Markov Face Models

Sarat C. Dass

Department of Statistics & Probability
Michigan State University
E. Lansing, MI 48224

Anil K. Jain

Department of Computer Science & Engineering
Michigan State University
E. Lansing, MI 48224

Abstract

The spatial distribution of gray level intensities in an image can be naturally modeled using Markov Random Field (MRF) models. We develop and investigate the performance of face detection algorithms derived from MRF considerations. For enhanced detection, the MRF models are defined for every permutation of site indices in the image. We find the optimal permutation that provides maximum discriminatory power to identify faces from nonfaces. The methodology presented here is a generalization of the face detection algorithm in [1, 2] where a most discriminating Markov chain model was used. The MRF models successfully detect faces in a number of test images in real time.

Key words and phrases: Markov Random Fields, face detection, maximum pseudolikelihood estimation, simulated annealing.

1. Introduction

The use of Markov Random Fields (MRFs) to model spatial processes on lattices has been popular and widespread. By using MRF models, one is able to model the behaviour of spatial processes locally via conditional distributions of attributes. In this paper, we use certain MRFs as models for the gray level intensities of facial images. Faces typically correspond to changes in gray level intensities along some spatial direction or at some special sites in the image. Our interest here is to determine whether MRFs capture these local changes in intensities for typical face images. There have been numerous attempts to detect faces in images using different techniques such as neural networks ([3, 4]), tree classifiers ([5]), distance from prototype criteria ([6]) and Markov Chains ([1, 2]). Although Markov Chains use some notion of pixel dependence, this dependence is only allowed in one direction in space. For this reason, we feel that MRFs will be viable models for face detection since dependence can be captured along several spatial directions for different sites in the image. MRFs have also been successfully used for texture modeling, see [7], for example. Since facial images can be viewed as a type of texture in some sense, this is another reason we feel that MRFs will also be good models for facial images.

The MRF models used here do not utilize high level feature extraction for the purpose of face detection. Indeed, our aim here is to provide an initial low-level detection algorithm. In the post processing stage, algorithms based on facial features can be utilized to finally decide if a face is indeed present in the test image. For this reason, we put greater emphasis in developing algorithms with low face reject rates in the detection framework.

In order to achieve better detection rates, we seek an optimal permutation of sites in the image for which the MRF model has the best fit. In other words, for detection between faces and nonfaces, it can turn out that a permutation of the sites in the image has better discriminatory power to distinguish between a face and a nonface compared to the original (unpermuted) sites. We call the resulting MRF the most discriminating MRF for detecting faces. Thus, the most discriminating MRF approach is a generalization of the most discriminating Markov Chain approach of [1, 2].

It is essential that low level detection algorithms be computationally efficient. For the most discriminating Markov Chain approach, this is definitely the case since sites can be updated sequentially utilizing the Markovian structure. It is well known that the normalizing constant arising from MRF models causes great difficulty in computations and may actually compromise the efficiency of the algorithm. For this reason, we avoid likelihoods resulting from MRF models. Instead, we use pseudolikelihoods and pseudolikelihood ratios for estimating model parameters and for subsequent detection. The resulting reduction in computational time and complexity is significant.

The remainder of this paper is organized as follows. In Section 2, we present the basic MRF models that we use. In Section 3, we discuss the procedures to train and cross validate the MRF models. We present MRFs defined via a permutation of sites in Section 4. The Chi-square criteria to find an optimal permutation of sites for face detection is given. We also present cross validation results of our detection algorithm in Section 4. Finally, the performance of our

detection algorithm on real images is illustrated in Section 5.

2. Markov Random Field Models For Face Detection

Let $S = \{1, 2, \cdots, N\}$ denote the collection of all sites in a $R \times C$ image, where $N = RC$. For each site s in S, let x_s denote the gray level intensity at that site which is an integer between 0 and $L-1$, both inclusive, and where L is the number of gray levels. We will assume that the spatial distribution of gray level intensities, $\mathcal{X} = \{x_s, s \in S\}$ on S follows a Markov Random Field (MRF) model. For any MRF, there is an associated neighborhood system $\mathcal{N} = \{N_s, s \in S\}$, where N_s denotes the neighbors of site s. We consider only the first order neighborhood structure for the MRF models (see Figure 1).

