Electronic Letters on Computer Vision and Image Analysis 4(2):1-14, 2004

Blocking Adult Images Based on Statistical Skin Detection

Huicheng Zheng*, Mohamed Daoudi† and Bruno Jedynak+

* *MIIRE Group, LIFL (CNRS UMR 8022) / INT*
ENIC-Telecom Lille 1, Rue G. Marconi, Cité Scientifique, 59655 Villeneuve d'Ascq, France
† *Laboratoire d'Informatique de l'Université de Tours (EA 2101)*
64 Avenue Jean Portalis, 37200 Tours, France
+ *Center for Imaging Science, The Johns Hopkins University, U.S.A.*
+ *Laboratoire de Mathématiques Appliquées, USTL, Bât M2, Cité Scientifique, 59655 Villeneuve d'Ascq, France*

Received 20 September 2004; accepted 3 November 2004

Abstract

This work is aimed at the detection of adult images that appear in Internet. Skin detection is of the paramount importance in the detection of adult images. We build a maximum entropy model for this task. This model, called the First Order Model in this paper, is subject to constraints on the color gradients of neighboring pixels. Parameter estimation as well as optimization cannot be tackled without approximations. With Bethe tree approximation, parameter estimation is eradicated and the Belief Propagation algorithm permits to obtain exact and fast solution for skin probabilities at pixel locations. We show by the Receiver Operating Characteristics (ROC) curves that our skin detection improves the performance over previous work [6] in the context of skin pixel detection rate and false positive rate. The output of skin detection is a grayscale skin map with the gray level indicating the belief of skin. We then calculate 9 simple features from this map which form a feature vector. Most of these features are based on fit ellipses, which are used to catch the characteristics of detected skin regions. Two fit ellipses are used for each skin map—the fit ellipse of all skin regions and the fit ellipse of the largest skin region. They are called respectively Global Fit Ellipse and Local Fit Ellipse in this paper. A multi-layer perceptron classifier is then trained for these features. Plenty of experimental results are presented, including photographs and a ROC curve calculated over a test set of $5,084$ photographs, which show stimulating performance for such simple features.

Key Words: Maximum Entropy Modeling, Markov Random Field, Belief Propagation Algorithm, Multi-Layer Perceptron, Skin Detection, Adult Image Detection.

1 Introduction

Images are an essential part of today's World Wide Web. The statistics of more than 4 million HTML webpages reveal that 70.1% of web pages contain images and that on average there are about 18.8 images per HTML

Correspondence to: <zheng@enic.fr>

This work is partially supported by European Community IAP 2117/27572-POESIA www.poesia-filter.org, http://sourceforge.net/projects/poesia/
Recommended for acceptance by Manuele Bicego
ELCVIA ISSN:1577-5097
Published by Computer Vision Center / Universitat Autònoma de Barcelona, Barcelona, Spain

webpage[23]. These images are mostly used to make attractive Web contents or to add graphical items to mostly textual content, such as navigational arrows.

However, images are also contributing to harmful (e.g. pornographic) or even illegal (e.g. paedophiliac) Internet content. So effective filtering of images is of paramount importance in an Internet filtering solution.

To block adult content, some representative companies as NetNanny and SurfWatch, operate by maintaining lists of URL's and newsgroups and require constant manual updating. Abundant literature is available, but the Internet is very rapidly evolving, not only quantitatively. Each day, 3 million pages are appearing on the Web. Detection based on image content analysis has the advantage to process equally all the images without the need for updating, so will produce more effective filtering.

By taking advantage of the fact that there is a strong correlation between images with large patches of skin and adult images we have to develop a skin detector. Skin color offers an effective and efficient way to detect the adult image content. There is already a large amount of work on this track.

The WIPE [4] system developed by Wang, Li, Wiederhold and Firschein uses a manually-specified color histogram model as a prefilter in an analysis pipeline. Input images whose average probability of skin is low are accepted as non-offensive. Images that contain considerable skin pass on to a final stage of analysis where they are classified using wavelet features. The algorithm uses a combination of Daubechies wavelets, normalized central moments, and color histograms to provide semantically-meaningful feature vector matching.

Forsyth's [5] research group has designed and implemented an algorithm to screen images of naked people. Their algorithms involve a skin filter and human figure grouper. The skin color model used by Fleck, Forsyth and Bregler consists of a manually specified region in a log-opponent color space. Detected regions of skin pixels form the input to a geometric filter based on skeletal structure. As indicated in their paper, 52.2% sensitivity and 96.6% specificity have been obtained for a test set of 138 images with naked people and 1401 assorted benign images. However, it takes about 6 minutes on a workstation for the figure grouper in their algorithm to process a suspect image passed by the skin filter. Most of the people in the images used in the experimental protocol are Caucasians and a small number of images are Blacks or Asians.

Jones and Rehg [6] propose techniques for skin color detection by estimating the distribution of skin and non-skin color in the color space using labeled training data. To detect adult images, some simple features are extracted. The discrimination performance based solely on skin is rather good for such simple features.

Bosson et al. [7] propose a pornography detection system which is integrated in a commercial system. This system is also based on skin detection. They compared the generalised linear model, the k-nearest neighbor classifier, the multi-layer perception (MLP) classifier and the support vector machine and found that the MLP gives the best classification performance.

