x) = \mathcal{C}_0(x) \cap \mathcal{D}$$

is then the following Gibbs distribution [16].

$$p(y|x) = \frac{p(y)}{p(x)} \prod_{s \in S} q(x_s|y_s) \qquad (2)$$

with

$$p(y) = \frac{1}{Z} \exp \sum_{<s,t>} (a_0(1 - y_s)(1 - y_t) + a_1 y_s y_t) \qquad (3)$$

where the sum ranges over all pairs of 4-neighbors pixels $< s, t >$, Z is a normalizing constant and a_0 and a_1 are two parameters that should be set up such that the constraints $\mathcal{D}$ are satisfied. The model in equation (3) is known as a Potts model [16].

4.2 Parameter estimation

Parameter estimation in the context of MaxEnt is still an active research subject, especially in situations where even the likelihood function cannot be computed for a given value of the parameters. This is the case here since the so-called partition function Z, viewed as a function of a_0 and a_1, cannot be evaluated even for very small size images. One line of research consists in approximating the model in order to obtain a formula where the partition function no longer appears: Pseudo-likelihood [1], [5] and mean field methods [20], [2] are among them. Another possibility is to use stochastic gradient as in [19]. Here we explore a related method based on the concept of Julesz ensembles defined in [18]. We learn from this work that one can sample an image from the model defined in (3) without knowing the parameters a_0 and a_1. This is true only in the asymptotic of an infinite image but we will apply the result for a large image, say 512x512 pixels. In a second step, we use this sample image in order to estimate the parameters a_0 and a_1. This is done using the quantity $p(y_s = 1|y_{(s)})$ which is the probability to observe the label 1 at pixel s given all the other values y_t, for $t \in S$ and $t \neq s$. For the model in (3), this quantity can be easily analytically computed as

$$p(y_s = 1|y_{(s)}) = \phi((a_1 + a_0)n_s(1) - 4a_0)$$

where $\phi(x) = (1 + e^{-x})^{-1}$ is the logistic[1] function and $n_s(1)$ is the number of neighbors of s that take the label 1. This sum can take only five different values. For each one, the quantity $p(y_s = 1|y_{(s)})$ can be estimated from the sample image, leading to five linearly independent equations from which parameters a_0 and a_1 can be estimated. Now, returning to how to obtain a sample from the model in (3). The key idea which originated in statistical physics [11], is that the MaxEnt model we are looking for is, in an appropriate asymptotic meaning, the uniform distribution over the set of images that respect the constraints $\mathcal{D}$. Now, sampling from this set can be achieved numerically using simulated annealing, see [6]. Details are presented in [8]. The obtained values are $a_0 = 3.76$ and $a_1 = 3.94$.

4.3 Experiments

For a new image x, skin detection requires to compute for each pixel the quantity $p(y_s|x)$. We do it for the model in (2) by Markov Chain Monte Carlo. We generate, using the Gibbs sampler algorithm [16], a sequence of label images

$$y^1, y^2, \ldots, y^{n_1}, \ldots, y^{n_2}$$

with stationary distribution the one in equation (5). Then, we estimate the quantity $p(y_s|x)$ by the empirical mean

$$\frac{1}{n2 - n1} \sum_{j=n_1+1}^{n_2} y_s^{(j)}$$

Our working parameters are $n_1 = 1$ and $n_2 = 100$. A output image is presented in Figure 3 Bottom Left. It compares favorably with the Baseline model. The ROC curve in Figure 2 indicates a drop of about 1% in false positive for the same detection rate as the Baseline model.

[1] also denoted sigmod

5 First Order Model

5.1 Defining the Model

The baseline model was built in order to mimic the one pixel marginal of the posterior, that is $q(y_s|x_s)$ as observed on the database. Then, in building the HMM model we added constraints on the prior $p(y)$ in order to smooth the model. Now, we constrain once more the MaxEnt model by imposing the two-pixel marginal of the posterior, that is $p(y_s, y_t|x_s, x_t)$, for 4-neighbor s and t, to match those observed in the training data. Hence we define for each image x, the following constraints:

$$\mathcal{C}_2(x) : \forall s \in S, \forall t \in \mathcal{V}(s), \forall x_s \in C, \forall x_t \in C,$$
$$\forall y_s \in \{0,1\}, \forall y_t \in \{0,1\},$$
$$p(y_s, y_t|x_s, x_t) = q(y_s, y_t|x_s, x_t)$$

The quantity $q(y_s, y_t|x_s, x_t)$ is the expected number of time we observe the values (y_s, y_t) for a couple of neighboring pixels among the couples of neighboring pixels with color values (x_s, x_t), regardless of the orientation of the pixels s and t in the training set.