Markovian models are, in general, parameterized by a certain number coefficients which govern the degree of spatial correlation between sites. These coefficients are unknown in typical applications and have to be estimated from training samples. It is well known that there is a conflict between the number of parameters estimated and reliability of the overall model fit to the data. We present two classes of MRFs for faces in this paper, with the number of parameters in the two classes being 3 and 234, respectively. We investigate the overall fit of these two classes of models for faces and non-faces in the training samples. Finally, the trained models are used to detect faces in test images.

2.1. Model I

The first class of MRF models that we consider is a variation of the auto-model (see [8]) with the local characteristics (conditional distributions) at site s given by

$$p(x_s \mid x_{-s}) = \frac{\exp\left\{\alpha x_s + \sum_{d=\{h,v\}} \beta_d \sum_{s \overset{d}{\sim} t} x_s\, x_t\right\}}{\sum_{x=0}^{L-1} \exp\left\{\alpha x + \sum_{d=\{h,v\}} \beta_d \sum_{s \overset{d}{\sim} t} x\, x_t\right\}}, \tag{1}$$

where the sum $s \overset{d}{\sim} t$ is taken over all neighbors of s in the direction $d, d = \{h, v\}$, x_{-s} denotes the gray level intensities of all sites except site s, and $\beta_d, d = \{h, v\}$ represent the strength of association along the horizontal and vertical directions, respectively. Figure 1 shows x_s in relation to its neighbors. The local characteristics specified in (1) uniquely determine a joint distribution on S, via Brooks' expansion (see [9]), given by

	x_{sn}	
x_{sw}	x_s	x_{se}
	x_{ss}	

Figure 1: First order neighbors of site s and corresponding gray level intensities

$$p(\underline{x}) = \frac{\exp\left\{\alpha T_{overall} + \sum_d \beta_d T_d\right\}}{\sum_{x_1=0}^{L-1} \sum_{x_2=0}^{L-1} \cdots \sum_{x_N=0}^{L-1} \exp\left\{\alpha T_{overall} + \sum_d \beta_d T_d\right\}} \tag{2}$$

where

$$T_{overall} = \sum_s x_s \quad \text{and} \quad T_d = \sum_s \sum_{s \overset{d}{\sim} t} x_s x_t$$

represent the overall average gray level intensity in the image and the overall product moment of neighboring gray level intensities in the direction d, respectively. Henceforth, the class of models given by the conditionals in (1) and having joint distribution (2) on S will be called the Model I class of MRF models. Model I captures overall changes in gray level values in 2 spatial directions (via β_h and β_v), and the overall gray level image intensity (via α). It is well known that the normalizing constant in the denominator of (2) is difficult to handle when estimating the parameters from data. For this reason, we will use pseudolikelihoods (pseudolikelihood ratios), instead of likelihoods (likelihood ratios), for the face detection problem. The pseudolikelihood (PL) for Model I is the product of local characteristics and is given by

$$PL(\text{Model I}) = \prod_{s=1}^{N} \frac{\exp\left\{\alpha x_s + \sum_d \beta_d \sum_{s \overset{d}{\sim} t} x_s x_t\right\}}{\sum_{x=0}^{L-1} \exp\left\{\alpha x + \sum_d \beta_d \sum_{s \overset{d}{\sim} t} x x_t\right\}} \tag{3}$$

2.2. Model II

The second class of MRF models that we consider is motivated by the autobinomial MRF model (see [8]) with site parameters $(\alpha_s, \beta_{st}, s, t \in S)$, specified by the local characteristics at each site s

$$p(x_s \mid x_{-s}) = \text{Bin}(L-1, \mu_s(\cdot \mid x_{-s}, \alpha_s, \beta_{st}, t \in N_s)), \tag{4}$$

where

$$\mu_s(1|x_{-s},\alpha_s,\beta_{st}) = \frac{\exp\left\{\alpha_s + \sum_{t\in N_s}\beta_{st}\,x_t\right\}}{\sum_{x=\{0,1\}}\exp\left\{\alpha_s x + \sum_{t\in N_s}\beta_{st}\,x_t\right\}} \tag{5}$$

and

$$\mu_s(0|x_{-s},\alpha_s,\beta_{st},t\in N_s) = 1-\mu_s(1|x_{-s},\alpha_s,\beta_{st},t\in N_s).$$

