Learning Inhomogeneous Gibbs Model of Faces by Minimax Entropy

Ce Liu[*]
Microsoft Research China

Song Chun Zhu
The Ohio State University

Heung-Yeung Shum
Microsoft Research China

Abstract

In this paper we propose a novel inhomogeneous Gibbs model by the minimax entropy principle, and apply it to face modeling. The maximum entropy principle generalizes the statistical properties of the observed samples and results in the Gibbs distribution, while the minimum entropy principle makes the learnt distribution close to the observed one. To capture the fine details of a face, an inhomogeneous Gibbs model is derived to learn the local statistics of facial feature points. To alleviate the high dimensionality problem of face models, we propose to learn the distribution in a subspace reduced by principal component analysis or PCA. We demonstrate that our model effectively captures important and subtle non-Gaussian face patterns and efficiently generates good face models.

1. Introduction

Many computer vision problems are concerned with human faces. While various methods have been proposed to locate, classify and recognize (*e.g.*, [12], [1], [13]) human faces, they inevitably make some underlying assumptions on the statistical models of faces.

One of the most dominant representations used for modeling faces is principle component analysis or PCA, which assumes the distribution a single Gaussian. For instance, Turk and Pentland [13] used PCA to construct eigenfaces that are then used for face recognition. Cootes *et al.* [8] established a two-layer model, and PCA is applied to model the key points on a face. In fact, there is a long history in study landmarks and distributions in statistics, which also use PCA extensively (*e.g.*, [5, 9]).

To capture more variations in the face models, a mixture of Gaussians model can be used to replace the single Gaussian model. A good example is the Active Shape Model [7]. Other deformable models have also been proposed by [2, 14] to effectively model faces using Gibbs models. However, such deformable models are manually defined without learning.

Existing face models work well for many applications, but they mostly capture the global characteristics of faces and often fail to reveal local details. Such local characteristics, however, are crucial for some tasks such as facial sketch or caricature generation. These tasks require learning some interesting face patterns that are often hidden in the non-Gaussian distributions. Although some non-Gaussian models have been proposed, they are very difficult to learn.

In fact, the mean and covariance (used in the Gaussian models) are the first-order and second-order statistics respectively. They are too simple to capture sufficient details of face models. The 1D marginal distributions or histograms, on the other hand, represent all orders and thus can be used as feature statistics. This motivates us to build an inhomogeneous statistical model of faces which can capture sophisticated local variations, by making use of 1D marginal distributions. Specifically, based on the minimax entropy principle, we build an inhomogeneous Gibbs model for faces. The forms of the potential functions (energy) are learned in a nonparametric way by a maximum entropy principle. We adopt a Markov chain Monte Carlo (MCMC) method, *i.e.*, inhomogeneous Gibbs sampling, to obtain the global optimal parameters. This makes our model sufficient to characterize the non-Gaussian properties in a distribution.

Our work is related to recent work on learning local statistics (*e.g.*, face hallucination [3], learning low-level vision using a Markov network [11, 10] and homogeneous Gibbs models such as FRAME [16, 15]). In particular, the minimax entropy principle proposed in [16] for homogeneous prior learning for texture analysis and synthesis inspired us to build inhomogeneous models for faces.

The remainder of this paper is organized as follows. After introducing the problem in Section 2, we construct an inhomogeneous Gibbs model by the minimax entropy principle in Section 3. Our method is illustrated by a toy problem in Section 4, and most importantly, applied to face modeling in Section 5. We present some discussions in Section 6 and conclude in Section 7.

[*] Visiting from Department of Automation, Tsinghua University, China. Email: lce@msrchina.research.microsoft.com

0-7695-1143-0/01 $10.00 (C) 2001 IEEE

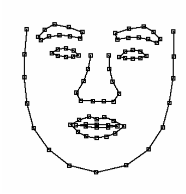

Figure 1. A face model with linked key points.

2. Problem Formulation

For each face, we define its model by a set of key points denoted by $w = \{(x_i, y_i), i = 1 : N/2\}$, and their connectivity, as shown in Figure 1.