Our approach is as follows. The first main step is skin detection. We build a model with Maximum Entropy Modeling (MaxEnt) for the skin distribution given the input color image. This model is a First Order Model (FOM) that imposes constraints on color gradients of neighboring pixels. Parameter estimation as well as optimization cannot be tackled without approximations. With Bethe tree approximation, parameter estimation is eradicated and the Belief Propagation (BP) algorithm permits to obtain exact and fast solution for skin probabilities at pixel locations. This model is referred to as TFOM for Tree First Order Model. The output of skin detection is a grayscale *skin map* with the gray levels being proportional to the skin probabilities. The second main step is pattern recognition. Some features are extracted from the skin map which compose a feature vector. We train a MLP classifier on $5,084$ patterns from the training set. In the test phase, the MLP classifier takes a quick decision on the input pattern in one pass.

The rest of this paper is organized as follows: in section 2 we build the model for skin detection and present the pixel classification performance of this model by Receiver Operating Characteristics (ROC) curves. Section 3 is devoted to feature extraction and pattern recognition. In section 4, some experimental results are presented. Section 5 concludes the paper.

2 Skin Detection

Skin detection consists in detecting human skin pixels from an image [2]. It plays an important role in various applications such as face detection [1] [3], searching and filtering image content on the web [4][5] [6]. A Self Organizing Map (SOM) for skin detection was proposed by [26]. In most experiments, skin pixels are acquired from a limited number of people under a limited range of lighting conditions.

Skin color segmentation approaches can be grouped into two basic categories:physical-based approaches and statistical approaches. Statistical approaches can be subdivided further into parametric approaches [28] [29] [3] and non parametric approaches [30] [31] . Parametric model approaches represent the skin color distribution in parametric form , such as Gaussian or Gaussian mixture [1][3].

In nonparametric approaches histograms are used to represent density in color space. The illumination conditions are often unknown in an arbitrary image, so the variation in skin colors is much less constrained than in controlled setup. 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 [6] proposed techniques for skin color detection by estimating the distribution of skin and non-skin color in the color space using labeled training data. Both parametric and nonparametric statistical approaches usually perform color-segmentation in color spaces that reduce the varying illuminant. A number of different color spaces have been used, however, normalized RGB and HSV are the most common color spaces used [27]. 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 [6].

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 publicly available Compaq Database [6]. It is a catalog of almost twenty thousand images. Each of them is manually segmented such that the skin pixels are labelled. Our goal is to infer a model from this set of data in order to perform skin detection on new images.

2.1 Maximum entropy model

The maximum entropy model (MaxEnt) has along history. It can be seen as a generalization of the "Principle of Insufficient Reason". This strategy was first proposed as a general inference procedure by [8]. Suppose we are about to estimate a distribution subject to some constraints. These contraints could be statistics we observed from a lot of samples from the underlying distribution. That is, we constraint our solution to those distributions conforming with our observations. The solutions could be enormous if we do not have any other prior information. Then which distribution should we choose? Maximum entropy principle solves this problem by the policy of honesty, that is, frankly acknowledge the full extent of its ignorance by taking into account all possibilities allowed by the knowledge [8]. We would not rule out any possibilities except the constraints explicitly tell us to do so. We would accept the solution that is in line with the constraints, and otherwise as uniform as possible. Under maximum entropy principle, the uniformity of a distribution is measured by the information entropy.

MaxEnt works as follows: (1) choose relevant features (2) compute their histograms on the training set

(3) write down the maximum entropy model within the ones that have the feature histograms as observed on the training set (4) estimate the parameters of the model (5) use the model for classification. This plan has been successfully completed for several tasks related to speech recognition and language processing. See for example [9] and the references therein. In these applications the underlying graph of the model is a line graph or even a tree but in all cases it has no loops. When working with images, the graph is the pixel lattice. It has indeed many loops. A breakthrough appeared with the work in [10] on texture simulation where (1)–(4) were performed for images and (5) replaced by simulation.

We adapt to skin detection as follows: in (1) we specialize in colors for two adjacent pixels and their "skinness". We choose RGB color space in our approach. In practice we know from [6][7] that the choice of color space is not critical given a histogram-based representation of the color distribution and enough training data. In (2) we compute the histogram of these features in the Compaq manually segmented database. Models for (3) are then easily obtained. In (4) we use the Bethe tree approximation, see [11]. It consists in approximating locally the pixel lattice by a tree. The parameters of the MaxEnt models are then expressed analytically as functions of the histograms of the features. This is a particularity of our features. In (5) we pursue the approximation in (4): we use the BP algorithm, see [12], which is exact in tree graph but only approximative in loopy graphs.

Indeed, one of us had already witnessed in a different context that tree approximation to loopy graph might lead to effective algorithms, see [13].

Let's fix the notations. The set of pixels of an image is S. We notate $C = \{0, \ldots, 255\}^3$. The color of a pixel $s \in S$ is x_s, $x_s \in C$. 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. The set of neighbors of s is notated as $\mathcal{V}(s)$. We use 4-neighbor system here. $< s, t >$ denotes such a pair of neighbors s and t, regardless of the orientation. The color image, which is the vector of color pixels, is notated x and the binary image made up of the y_s's is notated y.

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. Bayesian analysis tells us that, whatever cost function the user might think of, all that is needed is the joint distribution $p(x, y)$. In practice the model $p(x, y)$ is unknown. Instead, we have the segmented 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 using classical empirical estimators and are denoted with the letter q.

In what follows, we build a model for the probability distribution of the skinness image given the color image using MaxEnt.