It follows again from Brooks' expansion that the conditional distributions specified in (4) uniquely determine a joint distribution on S (provided $\beta_{st}=\beta_{ts}$) given by

$$p(\underline{x}) = \frac{\exp\left\{\sum_s \alpha_s x_s + \sum_{s\sim t}\beta_{st}x_s x_t\right\}}{\sum_{x_1}\sum_{x_2}\cdots\sum_{x_N}\exp\left\{\sum_s \alpha_s x_s + \sum_{s\sim t}\beta_{st}x_s\,x_t\right\}}, \tag{6}$$

where $s\sim t$ stands for all pairs of sites s and t that are neighbors in S. The form of the pseudolikelihood (PL) for the autobinomial MRF model is given by

$$PL(\text{Autobinomial}) = \prod_{s=1}^{N}\frac{\exp\left\{\alpha_s x_s + \sum_{t\in N_s}\beta_{st}\,x_s x_t\right\}}{\sum_{x=0}^{L-1}\exp\left\{\alpha_s x + \sum_{t\in N_s}\beta_{st}x x_t\right\}}. \tag{7}$$

Our main reason for considering the autobinomial MRF in (6) is to determine if the information present in special sites (for example, the location of eyes, nose and facial outline) are actually used by the MRF when distinguishing between a face and nonface. The importance of a site can be determined by relative magnitudes of the site coefficients, (α_s and β_{st}, $t\in N_s, s\in S$) in a typical face and nonface image. However, estimating the coefficients in (6) is computationally challenging because of difficulty in handling the normalizing constant. This problem is not alleviated when the pseudolikelihood in (7) is used since each parameter β_{st}, with $\beta_{st}=\beta_{ts}$, occurs in the conditional specifications of more than one site. Thus, we make the following simplification while retaining the ability to measure the importance of special sites. For each site s, we consider a parameter β_s that measures the overall importance of $\{\beta_{st}, t\in N_s\}$ in a face image. Thus, instead of (7), we consider the following approximation for the pseudolikelihood

$$PL(\text{Model II}) = \prod_{s=1}^{N}\frac{\exp\{\alpha_s U_s + \beta_s V_s\}}{\sum_{x=0}^{L-1}\exp\left\{\alpha_s x + \beta_s\sum_{t\in N_s}x x_t\right\}}. \tag{8}$$

where

$$U_s = x_s \quad \text{and} \quad V_s = \sum_{t\in N_s} x_s x_t$$

represent the gray level intensity of pixel s and the joint moment of neighboring gray level intensities, respectively. The pseudolikelihood in (8) is obtained by taking $\beta_{st}=\beta_s$ for $t\in N_s$ in (7). For each site s, β_s measures the "average" correlation of x_s with its neighbors. Thus, the approximate MRF model can assess the relative importance of site s via α_s and β_s in discriminating between a face and a nonface. The parameters in (8) can be maximized separately for each site s which entails great reduction in computational complexity. This is not available for (7). Henceforth, the class of models in (8) will be referred to as Model II.

3. Training the MRF Models and Cross Validation Results

The MRF models given in Section 2 are trained using a database of faces and nonfaces. Face examples are generated by extracting gray level values from a 20×15 window (which contains the central part of the human face in the case of positive examples). Each gray level value in the image is stored as one byte, and hence the 16 ($L=16$) possible values of gray levels can vary from $0-15$. The nonface examples are generated from images that resemble a face but are not actually so. The training database consists of 7,200 and 8,422 images of faces and nonfaces, respectively. Figures 2 and 3 each give 6 examples of face and nonface images in the training database. We fit each class of models (I and II) for faces and nonfaces training samples. We estimate the unknown parameters in each model by the Maximum Pseudolikelihood (MPL) method, that is, by maximizing the pseudolikelihoods given in (3) and (8), with respect to the unknown parameters.

3.1. Detection Algorithm

This is the next step once the parameters have been estimated using the training data set. We classify a test image as a face if

$$\sum_{s=1}^{N}\log\left(\frac{\hat{p}_{face}(x_s\mid x_{-s})}{\hat{p}_{nonface}(x_s\mid x_{-s})}\right) > 0. \tag{9}$$

Otherwise, the test image will be classified as a nonface. In (9), $\hat{p}(x_s\mid x_{-s})$ stands for the estimated value of the local characteristics at site s after the parameters have been estimated. The criteria stated in (9) is in terms of the sum of logarithms of pseudolikelihood ratios for faces and nonfaces, and will be called the log pseudolikehood (LPL) criteria.