We also collect a number of M face models, $\{w_i^{obs}, i = 1 : M\}$, which are possibly from different people, with different ages, shapes, and poses. We believe that there exists a natural but unknown distribution $f(w)$ and w_i^{obs} $i.i.d. \sim f(w)$. The objective of our work is to learn a prior distribution $p(w)$ from observed samples $\{w_i^{obs}\}$ such that $p(w)$ is close to $f(w)$. Most previous work have assumed that $p(w)$ can be a Gaussian or a mixture of Gaussions. In our work, we formulate $p(w)$ as an inhomogeneous Gibbs model.

2.1 Gaussian

The simplest model is Gaussian which is commonly used in PCA. Let $w = (x_1, y_1, x_2, y_2, \cdots, x_{N/2}, y_{N/2})^T$ be a vector of random variables. We assume

$$w = \sum_{i=1}^{L} \alpha_i \mathbf{e}_i + \mathbf{n} \qquad (1)$$

where $\mathbf{e}_i$ is i-th principal component of the covariance matrix of w, and $< \mathbf{e}_i, \mathbf{e}_k >= 0$ when $i \neq k$. L is the number of $\mathbf{e}_i$'s and $\mathbf{n}$ is Gaussian noise. Each component n_i of $\mathbf{n}$ is assumed to be independent such that $p(n_i, n_k) = p(n_i)p(n_k)$, where $p(n_i)$ is a Gaussian distribution with zero mean and variance σ_i^2 . PCA will have the distribution

$$p(w) = \frac{1}{Z} exp\{-\sum_{i=1}^{L} < w - \mu, \mathbf{e}_i >^2 / (2\sigma_i^2)\} \qquad (2)$$

where Z is the normalization factor. This model is easy to build from eigenvalues and eigenvectors of the covariance matrix, but the Gaussian and independence assumptions may not hold for modeling faces.

2.2 Mixture of Gaussians

A better model is a mixture of Gaussians *i.e.*,

$$p(w) = \sum_{i=1}^{C} \lambda_i G(w_i; \mu_i, \Sigma_i) \qquad (3)$$

where $\sum_{i=1}^{C} \lambda_i = 1$, $\lambda_i > 0$, $i = 1 : C$. C is the total number of Gaussian kernels and μ_i, Σ_i are the mean and covariance of Gaussian kernel i. Literally speaking, as $C \to \infty$, the above distribution can approximate any function. We need to estimate C, μ_i, Σ_i and λ_i, usually by an *Expectation-Maximization* (EM) [4] algorithm. But even some simple distributions are onerous to be modelled by a mixture of Gaussians and EM could only find a locally optimal solution, as will be shown in Section 4. Therefore, even a mixture of Gaussians model may not be sufficient to learn the distribution of facial key points.

2.3 Inhomogeneous Gibbs Model

Since the dimension N of w is large, it is hard to describe $f(w)$ directly. It is natural to build a model based on 1D statistics, as features of the distribution. These features are defined as $\{\phi^{(\alpha)}, \alpha = 1 : K\}$, where $\phi^{(\alpha)}(w)$ is a vector-valued function of w. Thus we can get the 1D marginal distribution

$$E_f[\phi^{(\alpha)}(w)] = \int \phi^{(\alpha)}(w) f(w) dw. \qquad (4)$$

From the observed sample set, we can get the corresponding empirical distribution (histogram):

$$\mu_{obs}^{(\alpha)} = \frac{1}{M} \sum_{i=1}^{M} \phi^{(\alpha)}(w_i) \qquad (5)$$

Then we define a distribution function set sharing the same marginal densities:

$$\Omega_f = \{p(w) | E_p[\phi^{(\alpha)}(w)] = \mu_{obs}^{(\alpha)}, \alpha = 1 : K\},$$

and regard two distributions as indiscriminating if they are in the same set. But how to select an optimal distribution $p(w)$ from this set? How to find its form and learn the parameters? How to choose the features to model the distribution? We show in next section that the optimal distribution turns out to be an inhomogeneous Gibbs one by the maximum entropy principle, the parameter can be learnt by Markov chain Monte Carlo and the feature set is gradually pursued by the minimum entropy principle.