Clearly, for each x, $\mathcal{C}_2(x) \subset \mathcal{C}_1(x)$. Using once more Lagrange multipliers, the solution to the MaxEnt problem under $\mathcal{C}_2(x)$ is then the following Gibbs distribution:

$$p(y|x) = \frac{1}{Z(x)} \exp\Big(\sum_{<s,t>} \lambda(x_s, x_t, y_s, y_t)\Big) \qquad (4)$$

where $Z(x)$ is a normalization function that depends on x but not on y and $\lambda(x_s, x_t, y_s, y_t)$ are parameters that should be set up to satisfy the constraints. Assuming that one color can take 256^3 values, the total number of parameters is $256^3 \times 256^3 \times 2 \times 2$. The previously mentioned parameter estimation methods clearly do not apply. In [17], the authors present a tree approximation to the pixel grid, called “Bethe tree”, after the physicist H.A. Bethe who used trees in statistical mechanics problems. Bethe trees permit us to compute analytically an approximation of the parameters in the model (4) as we shall see now.

5.2 Parameter estimation

The construction of Bethe trees is recursive. Figure 1 shows the first step. Successive steps are obtained by adding 3 new neighbors to each leaf.

Let us consider the following model

$$p(y|x) = \frac{1}{Z(x)} \exp H(x; y) \qquad (5)$$

with

$$H(x; y) = \sum_{<s,t>} \log q(y_s, y_t|x_s, x_t)) - 3 \sum_{s \in \mathring{S}} \log q(y_s|x_s)$$

where $Z(x)$ is a normalizing function of x and $\mathring{S}$ is the set of interior pixels of S, that is the ones that have exactly 4 neighbors. First, remark that the model in (5) is a special case of model in (4). Secondly, we verify that under the Beth tree approximation, with arbitrarily finite depth, the model in (5) satisfies the constraints. The proof is in the Appendix.

Now, let us see how in practice one can use the model in (5). As for the HMM model, the objective is to obtain simulations using the Gibbs sampler algorithm. This requires to compute the conditional distribution of a label y_s given all the other labels and the image of the colors x. For $s \in \mathring{S}$, we obtain

$$p(y_s = 1|y_{(s)}, x) = \phi(U(x; y)) \qquad (6)$$

with

$$U(x; y) = \sum_{t \in \mathcal{V}(s)} \log \frac{q(y_s = 1, y_t|x_s, x_t)}{q(y_s = 0, y_t|x_s, x_t)} - 3 \log \frac{q(y_s = 1|x_s)}{q(y_s = 0|x_s)}$$

Where ϕ is the sigmod function and $\mathcal{V}(s)$ are the 4 neighbors of s.

5.3 Experiments

Now let’s see how each term in (6) can be evaluated. First,

$$\frac{q(y_s = 1|x_s)}{q(y_s = 0|x_s)} = \frac{q(x_s|y_s = 1)}{q(x_s|y_s = 0)} \frac{q(y_s = 1)}{q(y_s = 0)}$$

and the quantities on the right side are easily obtained from the database as before. Second,

$$\frac{q(y_s = 1, y_t|x_s, x_t)}{q(y_s = 0, y_t|x_s, x_t)} = \frac{q(x_s, x_t|y_s = 1, y_t)}{q(x_s, x_t|y_s = 0, y_t)} \frac{q(y_s = 1, y_t)}{q(y_s = 0, y_t)}$$

Now the quantities on the right side involving the color values cannot be directly extracted from the database without drastic over-fitting since the histogram involved has a support of dimension six. Hence some kind of dimension reduction is needed.

One natural solution is to assume conditional independence, that is

$$\frac{q(x_s, x_t|y_s = 1, y_t)}{q(x_s, x_t|y_s = 0, y_t)} = \frac{q(x_s|y_s = 1)}{q(x_s|y_s = 0)}$$

The obtained model is then a HMM model, as in equation (2). Hence, Bethe tree method gives another way to estimate parameters a_0 and a_1. Obtained values are $a_0 = 3.94$ and $a_1 = 4$, which are close to the values obtained in section 4.