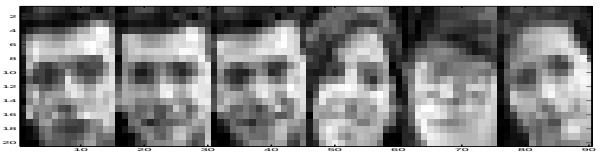


Figure 2: Examples of faces in the training data (20×15 images with 16 gray levels).

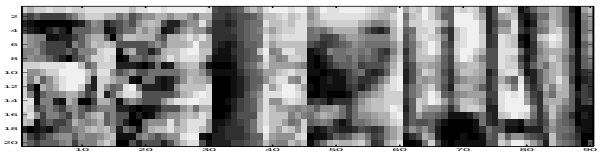


Figure 3: Examples of nonfaces in the training data (20×15 images with 16 gray levels).

3.2. Cross Validation Results

Two types of errors can arise when using the MRF models for face detection. Type I error is made when the detection procedure fails to detect a true face whereas Type II error refers to detecting a false face. We view Type I error as the more serious of the two, since a post processing stage which detects facial features can eliminate most of the falsely detected faces. We use cross validation to obtain estimates of Type I and Type II errors for each model as follows. Both the training data set of faces and nonfaces are randomly divided into two groups, the first group for training the MRF models, and the second group for detection. Using the training images (faces and nonfaces) from the first group, the optimal permutation and the corresponding parameters of the MRF model is found. The LPL criteria for face detection is used on the remaining training face and nonface images to obtain estimates of Type I and Type II errors, respectively. The results of the cross validation procedure is given in Tables 1 and Tables 2. A measure of overlap between the two histograms (faces and nonfaces) is given by

$$D(f,g) = \int_R \left(\sqrt{f(x)} - \sqrt{g(x)}\right)^2 dx$$

for f and g being the estimates of face and nonface densities from cross validation. It can be shown that $0 \leq D(f,g) \leq 2$, with $D(f,g) = 0$ iff $f = g$, and $D(f,g) = 2$ if f and g are completely separated. Small values of D in the fourth column of Tables 1 and 2 indicate the the distributions of face and nonface are not well separated.

4. Most Discriminating MRF Models via Permutations

For better detection purposes, we investigate if the MRF models are a better fit to a *permutation* of the sites in the image, instead of the natural ordering. We consider the class of all permutations of sites 1 to N, and choose that permutation which gives maximum discriminatory power for detecting faces. One argument for considering permutations of site indices is that the joint association of x_{π_s} and x_{π_t}, for a permutation π, may be better at discriminating between faces and nonfaces compared to x_s and x_t. Thus, following the construction of joint MRF models on S using conditional specifications for Model I class, one can similarly define local characteristics for a given permutation π by

$$p(x_{\pi_s} \mid x_{\pi_{-s}}) = \frac{\exp\left\{\alpha x_{\pi_s} + \sum_d \beta_d \sum_{s \overset{d}{\sim} t} x_{\pi_s}\, x_{\pi_t}\right\}}{\sum_{x=0}^{L-1} \exp\left\{\alpha x + \sum_d \beta_d \sum_{s \overset{d}{\sim} t} x\, x_{\pi_t}\right\}} \tag{10}$$

which gives rise to the joint probability density

$$p(\underline{x}) = \frac{\exp\left\{\alpha T^\pi_{overall} + \sum_d \beta_d\, T^\pi_d\right\}}{\sum_{x_1}\sum_{x_2}\cdots\sum_{x_N} \exp\left\{\alpha T^\pi_{overall} + \sum_d \beta_d\, T^\pi_d\right\}} \tag{11}$$

where $T^\pi_{overall}$ and T^π_d given by

$$T^\pi_{overall} = \sum_s x_{\pi_s} \quad \text{and} \quad T^\pi_d = \sum_s \sum_{s \overset{d}{\sim} t} x_{\pi_s} x_{\pi_t}$$

are the counterparts of $T_{overall}$ and T_d in Section 2 for a given permutation π.