2.2 First Order Model (FOM)

Our MaxEnt model respects the two-pixel marginal of the joint distribution of color and skinness as observed in the training database, that is, for two adjacent sites s and t, $p(x_s, x_t, y_s, y_t)$ should match those observed in the training data. Hence we define the following constraints:

$$\begin{gathered} \mathcal{C} : \forall s \in S, \forall t \in \mathcal{V}(s), \forall x_s \in C, \forall x_t \in C, \\ \forall y_s \in \{0, 1\}, \forall y_t \in \{0, 1\}, \\ p(x_s, x_t, y_s, y_t) = q(x_s, x_t, y_s, y_t) \end{gathered} \tag{1}$$

The quantity $q(x_s, x_t, y_s, y_t)$ is the proportion of times we observe the values (x_s, x_t, y_s, y_t) for a couple of neighboring pixels, regardless of the orientation of the pixels s and t in the training set.

Here we shall derive a MaxEnt solution for the joint distribution $p(x, y)$ under the constraints $\mathcal{C}$.

Remark that the constraints in (1) are expectations with respect to p. Indeed,

$$p(x_s, x_t, y_s, y_t) = E_p[\delta_{x_s}(X_s)\delta_{x_t}(X_t)\delta_{y_s}(Y_s)\delta_{y_t}(Y_t)] \tag{2}$$

with

$$\delta_a(b) = \begin{cases} 1 & \text{if } a = b \\ 0 & \text{if } a \neq b \end{cases}$$

Using Lagrange multipliers[8], the solution to the MaxEnt problem under $\mathcal{C}$ is then the following Gibbs distribution:

$$p(x, y) \approx \prod_{<s,t>} \lambda(x_s, x_t, y_s, y_t) \tag{3}$$

where $\lambda(x_s, x_t, y_s, y_t) > 0$ are parameters that should be set up to satisfy the constraints. The sign $\approx$ indicates here and after equality up to a function that depends possibly on x but not on y. This function is called the partition function in statistical mechanics.

Using Bayes formula, one then obtains:

$$p(y|x) \approx \prod_{<s,t>} \lambda(x_s, x_t, y_s, y_t) \tag{4}$$

Assuming that one color can take 256^3 values, the total number of parameters is $256^3 \times 256^3 \times 2 \times 2$.

Parameter estimation in the context of MaxEnt is still an active research subject, especially in situations where the likelihood function cannot be computed for a given value of the parameters. This is the case here, since the partition function 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 [14], [15] and mean field methods [16], [17] are among them. Another possibility is to use stochastic gradient as in [18]. However, due to the large number of parameters in the FOM model, this is a real challenge.

Moreover, recall that the quantities of interest for the users are the one pixel marginal of the posterior, that is for each s the quantity $p(y_s = 1|x)$. These quantities are not easily available due once more to the impossibility of evaluating the partition function. One has then to use stochastic algorithm as the Gibbs sampler which is time consuming or to rely on an approximate model.

Bethe Tree approximation deals with parameter estimation by giving us a simple analytical model free of parameter. We can further implement BP algorithm to achieve fast computation as we shall see now.

2.3 Bethe Tree Approximation of FOM (TFOM)

The FOM defined in (4) is a Markov Random Field (MRF) on the non-oriented pixel graph with 4-neighbor connectivity. See Appendix for an introductory description of such models. Let us assume for now that this graph was a tree: that is a connected graph without loops. Then, the MaxEnt solution for fixed x under $\mathcal{C}$ would be

$$p(x, y) \approx \prod_{<s,t>} \frac{q(x_s, x_t, y_s, y_t)}{q(x_s, y_s)q(x_t, y_t)} \prod_{s \in S} q(x_s, y_s) \tag{5}$$

The proof is as follows: we know from [25] that any pairwise MRF on a tree graph can be written

$$p(z) \approx \prod_{<s,t>} \frac{q(z_s, z_t)}{q(z_s)q(z_t)} \prod_{s \in S} q(z_s) \tag{6}$$

where $q(z_s)$ is the one-site marginal of p and $q(z_s, z_t)$ is its two-site marginal.

Applying this result to $z = (x, y)$ permits to obtain the model in equation (5). By construction it is in $\mathcal{C}$. Moreover it has the same form as the one in equation (3) which concludes the proof.

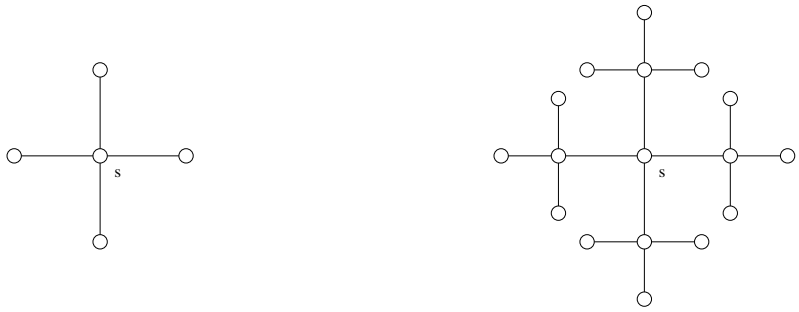

Figure 1: Left: a Bethe tree of depth 1 rooted at s. Right: a Bethe tree of depth 2 rooted at s

Using Bayes formula, one then gets:

$$p(y|x) \approx \prod_{<s,t>} \frac{q(x_s, x_t|y_s, y_t)q(y_s, y_t)}{q(x_s|y_s)q(x_t|y_t)q(y_s)q(y_t)} \prod_{s\in S} q(x_s|y_s)q(y_s) \tag{7}$$

The main point to be made here is that the model in equation (7) is parameter free.