3. Learning Inhomogeneous Gibbs Model

3.1. Maximum Entropy Principle

We apply the maximum entropy principle to learn the $p(w)$ so that the learnt model can be generalized, or present no more information than what is available [16],

$$p(w) = \text{argmax}\{-\int p(w) \log p(w) dw\} \qquad (6)$$

subject to

$$E_p[\phi^{(\alpha)}(w)] = \int \phi^{(\alpha)}(w)p(w)dw = \mu_{obs}^{(\alpha)},\ \alpha = 1:K$$

and $\int p(w)dw = 1$.

It is well known that the solution to above problem is Gibbs distribution with the form [16]:

$$p(w;\Lambda) = \frac{1}{Z(\Lambda)}\exp\{-\sum_{\alpha=1}^{K} < \lambda^{(\alpha)}, \phi^{(\alpha)}(w) >\}, \qquad (7)$$

where Λ is the parameter set and $Z(\Lambda)$ is the normalization factor. $\lambda^{(\alpha)}$ is a vector-valued parameter according to $\phi^{(\alpha)}(w)$. The exponential term of (7) is the *Gibbs potential energy* to be learnt. The The *maximum likelihood estimation* (MLE) is utilized to estimate $\hat{\Lambda}$. Let

$$L(\Lambda) = \frac{1}{M}\sum_{i=1}^{M}\log p(w_i^{obs};\Lambda) \qquad (8)$$

be the log-likelihood function. The optimization is a stochastic ascent method with the gradient of $L(\Lambda)$ [16], leading to an iterative update of $\lambda^{(\alpha)}$,

$$\frac{\partial\lambda^{(\alpha)}}{\partial t} = E_{p(w;\Lambda)}[\phi^{(\alpha)}(w)] - \mu_{obs}^{(\alpha)},\ \alpha = 1:K, \qquad (9)$$

where $E_{p(w;\Lambda)}[\phi^{(\alpha)}(w)]$ is calculated from the synthesized sample set $\{w_i^{syn}, i = 1:M'\}$:

$$E_{p(w;\Lambda)}[\phi^{(\alpha)}(w)] \approx \frac{1}{M'}\sum_{i=1}^{M'}\phi^{(\alpha)}(w_i^{syn}) = \mu_{syn}^{(\alpha)}. \qquad (10)$$

The synthesized sample set $\{w_i^{syn}\}$ is drawn from an inhomogeneous Gibbs sampling.

Unlike in the homogeneous model [16], the synthesized histogram $\mu_{syn}^{(\alpha)}$ in our inhomogeneous model must be estimated from a number of independent samples in the Markov chain, not from a Markov random field.

3.2. Minimum Entropy Principle

Since our goal is to make an inference about the underlying distribution $f(w)$, the goodness of the model can be measured by the *Kullback-Leibler*(KL) divergence from $f(w)$ to $p(w;\Lambda)$:

$$KL(f(w), p(w;\Lambda)) = \int f(w)\log\frac{f(w)}{p(w;\Lambda)}dw$$

$$\begin{aligned} &= -E_f[\log p(w;\Lambda)] + E_f[\log f(w)] \\ &= entropy(p(w;\Lambda)) - entropy(f(w)) \end{aligned}$$

Note that $entropy(f(w))$ is fixed and the entropy of $p(w;\Lambda)$ depends on the set of features $\{\phi^{(\alpha)}, \alpha = 1, 2, \cdots\}$ included in the distribution $p(w;\Lambda)$. We should find the best features to constrain $p(w;\Lambda)$ such that it has minimum entropy. We call this the *minimum entropy principle*.

Let B be the set of all possible features, and $S \subset B$ an arbitrary set of K features. Let

$$\Omega_S = \{p(w) | E_p[\phi^{(\alpha)}(w)] = E_f[\phi^{(\alpha)}(w)], \forall\phi^{(\alpha)} \in S\}$$

be the set of probability distributions which can reproduce the expected feature statistics in S. Then by the minimax entropy principle, the optimal set of features is

$$S^* = \arg\min_{|S|=K}\{\max_{p\in\Omega_S} entropy(p)\}. \qquad (11)$$

We must devise a feature pursuit method to find the above optimal feature set.