Similarly, for a given permutation π, the local characteristics of the Model II class becomes

$$p(x_{\pi_s} \mid x_{\pi_{-s}}) = \text{Bin}(L-1, \mu_s(\cdot \mid x_{\pi_{-s}}, \alpha_s, \beta_{st}) \tag{12}$$

where

$$\mu_s(1 \mid x_{\pi_{-s}}, \alpha_s, \beta_{st}) = \frac{\exp\left\{\alpha_s + \sum_{t \in N_s} \beta_{st} x_{\pi_s}\, x_{\pi_t}\right\}}{\sum_{x=\{0,1\}} \exp\left\{\alpha_s x + \sum_{t \in N_s} \beta_{st}\, x x_{\pi_t}\right\}} \tag{13}$$

and

$$\mu_s(0 \mid x_{\pi_{-s}}, \alpha_s, \beta_{st}) = 1 - \mu_s(1 \mid x_{\pi_{-s}}, \alpha_s, \beta_{st}).$$

The joint MRF model specified by the local characteristics

Table 1: Crossvalidation results for Model I (natural order)

Run No.	Type I Error	Type II Error	D
1	0.3817	0.4483	0.29
2	0.3433	0.4883	0.27
3	0.3400	0.4900	0.26
4	0.3567	0.4900	0.25
5	0.4117	0.4683	0.26

Table 2: Crossvalidation results for Model II (natural order)

Run No.	Type I Error	Type II Error	D
1	0.1587	0.1007	0.92
2	0.1553	0.1067	0.91
3	0.1753	0.0960	0.93
4	0.1573	0.0987	0.94
5	0.1420	0.1080	0.91

Table 3: Crossvalidation results for Model I (permuted)

Run No.	Type I Error	Type II Error	D
1	0.0750	0.1317	1.19
2	0.0783	0.1217	1.23
3	0.0900	0.0967	1.18
4	0.0850	0.1350	1.16
5	0.1064	0.1400	1.12

Table 4: Crossvalidation results for Model II (permuted)

Run No.	Type I Error	Type II Error	D
1	0.0920	0.0787	1.24
2	0.0947	0.0767	1.22
3	0.1027	0.0780	1.20
4	0.1060	0.0773	1.22
5	0.0907	0.0767	1.26

in (12) becomes

$$p(\underline{x}) = \frac{\exp\left\{\sum_s \alpha_s x_{\pi_s} + \sum_{s\sim t} \beta_{st} x_{\pi_s} x_{\pi_t}\right\}}{\sum_{x_1}\sum_{x_2}\cdots\sum_{x_N} \exp\left\{\sum_s \alpha_s x_{\pi_s} + \sum_{s\sim t} \beta_{st} x_{\pi_s} x_{\pi_t}\right\}} \tag{14}$$

Similarly, the approximate PL for the autobinomial model is given by

$$PL(\text{Model II}) = \prod_{s=1}^{N} \frac{\exp\{\alpha_s U_s^\pi + \beta_s V_s^\pi\}}{\sum_{x=0}^{L-1} \exp\left\{\alpha_s x + \beta_s \sum_{t\in N_s} x x_t^\pi\right\}}. \tag{15}$$

where U_s^π and V_s^π given by

$$U_s^\pi = x_{\pi_s} \quad \text{and} \quad V_s^\pi = \sum_{t\in N_s} x_{\pi_s} x_{\pi_t}$$

are the counterparts of U_s and V_s in Section 2 for a given permutation π.

4.1. Chi-square Metric for Model I

For Model I, the statistics $T^\pi_{overall}$ and $T^\pi_d, d = \{h, v\}$ are sufficient for the model parameters $(\alpha, \beta_d, d = \{h, v\})$. Also, there is a one-to-one correspondence between the parameter values $(\alpha, \beta_d, d = \{h, v\})$ and the expected values of $T^\pi_{overall}$ and $T^\pi_d, d = \{h, v\}$. Thus, if a face corresponds to the values $face = \{\alpha_f, \beta_{d,f}\}$, and a nonface corresponds to the values $nonface = \{\alpha_{nf}, \beta_{d,nf}\}$, we would want the distance between the parameter values for face and nonface to be furthest apart for maximum discrimination. Equivalently, we require that the distance between $(E_{face}(T^\pi_{overall}), E_{face}(T^\pi_d))$, to be furthest away from $(E_{nonface}(T^\pi_{overall}), E_{nonface}(T^\pi_d))$ according to some measure of distance. In [1] and [2], the Kulback-Leibler distance between two distributions was chosen in the case of Markov chains. However, since the likelihoods are difficult to handle in the case of MRFs, we resort to a different distance measure, namely, the Chi-square distance, given by

$$\chi^2(\text{ Model I }) = \frac{\{E_{face}(T^\pi_{overall}) - E_{nonface}(T^\pi_{overall})\}^2}{E_{face}(T^\pi_{overall})} + \sum_{d\in\{h,v\}} \frac{\{E_{face}(T^\pi_d) - E_{nonface}(T^\pi_d)\}^2}{E_{face}(T^\pi_d)}. \tag{16}$$