We use Bethe trees to simulate the pixel graph. Bethe trees are named after the physicist H.A. Bethe, who used trees in statistical mechanics problems. They have been introduced in computer vision as a way of approximating estimators in MRF models in [11]. We shall revisit this work in connection with maximum entropy models.

The key idea is to provide a tree that approximates locally the pixel lattice. More precisely, for each pixel s, we consider a sequence of trees $\mathcal{T}_1^s, \mathcal{T}_2^s, \ldots$ of increasing depth. The construction is as follow: the root node of the tree is associated with s. For each neighbor t of s in the pixel-graph, a child node indexed by t is add to the root node. This defines $\mathcal{T}_1$. Subsequently, for each u, neighbor of a neighbor of s, (excluding s itself), a grandchild node indexed by u is added to the appropriate child node. This defines $\mathcal{T}_2$, and so on, see [11] for a detailed account. An important remark is that a single pixel might lead to several different nodes in the tree! For example $\mathcal{T}_2^s$ is build with s, the neighbors of s and the neighbors of these. Using 4-neighbors , and assuming that s is not in the border of the image, this makes up 13 pixels, but the associated tree has 17 nodes, 4 pixels being replicated twice each, see Fig. 1.

Nevertheless, the TFOM model in (7) cannot be directly used as it is. Indeed, the quantities $q(x_s, x_t|y_s, y_t)$ cannot be directly extracted from the database without drastic over-fitting. In effect the four histograms involved have a support of dimension six, three dimensions for each pixel. Hence, some kind of dimension reduction is needed. We have experimented the following:

$$q(x_s, x_t|y_s, y_t) \sim q(x_s|y_s)q(x_t - x_s|y_s, y_t) \tag{8}$$

That is, we assume that the color gradient at s, measured by the quantity $x_t - x_s$, is, given the skinness at s and t, independent of the actual color x_s. Evaluation of the right side of the sign $\sim$ requires to compute 6 histograms with a support of dimension 3 and with 32 bins for each dimension.

2.4 Belief Propagation (BP) Algorithm

In the previous section we proved that the quantity $p(y|x)$ can be expressed as follows:

$$p(y|x) \approx \prod_{<s,t>} \psi(x_s, x_t, y_s, y_t) \prod_{s\in S} \phi(x_s, y_s) \tag{9}$$

where

$$\psi(x_s, x_t, y_s, y_t) = \frac{q(x_s, x_t|y_s, y_t)q(y_s, y_t)}{q(x_s|y_s)q(x_t|y_t)q(y_s)q(y_t)} \tag{10}$$

Figure 2: First row: original color image. Second row: the corresponding skin map output by TFOM

and

$$\phi(x_s, y_s) = q(x_s|y_s)q(y_s) \tag{11}$$

Our aim is to compute for each pixel s, the quantity $p(y_s|x_s, s \in \mathcal{T}_k)$, for p in the model above, and for k ranging from 1 to say 5. This computation can be done exactly. Moreover, it can be done efficiently using the BP algorithm. This algorithm has been discovered in different scientific communities. It is called BP in A.I., Viterbi algorithm in the special case of line graphs and dynamic programming in combinatorial optimization. See [12] and the references therein for a detailed account.

The BP algorithm consists in computing k times :

$$m_{ts}(y_s) \leftarrow \sum_{y_t} \phi(x_t, y_t)\psi(x_s, x_t, y_s, y_t) \prod_{u \in \mathcal{V}(t), u \neq s} m_{ut}(y_t) \tag{12}$$

where m_{ts} are interpreted as a message coming from t to s and are initialized with the value 1. We then obtain, for $y_s = 0, 1$:

$$p(y_s|x_s, s \in \mathcal{T}_k^s) \approx \phi(x_s, y_s) \prod_{t \in \mathcal{V}(s)} m_{ts}(y_s) \tag{13}$$

2.5 Performance of the TFOM

The output of skin detection is a map indicating the probabilities of skin on pixels. Through normalization we get a grayscale image on the same grid as the input image and with the gray levels proportional to the skin probabilities. It is called *skin map* in this paper. Some skin detection results of TFOM are shown in Fig. 2. Most of the skin regions are detected correctly. However, there are also some false alarms when some objects with skin-like color appear in the background. For example, parts of the red-brick wall are marked as skin in the third image. When skin tone is changed by strong light, we could miss some skin regions. For example, our skin detector misses partly the right hand of the left man in the second image and the left arm of the right woman in the third image.