3.3. Feature Pursuit

The feature bank B in our model chooses linear unit vectors as features, which form a unit hypersphere in N-dimensional linear space. Let $S = \{\phi^{(\alpha)},\ \alpha = 1:K\}$ be the current selected feature set, and $p_S(w)$ be the maximum entropy distribution of $f(w)$. We want to find a new feature $\phi^{(\beta)} \in B$ such that $E_{p_S}[\phi^{(\beta)}(w)]$ and $E_f[\phi^{(\beta)}(w)]$ are the most different. Since $\phi^{(\beta)}$ is linear, we have $\phi^{(\beta)}(w) = w^T\phi^{(\beta)}$. The difference is measured by the KL divergence

$$KL(E_{p_S}[w^T\phi^{(\beta)}], E_f[w^T\phi^{(\beta)}])$$

$$= \int E_{p_S}[w^T\phi^{(\beta)}]\log\frac{E_{p_S}[w^T\phi^{(\beta)}]}{E_f[w^T\phi^{(\beta)}]}d(w^T\phi^{(\beta)}) \qquad (12)$$

The above integration could not be analytically calculated. But we can get the empirical distribution of $E_{p_S}[w^T\phi^{(\beta)}]$ and $E_f[w^T\phi^{(\beta)}]$ by independent samples from $p_S(w)$ and $f(w)$ respectively, using the Parzen window method with Gaussian kernels

$$\begin{aligned} h_{syn}^{\phi^{(\beta)}}(z) &= \frac{1}{M'}\sum_{i=1}^{M'}G(z - (w_i^{syn})^T\phi^{(\beta)}) \\ h_{obs}^{\phi^{(\beta)}}(z) &= \frac{1}{M}\sum_{i=1}^{M}G(z - (w_i^{obs})^T\phi^{(\beta)}) \end{aligned}$$

where $G(z-\mu)$ is a 1-D Gaussian kernel function with mean μ and a certain variance σ. The KL divergence can thus be estimated by Monte Carlo integration:

$$KL(E_{p_S}[w^T\phi^{(\beta)}], E_f[w^T\phi^{(\beta)}]) \approx KL(h_{syn}^{\phi^{(\beta)}}(z), h_{obs}^{\phi^{(\beta)}}(z))$$

$$= \int h_{syn}^{\phi^{(\beta)}}(z) \log \frac{h_{syn}^{\phi^{(\beta)}}(z)}{h_{obs}^{\phi^{(\beta)}}(z)} dz. \qquad (13)$$

Now, the feature pursuit process becomes a search for a new linear feature $\phi^{(\beta)}$ such that

$$\phi^{(\beta)} = \arg \max_{\phi^{(\beta)}} KL(h_{syn}^{\phi^{(\beta)}}(z), h_{obs}^{\phi^{(\beta)}}(z)) \qquad (14)$$

subject to $\quad ||\phi^{(\beta)}|| = 1.$

The gradient of the above objective function is

$$\frac{\partial KL(h_{syn}^{\phi^{(\beta)}}(z), h_{obs}^{\phi^{(\beta)}}(z))}{\partial \phi^{(\beta)}} = \int$$

$$\{\frac{\partial h_{syn}^{\phi^{(\beta)}}(z)}{\partial \phi^{(\beta)}}[\log \frac{h_{syn}^{\phi^{(\beta)}}(z)}{h_{obs}^{\phi^{(\beta)}}(z)} + 1] + h_{syn}^{\phi^{(\beta)}}(z) \frac{\partial h_{obs}^{\phi^{(\beta)}}(z)}{\partial \phi^{(\beta)}}\} dz$$

where

$$\frac{\partial h_{obs}^{\phi^{(\beta)}}(z)}{\partial \phi^{(\beta)}} = -\frac{1}{M\sigma^2}$$

$$\sum_{i=1}^{M} (z - (w_i^{obs})^T \phi^{(\beta)}) w_i^{obs} G(z - (w_i^{obs})^T \phi^{(\beta)})$$

and $\partial h_{syn}^{\phi^{(\beta)}}(z) / \partial \phi^{(\beta)}$ can be calculated similarly. Since the constraint of the optimization problem is a hypersphere, the gradient can be projected to this sphere. The gradient and the KL divergence are approximated by *Compound-Simpson* numeral integration. While the solution of gradient ascent is only locally optimal, satisfactory results can be obtained by randomly choosing multiple initial values.