Since the quantities involved in (16) are unknown, we estimate them using the training data set. Thus, for each permutation π, we estimate $E_{face}(T^\pi_{overall})$ by the overall average gray level intensity over the face training data,

$$\hat{E}_{face}(T^\pi_{overall}) = \frac{1}{N_f}\sum_{k=1}^{N_f}\sum_{s=1}^{N} x^{(k)}_{\pi_s}$$

where the sum ranges through all images in the face training data set and N_f is the number for training face images. We estimate $E_{face}(T^\pi_d)$ by

$$\hat{E}_{face}(T^\pi_d) = \frac{1}{N_f}\sum_{k=1}^{N_f}\sum_{s}\sum_{s\overset{d}{\sim}t} x^{(k)}_{\pi_s} x^{(k)}_{\pi_t}$$

In a similar fashion, the estimates corresponding to the nonface training data set is

$$\hat{E}_{nonface}(T^{\pi}_{overall}) = \frac{1}{N_{nf}} \sum_{k=1}^{N_{nf}} \sum_{s=1}^{N} x^{(k)}_{\pi_s}$$

and

$$\hat{E}_{nonface}(T^{\pi}_{d}) = \frac{1}{N_{nf}} \sum_{k=1}^{N_{nf}} \sum_{s} \sum_{\substack{d \\ s \sim t}} x^{(k)}_{\pi_s} x^{(k)}_{\pi_t},$$

where N_{nf} is the total number of nonfaces in the training data set.

4.2. Chi-square Metric for Model II

For the approximate PL in (8), the relevant site statistics are given by U^{π}_s and V^{π}_s for each site s. We use the following Chi square criteria for discrimination

$$\chi^2(\text{ Model II }) = \sum_{s=1}^{N} \frac{\{E_{face}(U^{\pi}_s) - E_{nonface}(U^{\pi}_s)\}^2}{E_{face}(U_s)}$$

$$+ \sum_{s=1}^{N} \frac{\{E_{face}(V^{\pi}_s) - E_{nonface}(V^{\pi}_s)\}^2}{E_{face}(V^{\pi}_s)}. \qquad (17)$$

The unknown quantities in (17) are estimated from the face and nonface training data set. For every permutation π, the estimate of $E_M(U^{\pi}_s)$ is

$$\hat{E}_M(U^{\pi}_s) = \frac{1}{N_M} \sum_{k=1}^{N_M} x^{(k)}_{\pi_s}$$

for $M = \{face, nonface\}$ and $N_M = \{N_f, N_{nf}\}$ accordingly, and the estimate of $E_M(V^{\pi}_s)$ is

$$\hat{E}_M(V^{\pi}_s) = \frac{1}{N_M} \sum_{k=1}^{N_M} \sum_{t \in N_s} x^{(k)}_{\pi_s} x^{(k)}_{\pi_t}$$

for M and N_M as before.

Using the likelihood ratio as a discrimination criteria is not feasible in the case of MRF models, since the normalizing constants cannot be broken down in simpler sum components as was done in the case of the Markov chain model. Using the ratio of pseudolikelihoods is easier compared to the full likelihood but it is still computationally time consuming. Therefore, we resort to the estimated Chi-square discrimination criteria that we discussed above when using MRF models for face detection problems.

4.3. Finding the Best Permutation using the Chi-square Criteria

Since the space of all permutations is extremely large, ($\mathcal{O}(N!)$, for N sites), we resort to simulated annealing (SA) to find the best permutation according to (16) and (17) for Models I and II, respectively. The SA algorithm ([11]) is described as follows. Start with an initial permutation, π_0, and initial temperature, $T = t_0$, say. Randomly select two sites for interchange and obtain the updated permutation, π_1. For π_1, calculate the Chi-square distance between faces and nonfaces in the training set. If this distance is larger than the initial Chi-square distance for π_0, accept the new permutation, π_1. Otherwise, accept the new permutation, π_1, with probability e^{δ}, where δ is the difference in (16) (or (17)) between π_1 and π_0. The acceptance-rejection scheme is carried out for a large number of runs. Subsequently, T is reduced to, say, t_1, and the above algorithm is repeated for the temperature, t_1. The SA procedure reaches a solution that is close to the global optimal solution when T is small. The acceptance-rejection scheme for each temperature level was carried out for $n = 1000$ times. The cooling schedule was taken to be $T = T * 0.97$.