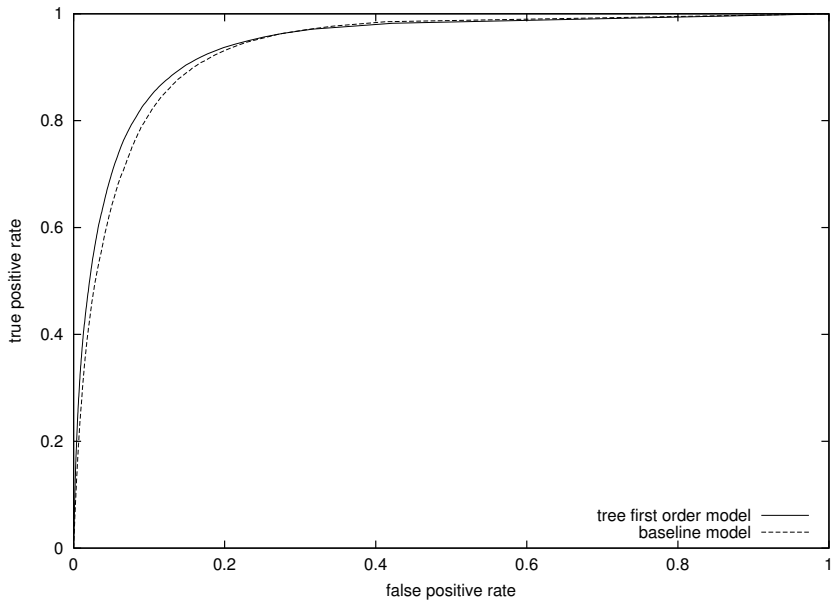

Figure 3: Receiver Operating Characteristics (ROC) curves for the Tree First Order Model (TFOM) and the baseline model

The TFOM model outperforms the baseline model implemented in [6] as shown in our previous work [19]. Figure 3 compares TFOM and the baseline model with ROC curves in the context of skin pixel detection rate and false positive rate. The results are computed over the Compaq database containing about $18,696$ photographs. It is split into two almost equal parts randomly. The first part, containing nearly 2 billion pixels is used as the training set while the other one, the test set, is left aside for ROC curve computation.

3 Adult Image Detection

3.1 Feature Extraction

There are propositions for high-level features based on grouping of skin regions[5] that might distinguish adult images from those not, but here we have a requirement to process the images speedily so, along with [6][4], we are interested to try simpler features.

Since skin distribution is of the paramount importance for the detection of adult images[6][7], all our current features are based on the skin map. For the sake of practicality, the features should be simple and easy to calculate.

We first binarize the skin map by simple thresholding. We then implement morphological open/close operations to remove noise and connect broken regions. Small skin regions are considered insignificant and discarded. Many of our features are based on the fit ellipses[21] calculated on the skin map, since they could meet our requirement for simplicity and capture some important shape information. We observed from experiments that for approaches based on skin detection, portraits have a tendence to be detected as adult images since generally portraits expose plenty of skin as adult ones. The fit ellipses will hopefully at least help discriminate portraits from adult images. We will calculate two fit ellispes for each skin map—the Global Fit Ellipse (GFE) and the Local Fit Ellipse (LFE). The GFE is computed on the whole skin map, while the LFE only on the largest skin region in the skin map. The GFE and LFE capture most of the skin distributions in the whole image and in the largest skin region respectively. Figure 4 shows the GFE as well as the LFE for a skin map.

With the skin map, we extract 9 features from the input image. The first 3 features are global:

- the average skin probability of the whole image

Figure 4: First: the original input image. Second: the Global Fit Ellipse (GFE) on the skin map. Third: the Local Fit Ellipse (LFE) on the skin map

- the average skin probability inside the GFE
- number of skin regions in the image

The other 6 features are computed on the largest skin region of the input image.

- distance from the centroid of the largest skin region to the center of the image
- angle of the major axis of the LFE from the horizontal axis
- ratio of the minor axis to the major axis of the LFE
- ratio of the area of the LFE to that of the image
- average skin probability inside the LFE
- average skin probability outside the LFE

All these features compose a simple feature vector. No effort was done to find the correlation between features.

3.2 Pattern Recognition

The feature extraction steps described in the previous subsection produce a feature vector for each image. The task is then to find the decision rule on this feature vector that optimally separates adult images from those not.

Evidence from [7] shows that the MLP classifier offers a statistically significant performance over several other approaches such as the generalized linear model, the k-nearest neighbor classifier and the support vector machine.

The semilinear feedforward net as reported by Rumelhart, Hinton, and Williams [24] has been found to be an effective system for learning discriminants for patterns from a body of examples.

We denote the input layer as i, the hidden layer as j and the output layer as k. The learning procedure starts off with a random set of weight values w_{ji}, which denote the weights of the net input from the ith layer to the jth layer. One of the training-set patterns p is used as input to evaluate the output(s) o_{pk} in a feedforward manner. The errors at the output(s) E_p generally will be quite large, which necessitates changes $\Delta_p w_{ji}$ in the weights. Using the backpropagation procedure, the net calculates $\Delta_p w_{ji}$ for all the w_{ji} in the net for the pattern p. This procedure is repeated for all the patterns in the training set to yield the resulting Δw_{ji} for all the weights for that one presentation. The corrections to the weights are made and the output(s) are again evaluated in a feedforward manner. Discrepancies between actual and target output values again result in evaluation of weight changes. After complete presentation of all patterns in the training set, a new set of weights is obtained and new output(s) are again evaluated in a feedforward manner. In a successful learning exercise, the system error