But how many features are enough? When the optimal feature found by (14) is added to the feature set, *i.e.* $S' = \{S, \phi^{(\beta)}\}$, the entropy of new Gibbs distribution $p_{S'}(w; \Lambda)$ will decrease compared to that of $p_S(w)$. This entropy decrease is indeed the information gained by the new feature $\phi^{(\beta)}$ and can be approximately measured by the KL divergence

$$d[\phi^{(\beta)}] = KL(E_{p_S}[w^T \phi^{(\beta)}], E_f[w^T \phi^{(\beta)}]) \qquad (15)$$

On the other hand, the model complexity will increase as a new feature is selected. By the minimum description length (MDL) principle, the feature pursuit procedure stops as soon as the entropy decrease does not compensate for the increase in model complexity. In practice, if $d[\phi^{(\beta)}] < \varepsilon$, where ε is a small value, the feature pursuit can stop. The feature set is sufficient to represent the observed distribution.

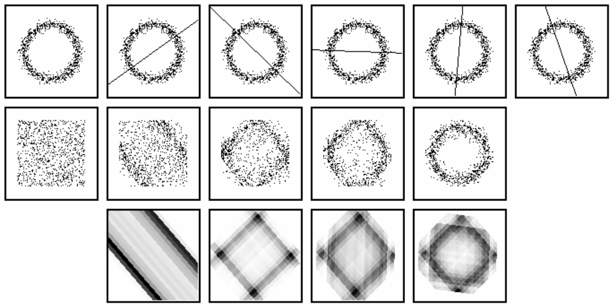

Figure 2. Feature pursuit process with a circle-like distribution.

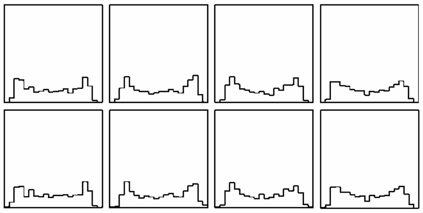

Figure 3. Histogram comparison between the observed samples (the top row) and the synthesized samples (the bottom row).

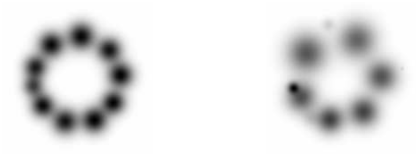

Figure 4. Synthesized samples using EM. Left: good initialization; Right: bad initialization.

4. An Example

We use a simple example to illustrate our inhomogeneous prior learning method. Suppose the distribution is a circle-like one that can be modeled as (in polar coordinates)

$$r \sim Gaussian(r_0, \sigma), \theta \sim Uniform(0, 2\pi].$$

We can easily draw a set of independent samples from this distribution, shown in row 1, column 1 (denoted as $L(1,1)$ for simplicity) of Figure 2. Before any features are selected, the synthesized samples $L(2,1)$ are uniformly drawn on the space, because the energy term of the Gibbs distribution is zero. After comparing the observed and synthesized samples, the first feature on which their KL divergence is maximized is chosen, with the observed samples shown in $L(1,2)$. Samples drawn according to this feature are then shown in $L(2,2)$ and the corresponding Gibbs energy is shown in $L(3,2)$. The darker a pixel is, the smaller Gibbs energy it carries. Due to the symmetry of circle, the position of the first feature is arbitrary.