Once the best permutation was found, the parameters of the MRF for faces and nonfaces were estimated using the Maximum Pseudolikelihood (MPL) method.

4.4. Detection Algorithm

For the optimal permutation, π^{opt}, and the corresponding estimated parameters (for both the face and nonface MRF models), an image is classified as a face if

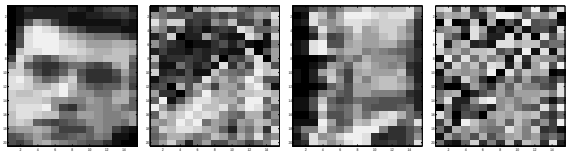
$$\sum_{s=1}^{N} \log \left(\frac{\hat{p}_{face}(x_{\pi^{opt}_s} \mid x_{\pi^{opt}_{-s}})}{\hat{p}_{nonface}(x_{\pi^{opt}_s} \mid x_{\pi^{opt}_{-s}})} \right) > 0. \qquad (18)$$

Otherwise, the test image will be classified as a nonface. In (18), $\hat{p}(x_{\pi^{opt}_s} \mid x_{\pi^{opt}_{-s}})$ stands for the estimated value of the local characteristics at site s after the optimal permutation π^{opt} has been found and the parameters have been estimated. This is again the log pseudolikelihood (LPL) criteria for the permuted sites.

4.5. Cross Validation Results

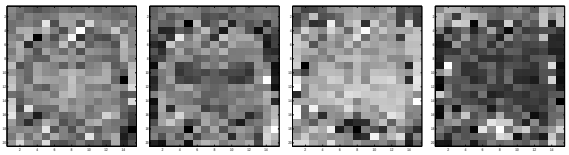
The results of the cross validation procedure for permuted sites are given in Tables 3 and Tables 4 for the permuted MRF models. The cross validation procedure is run 5 times for Model I and 5 times for Model II to ascertain the variability of both kinds of errors.

It is clear from Tables 3 and 4 that Model II has more consistent detection properties compared to Model I. The average Type I and Type II error probabilities for Model II are 9% and 7%, respectively, whereas for Model I, the range of Type I and Type II error probabilities are from 7%-11%



(a) (b) (c) (d)

Figure 4: Permutations of sites. (a) Face example, (b) Permuted face, (c) Nonface example, (d) Permuted nonface.



(a) (b) (c) (d)

Figure 5: Parameter values for faces and nonfaces. (a) $\{\alpha_s\}$ for faces, (b) $\{\beta_s\}$ for faces, (c) $\{\alpha_s\}$ for nonfaces, (d) $\{\beta_s\}$ for nonfaces.

and 9% - 14%, respectively. Model I sometimes performs better than Model II for true face images but Model I always gives more false alarms compared to Model II. We see from the fourth column entries of Tables 3 and 4 that the distributions are better separated compared to the case when the sites were not permuted.

For Model II, it is interesting to see how the optimal permutation rearranges gray level intensities in an image. Figure 4 (a) shows a typical face image from the training data base. The optimal permutation is applied to the face image and the resulting image is presented in Figure 4 (b). It is clear that the optimal permutation forms two distinct clusters of gray level intensities, one cluster of low gray level intensities while another cluster of higher gray level intensities. The relative positions of these clusters in a face image are also fixed for different face images. No such cluster forms when a nonface image is considered. See Figures 4 (c) and 4 (d), for example.

We also display the site coefficients, $\{\alpha_s\}$ and $\{\beta_s\}$, of Model II for faces and nonfaces. The image plots are obtained first by rescaling the coefficients to the $0-255$ range, and then reordering the permuted sites back to the natural order. Figures 5 (a) and (b) show the relative magnitude of the $\{\alpha_s\}$ and $\{\beta_s\}$, respectively, for a face image. Observe that the $\{\beta_s\}$ image extracts the distinguishing features of a face, namely the face outline, and the positions of the eyes and nose. Since the eyes and nose are relatively darker regions compared to the surrounding sites, $\{\beta_s\}$ at the boundaries of the eyes, nose and face outlines capture this change in gray level intensity. Since the intensities change in opposite directions (from lighter to darker, or vice versa), this is reflected in the $\{\beta_s\}$ coefficients by their low negative values.