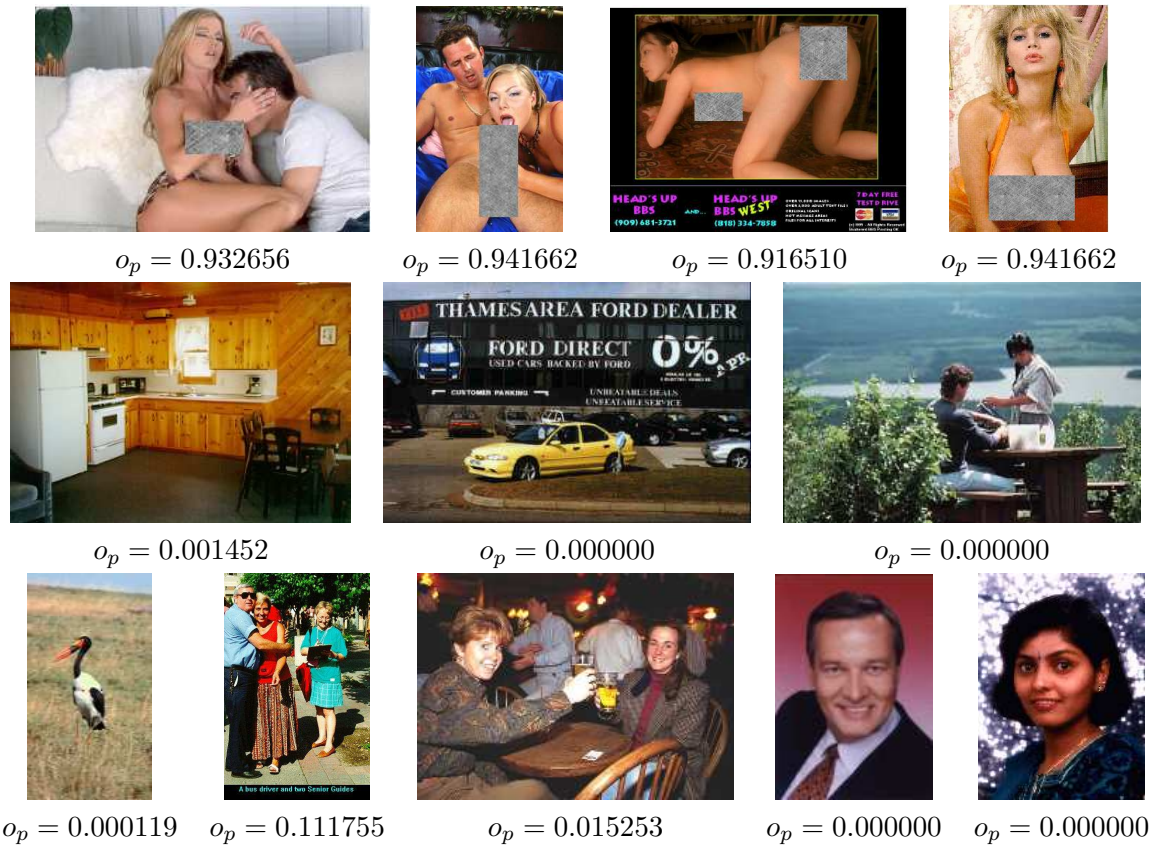

Figure 5: First row: Experimental results on adult images. Second and third rows: Experimental results on non adult images. Below the images are the associated outputs of the MLP

will decrease with the number of iterations, and the procedure will converge to a stable set of weights, which will exhibit only small fluctuations in value as further learning is attempted[22].

This net outputs a number between 0 and 1. The nearer the number is to 1, the more possibly the input pattern corresponds to an adult image. We then set a threshold T, $0 < T < 1$, to get the binary decision. In the test phase, the net takes a quick decision on the input pattern in one pass.

4 Experiments

All experiments are made using the following protocol. The database contains $10,168$ photographs, which are imported from the Compaq database[6] and the Poesia database[23]. It is split into two equal parts randomly, with $1,297$ adult photographs and $3,787$ other photographs in each part. Then these two parts are used as the training set and the test set respectively. In Fig. 5 we show some examples of the results for adult images and non adult images from the test set. The outputs of the MLP are shown just below the corresponding images.

There are some cases where our detector does not work well. In Fig. 6 several such examples are presented. The first adult image is not detected since the skin appears almost white due to over-exposure. We see that most of the skin is not detected on the skin map. The second adult image contains two connected big frames. The LFE of this image will then be very big, and the average skin probability inside this LFE will be very small. The third image is benign, but it is detected adult since the toy dog takes a skin-like color and the average skin probabilities inside the GFE and the LFE are very high. The fourth image is a portrait but decided adult since it exposes a lot of skin and even the hair and the clothes take skin-like colors. We believe skin

$o_p = 0.006828$ $o_p = 0.000005$ $o_p = 0.899044$ $o_p = 0.938251$

Figure 6: First row: original images. Second row: the corresponding skin maps. Below the skin maps are the corresponding outputs of the MLP. The first two columns are adult images with low outputs, while the other two columns are benign images with high outputs

detection based solely on color information cannot do much more, so maybe some other sorts of information is needed to improve the adult image detection performance. For example, some kind of face detector could be implemented to improve the results. However, generally adult images in webpages tend to appear together, and are surrounded by text, which could be an important clue for the adult content detector.

By varying the threshold T, a ROC curve is achieved as shown in Fig. 7. The elapse time is about 1.51×10^{-5}second/pixel, i.e., about 1 second for a 256×256 image.

5 Summary and Conclusions

This work is aimed at filtering adult images that appear in Internet. The first step of our approach is skin detection. Maximum entropy modeling is used to evaluate the skinness of pixels in the input image. We build a First Order Model that introduces constraints on color gradients of neighboring pixels. We then use Bethe tree approximation to eradicate parameter estimation. It gives us a simple analytical expression of the maximum entropy model, which is then called TFOM for Tree First Order Model in this paper. The Belief Propagation algorithm could be further implemented to accelerate the process. We show by the Receiver Operating Characteristics (ROC) curves that our skin detection improves the performance over previous work [6] in the context of skin pixel detection rate and false positive rate.