A more promising dimension reduction procedure is the following approximation

$$q(x_s, x_t|y_s, y_t) \approx q(x_s|y_s) q(x_t - x_s|y_s, y_t)$$

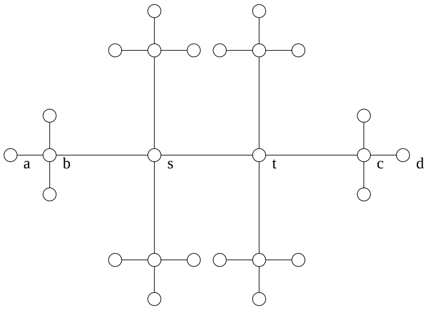

Figure 1: A Bethe tree approximation of the pixel graph at the neighborhood of pixels s and t

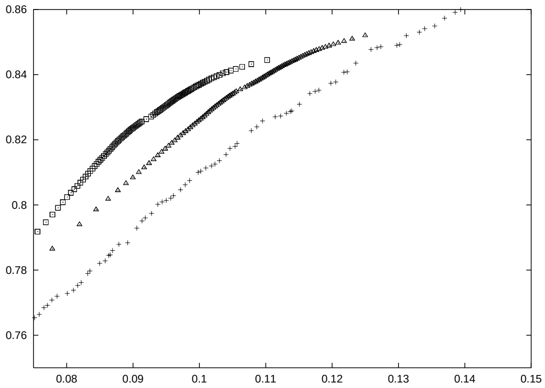

Figure 2: Receiver Operating Characteristics (ROC) curve for each model. x-axis is the false positive rate, y-axis is the detection rate which is the complement to 1 of the false negative rate. Baseline model is shown with crosses, HMM model with triangles, while the first order model is shown with squares.

That is, we assume that the color gradient at s, measured by the quantity $x_t - x_s$, is, given the labels at s and t, independent of the actual color x_s. Evaluation of the right side of the sign $\approx$ requires to compute 6 histograms with a support of dimension 3 only.

Experiments with this model are presented in figures 2 and 3. The setup is the same as for the HMM model. In figure 3, Bottom Right, one can visually appreciate the improvement in localization of the skin zones compared to the HMM model. Bulk results in the ROC curve of Figure 2 shows a slight improvement of performance too.

6 Conclusions

We have considered a sequence of three models for skin detection built from a large collection of labelled images. For a given color image, such a model puts weight on binary images defined on the same pixel grid. Each model is a maximum entropy model with respect to constraints. These constraints concern marginal distributions. Our models are nested. The first model, called the baseline model is well known from practitioners. Pixels are considered as independent. Performance, measured by the ROC curve on the Compaq database is impressive for such a simple model. However, single image examination reveals very irregular results. The second model is a Hidden Markov Model. It includes constraints that force smoothness of the solution. The ROC curve obtained shows better performance than the baseline model. Finally, color gradient is included in the set of constraints. Thanks to Bethe tree approximation, we obtain a simple analytical expression for the coefficients of the associated MaxEnt model. Performance, compared with previous model is once more improved.

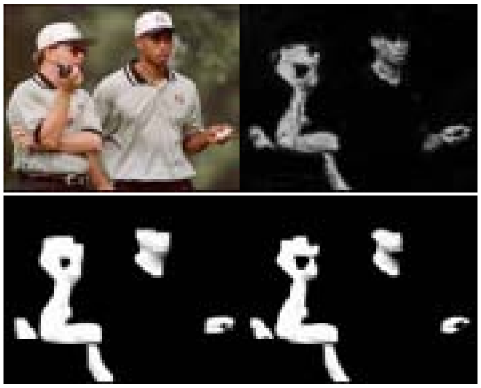

Figure 3: Top left: the original image. Top right: the result of the Baseline model. Bottom left: the result of the HMM model. Bottom right: the result of the first order model

7 Appendix

Following is the proof that under the Bethe tree approximation, the model in (5) satisfies the constraints $\mathcal{C}_2$. We restrict the proof to s and t in the interior of S. First, consider the Bethe tree of depth 1 shown in Figure 1. In order to simplify the writing, we write $r(y_s, y_t)$ for $q(y_s, y_t | x_s, x_t)$, and $r(y_s)$ for $q(y_s | x_s)$.