5. Face Detection for Real Images

We apply the face detection algorithm based on (permuted) Models I and II to real images. We consider images of arbitrary sizes with gray level intensities ranging from $0-255$. These images (see Figure 7) consists of one or more faces of an arbitrary size.

First, the gray level intensities of the original image are converted to the $0-15$ range by division modulo 16. Some blocking effect in the original image is observed after performing this step (see Figure 6). In order to fit a face in these images into our 20×15 detection frame, we scale (up or down) the original image so that the faces approximately fit into the detection frame. Then, we slide a 20×15 window in a raster scan fashion over the rescaled image. The LPL values are calculated for each position of the detection window. If an LPL value is greater than 0, a face frame (red frame) is placed over the window. Several threshold values, other than 0 (in (18)), such as 5 and 10, are also considered. Possible faces correspond to high positive LPL values. Both models detect all the faces in the four test images with single and multiple faces. Some spurious faces are detected and they disappear when the threshold level is raised. In general, Model II performed better at detecting faces compared to Model I. This was also established based on cross validation results. For test images, we empirically determine a good value of the threshold. Figure 7 show the results of the detection algorithm based on Model II for some of the images. The detection algorithm was written in MATLAB and was run on a PC with a 750 Mhz Pentium III processor. The detection times (in seconds) for these images ((a),(b),(c) and (d)) are 8, 10, 67 and 300, respectively.

Original image Modulo 16 reduction Scaled Image

Figure 6: Effects of Blocking and Scaling

6. Summary and Conclusions

We have presented two Markov models for face detection. Better detection properties are obtained for a permutation of the sites, instead of the natural ordering. Model II results in smaller error probabilities of detection compared to Model I. Moreover, Model II distinguishes faces from nonfaces by

(a) (b) (c) (d)

Figure 7: Sample input images and detection results. Image sizes (a) 200×170, (b) 270×150, (c)410×450, (d) 350×550

identifying regions that form the outlines of the eyes, nose and the face. For future work, we plan to investigate multiple MRF models for faces and extend our detection algorithm to color images.

Acknowledgments

The authors would like to thank N. Duta for helpful discussions and for making the face/nonface database available to the authors.

References

[1] N. Duta, *Learning based Detection, segmentation and matching of objects*. PhD thesis, Michigan State University, 2000.

[2] A. Colmenarez and T. Huang, "Face detection with information-based maximum discrimination," *Proceedings of CVPR '97*, 1997.

[3] H. Rowley, *Neural Network-based Face Detection*. PhD thesis, Carnegie Mellon University., 1999.

[4] H. Rowley, S. Baluja, and T. Kanade, "Neural network-based face detection.," *IEEE Trans. Pattern Anal. and Machine Intelligence*, vol. 20, no. 4, pp. 1019–1031, 1997.

[5] Y. Amit, D. Geman, and B. Jedynak, "Efficient focusing and face detection," in *Face Recognition: From Theory to Applications* (H. Wechsler, ed.), pp. 143–158, NATO ASI Series F, Springer Verlag, Berlin, 1997.

[6] K. Sung and T. Poggio, "Example-based learning for view-based human face detection," *IEEE Trans. Pattern Anal. and Machine Intelligence*, vol. 20, no. 1, pp. 39–52, 1998.

[7] G. R. Cross and A. K. Jain, "Markov random field texture models," *IEEE Trans. Patt. Anal. Mach, Intell.*, vol. 5, pp. 25–39, 1983.

[8] X. Guyon, *Random Fields on a Network*. Springer-Verlag, 1995.

[9] J. P. Hobert and G. Casella, "Functional compatibility, markov chains and gibbs sampling with improper posteriors," *J. Comput. Graph. Statist.*, vol. 7, no. 1, pp. 42–60, 1998.

[10] R. Chellappa and A. Jain, eds., *Markov Random Fields. Theory and Application*. Academic Press, Inc., 1991.

[11] E. Aarts and J. Korst, *Simulated Annealing and Boltzmann Machines: a Stochastic Approach to Combinatorial Optimization and Neural Computing*. Wiley, Chichester, 1989.