The output of skin detection is a grayscale skin map. We compute a sequence of 9 features from this skin map which form a feature vector. We use the fit ellipses to catch the characteristics of skin distribution. Two ellipses are used for each skin map—the Global Fit Ellipse (GFE) and the Local Fit Ellipse (LFE). A multilayer perceptron classifier is trained for these features. It is a semilinear feedforward net with backpropagation of error.

Many experimental results are presented including photographs and a ROC curve calculated over a test set of $5,084$ photographs, which show stimulating performance for such simple features. To improve the results one can use a face detector. However, in general adult images tend to appear together and are surrounded by text in webpages. The analysis of text is known to improve the performance of adult webpage detection[6][23].

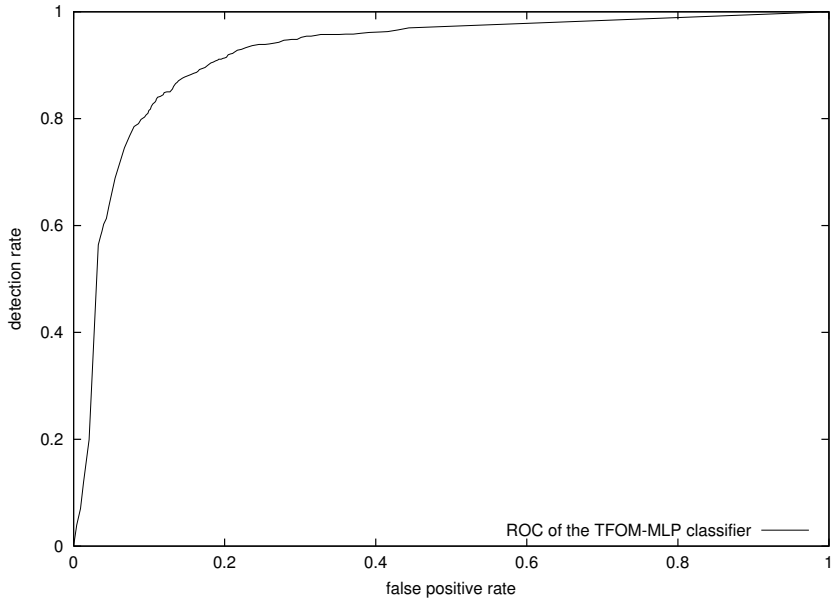

Figure 7: ROC curve of the TFOM-MLP adult image detector

Appendix: MRF on tree graphs

Let S be finite set of vertices. A neighboring system $\mathcal{V}$ is a collection of subsets of S indexed by a vertex : $\mathcal{V} = \{V_s \subset S, s \in S\}$ such that

- a) $\forall s \in S, s \notin V_s$
- b) $\forall s, t \in S, s \in V_t \Rightarrow t \in V_s$

One can then check that $(S, \mathcal{V})$ is a non oriented graph.

Let $X = (X_s)_{s \in S}; X_s \in \{1, .., K\}$ be a stochastic process defined over a set S and taking a finite number of values. It is a Markov Random Field (MRF) with respect to (S,V) if and only if $\forall x \in \{1, .., K\}^S$, $p(X = x) > 0$ and

$$\forall s \in S, p(X_s = x_s | X_t = s_t; t \neq s) = p(X_s = x_s | X_t = x_t; t \in V_s) \tag{14}$$

In words, (14) says that in order to guess the value of X at s, it is equivalent to know the values of the entire field except at s or to know the values at the neighboring locations of s.

Of particular interest are the so-called pairwise MRF. These are the simplest and most used in computer vision. Their distribution is given by

$$\pi(x) = p(X_s = x_s; s \in S) = \frac{1}{Z} \prod_{<s,t>} \psi_{st}(x_s, x_t) \prod_{s \in S} \phi_s(x_s) \tag{15}$$

where the product over $< s, t >$ means the product over all the couples of mutual neighbor vertices and Z is a normalizing constant.

Now, in the very special case where the underlying graph is loop free, for example a tree, then the functions (also called potentials) ψ_{st} and ϕ_s can be expressed as simple functions of the marginals of $\pi(.)$ and moreover when expressed in such a way, the normalizing constant Z equals to one. Specifically,

$$\pi(x) = \prod_{<s,t>} \frac{p_{st}(x_s, x_t)}{p_s(x_s) p_t(x_t)} \prod_{s \in S} p_s(x_s) \tag{16}$$

where $p_s(x_s)$ is the marginal distribution of $\pi(.)$ for vertex s and $p_{st}(x_s, x_t)$ is the marginal distribution of $\pi(.)$ for neighbor vertices s and t. That is

$$p_s(x_s) = \sum_{x_t; t \in S; t \neq s} \pi(x) \quad (17)$$

$$p_{st}(x_s, x_t) = \sum_{x_u; u \in S; u \neq s,t} \pi(x) \quad (18)$$

$$1 = \sum_{x_u; u \in S} \pi(x) \quad (19)$$

A proof can be found in [25].

References

[1] Jean-Christophe Terrillon and M. N. Shirazi and 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", *Fourth International Conference On Automatic Face and gesture Recognition*, 54–61,2000.

[2] V. Vezhnevets and V. Sazonov and A. Andreeva, "A survey on pixel-based skin color detection techniques", *Graphicon2003, 13th International Conference on the Computer Graphics and Vision*, 2003,Moscow, Russia,September.

[3] L.M. Bergasa and M.Mazo. and A. Gardel and M.A. Sotelo and L. Boquete, "Unsupervised and adapative Gaussian skin-color model", *Image and Vision Computing*, 2000,18,987-1003.