Starting from (5), we have

$$p(y_s, y_t) = \frac{N(y_s, y_t)}{D}$$

with

$$D = \sum_{y_s=0}^{1} \sum_{y_t=0}^{1} N(y_s, y_t)$$

and

$$N(y_s, y_t) = \sum_{y_{(s,t)}} \exp H(x; y)$$

This last sum ranges over all the possible values for all the labels except the ones at s and t. Now, from the Bethe tree in Figure 1,

$$N(y_s, y_t) = N_1(y_s, y_t)N_2(y_s)N_3(y_t)$$

with

$$N_1(y_s, y_t) = r(y_s, y_t)r^{-3}(y_s)q^{-3}(y_t)$$

$$N_2(y_s) = (\sum_{y_b=0}^{1} r(y_b, y_s)r^{-3}(y_b)(\sum_{y_a=0}^{1} r(y_b, y_a))^3)^3$$

$$N_3(y_t) = (\sum_{y_c=0}^{1} r(y_t, y_c)r^{-3}(y_c)(\sum_{y_d=0}^{1} r(y_c, y_d))^3)^3$$

Hence

$$N(y_s, y_t) = r(y_s, y_t)$$

which concludes the proof. One can easily extend the argument to Bethe tree of arbitrary fixed depth.

References

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

[2] G. Celeux, F. Forbes, and N. Peyrard. Em procedures using mean field-like approximations for markov model-based image segmentation. *to appear in Pattern Recognition*, 2002.

[3] R. Chellappa and A. Jain, editors. *Markov Random Fields: Theory and Applications*. Academic Press, 1996.

[4] Cover and Thomas. *"Elements of Information Theory"*. Wiley, 1991.

[5] F. Divino and A. Frigessi. Penalized pseudolikelihood inference in spatial interaction models with covariates. *to appear on the Scandinavian Journal of Statistics*, 2000.

[6] S. Geman and D. Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on PAMI*, 6:721–741, 1984.

[7] E. Jaynes. Information theory and statistical mechanics. *Phisical Review*, 106, 1957.

[8] B. Jedynak, H. Zheng, M. Daoudi, and D. Barret. Maximum entropy models for skin detection. Technical Report publication IRMA, Volume 57, number XIII, Université des Sciences et Technologies de Lille, France, 2002.

[9] M. Jones and J. M. Rehg. Statistical color models with application to skin detection. In *Computer Vision and Pattern Recognition*, pages 274–280, 1999.

[10] M. J. Jones and J. M. Rehg. Statistical color models with application to skin detection. Technical Report CRL 98/11, Compaq, 1998.

[11] A. Martin-Lof. The equivalence of ensembles and gibbs'phase rule for classical lattice-systems. *Journal of Statistical Phisics*, 20:557–569, 1979.

[12] J.-C. Terrillon, M. David, and S. Akamatsu. Automatic detection of human faces in natural scene images by use of a skin color model and of inavariant moments. In *IEEE Third International Conference on Automatic Face and gesture Recognition*, pages 112–117, 1998.

[13] J.-C. Terrillon, M. N. Shirazi, H. Fukamachi, and S. Akamatsu. Comparative performance of different skin chrominance models and chrominance spaces for the automatic detection of human faces in color images. In *Fourth International Conference On Automatic Face and gesture Recognition*, pages 54–61, 2000.

[14] J. Z. Wang, J. Li, G. Wiederhold, and O. Firschein. Classifying objectionable websites based on image content. *Notes in Computer Science, Special issue on iteractive distributed multimedia systems and telecommunication services*, 21/15:113–124, 1998.

[15] J. Z. Wang, J. Li, G. Wiederhold, and O. Firschein. System for screening objectionable images. *Images, Computer Communications Journal*, 1998.

[16] G. Winkler. *Image Analysis, Random Fields and Dynamic Monte Carlo Methods*. Springer-Verlag, 1995.

[17] C. Wu and P. C. Doerschuk. Tree approximations to markov random fields. *IEEE Transactions on PAMI*, 17(4):391–402, April 1995.

[18] Y. Wu, S. Zhu, and X. Liu. Equivalence of julesz ensemble and frame models. *International Journal of Computer Vision*, 38(3):247–265, July 2000.

[19] L. Younes. Estimation and annealing for gibbsian fields. *Annales de l'Institut Henry Poincare, Section B, Calcul des Probabilits et Statistique*, 24:269–294, 1998.

[20] J. Zhang. The mean field theory in em procedure for markov random fields. *IEEE Transactions on Signal Processing*, 40(10):2570–2583, October 1992.