[4] J.Z. Wang, J. Li, G. Wiederhold, O. Firschein, "System for Screening Objectionable Images", *Computer Communications* (21)15:1355–1360, 1998.

[5] M.M. Fleck, D.A. Forsyth, C. Bregler: "Finding naked people", *Proc. European Conf. on Computer Vision*, B. Buxton, R. Cipolla, Springer-Verlag, Berlin, Germany, 2:593–602, 1996.

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

[7] A. Bosson, G.C. Cawley, Y. Chian, R. Harvey, "Non-retrieval: blocking pornographic images", *Proc. Intl. Conf. on the Challenge of Image and Video Retrieval, Lecture Notes in Computer Science*, Springer-Verlag, London, 2383:50–60, 2002.

[8] E. Jaynes, *Probability Theory: The Logic of Science*, http://omega.albany.edu:8008/JaynesBook

[9] A. Berger, S.D. Pietra, V.D. Pietra, "A maximum entropy approach to natural language processing", *Computational Linguistics*, 22:39–71, 1996.

[10] S.C. Zhu, Y. Wu, D. Mumford, "Filters, Random Fields and Maximum Entropy (FRAME): towards a unified theory for texture modeling", *International Journal of Computer Vision*, 27:107–126, 1998.

[11] C. Wu, P.C. Doerschuk, "Tree Approximations to Markov Random Fields", *IEEE Transactions on PAMI*, 17:391–402, April, 1995.

[12] J.S. Yedida, W.T. Freeman, Y. Weiss, "Understanding Belief Propagation and its Generalisations", *Technical Report TR-2001-22*, Mitsubishi Research Laboratories, January, 2002.

[13] D. Geman, B. Jedynak, "An Active Testing Model for Tracking Roads in Satellite Images", *IEEE Trans. on PAMI*, 18(1):1–14, January, 1996.

[14] J. Besag, "On the Statistical Analysis of Dirty Pictures", *Journal of the Royal Statistical Society, Series B*, 48(3):259–302, 1986.

[15] F. Divino, A. Frigessi, "Penalized pseudolikelihood inference in spatial interaction models with covariates", *Scandinavian Journal of Statistics*, 27(3):445-458, 2000.

[16] J. Zang, "The Mean Field Theory in EM Procedure for Markov Random Fields", *IEEE Transactions on Signal Processing*, 40(10):2570–2583, October, 1992.

[17] G. Celeux, F. Forbes, N. Peyrard, "EM Procedures Using Mean Field-Like Approximations for Markov Model-Based Image Segmentation", *Pattern Recognition*, 36(1):131–144, 2003.

[18] L. Younes, "Estimation and annealing for Gibbsian fields", *Annales de l'Institut Henry Poincaré, Section B, Calcul des Probabilités et Statistique*, 24:269–294, 1998.

[19] B. Jedynak, H. Zheng, M. Daoudi, "Statistical Models for Skin Detection", *IEEE Workshop on Statistical Analysis in Computer Vision*, in conjunction with CVPR 2003 Madison, Wisconsin, June 16–22, 2003.

[20] B. Jedynak and H. Zheng and M. Daoudi and D. Barret, "Maximum Entropy Models for Skin Detection", *publication IRMA*, Université des Sciences et Technologies de Lille, France, 2002, Volume 57, number XIII.

[21] R.M. Haralick, L.G. Shapiro, *Computer and Robot Vision*, 1:639–658, 1992.

[22] Y.-H. Pao, *Adaptive Pattern Recognition and Neural Networks*, Reading, Addison-Wesley, Massachusetts, 121–129, 1989.

[23] B. Starynkevitch, M. Daoudi et al., *POESIA Software Architecture Definition Document*, http://www.poesia-filter.org/pdf/Deliverable_3_1.pdf, Deliverable 3.1:7–9, December, 2002.

[24] D.E. Rumelhart, G.E. Hinton and R.J. Williams, "Learning internal representations by error propagation", In D.E. Rumelhart and J.L. McClelland (Eds.), *Parallel Distributed Processing: Explorations in the Microstructures of Cognition*, MIT Press, Cambridge, MA., 1:318–362, 1986

[25] J. Pearl, *Probabilistic Reasoning in intelligent systems: networks of plausible inference*, Morgan Kaufmann, 1988.

[26] D. Brown, I. Craw, J. Lewthwaite, "A som based approach to skin detection with application in real time systems" *In Proc. of the British Machine Vision Conference*, 2001.

[27] L. Sigal, S. Sclaroff and V. Athitsos, "Skin Color-Based Video Segmentation under Time-Varying Illumination" *IEEE Trans. on PAMI, 26(7):862-877, July, 2004*

[28] T. Darrell, G.G. Gordon, M. Harville and J. Woodfill, "Integrated Person Tracking Using Stereo, Color, and Pattern Detection" *In Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 601-607, 2001*

[29] W. Hafner and O. Munkelt, "Using Color for Detecting Persons in Image Sequences" *Pattern Recognition and Image Analysis, Vol. 7, No. 1, pp. 47-52, 1997*

[30] S.T. Birchfield, "Elliptical Head Tracking Using Intensity Gradients and Color Histograms" *In Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 232-237, 1998*

[31] K. Schwerdt and J.L. Crowly, "Robust Face Tracking System for Sign Language Recognition" *In Proc. Int'l Conf. Automatic Face and Gesture Recognition, pp. 90-95, 2000*