After comparing the synthesized and observed samples again, the second feature is chosen. The second feature turns out to be a vector perpendicular to the first feature. Samples synthesized based on these two features are shown in $L(2,3)$ and the Gibbs energy in $L(3,3)$. Similarly the third feature is found, and it is right the middle of the first and second features. The fourth is again perpendicular to the third one. Because the information gain (KL divergence) by adding the fifth feature is insignificant (lower than a given threshold of 0.05), the feature pursuit process stops and the selected four features are regarded as nearly sufficient to describe the distribution. We observe that the final synthesized sample set using four features (shown in $L(2,5)$) is very similar to the observed one, *i.e.*, a circle. This is also demonstrated by the histograms for both the observed and synthesized samples with these two features, as shown in Figure 3. The whole learning process takes three minutes on a $PIII$ $667Hz$ PC with $256MB$ memory.

It is obvious that the single Gaussian method fails to deal with this problem. But it may be learnt with a mixture of Gaussians. For example, if we set the kernel number to be 10, the EM algorithm will generate some results depending on the initialization. Figure 4 shows a "good" estimation on the left and a "bad" one on the right. Even the "good" one cannot accurate describe the distribution because of insufficient number of kernals.

5. Learning the Face Model

5.1. Dealing with High Dimensions

Before we apply the inhomogeneous prior learning technique to modeling faces, we must deal with the high dimensionality problem. We use 83 feature points in our face model ($N = 166$). There are two problems why the process could be very slow.

1. The number of energy terms in the Gibbs distribution is proportional to the number of features selected. If too many features, it is time consuming to calculate the conditional density and to run the Markov chain.

2. In the first few steps of feature pursuit, the possible state space constrained by few linear features will be so huge in high dimensional space that it will take the Markov chain a long time to walk through the space.

Recall that, in feature pursuit, the feature projected in which the synthesized and observed densities have the maximum KL divergence is selected. It is obvious that the KL divergence between the delta function $\delta(x - x_0)$ and any other density function $g(x) \neq \delta(x - x_0)$ is infinite. At the beginning of feature pursuit, the feature set is null so the Gibbs distribution is in fact a uniform one in a hypercubic. Let $U(w)$ be the initial uniform distribution. We want to find a feature $\phi^{(\beta)}$ such that $KL(h_{U(w)}^{\phi^{(\beta)}}, h_{f(w)}^{\phi^{(\beta)}})$ is maximized. If

$$h_{f(w)}^{\phi^{(\beta)}} \approx \delta(z - z_0) \text{ and } h_{U(w)}^{\phi^{(\beta)}} \neq \delta(z - z_0)$$

then $\phi^{(\beta)}$ must be the optimal one to be selected. In practice the above condition can be approximated by

$$Var(h_{f(w)}^{\phi^{(\beta)}}) \ll Var(h_{U(w)}^{\phi^{(\beta)}})$$

Or simply

$$Var(h_{f(w)}^{\phi^{(\beta)}}) < \varepsilon \qquad (16)$$

where ε is a very small value, then $\phi^{(\beta)}$ is a very important feature which must be selected.

PCA is naturally employed to find the most important features at the first stage. In fact, the smallest eigenvectors (which correspond to the smallest eigenvalues) record the basic properties of the distribution, while the largest eigenvectors record the main variation. We select the smallest eigenvectors contributing to less than 3% of the total eigenvalue as primitive features. This solves problem (2) because the Markov chain now only needs to walk through the constrained space by the primitive features.

To efficiently solve the problem (1), we again reduce the number of features by PCA, and learn the distribution in a much smaller subspace. If we use the smallest eigenvectors as initial features, the marginal distributions on these features are very rigid which can be regarded as *hard constraints* that we must satisfy. This is approximately equivalent to reducing the dimensions to a subspace constructed by the largest principal components. In this way, the original problem is greatly simplified.

5.2. Inhomogeneous Gibbs Models of Faces

We now apply our method to learn the face model of key points. First we reduce the dimension (166) of the original problem using PCA and transform the key points to an 18-dimensional subspace. Some typical training faces from our 73 training examples are shown in Figure 5(d).

At the beginning, the feature set is null. All the faces are uniformly sampled in the PCA-constrained space. Some typical synthesized samples are listed in Figure 5(a). Most of them are not acceptable as human faces. Samples drawn from the learnt distribution with 4 features are shown in Figure 5(b). The results are much improved because some of drawn samples do look like faces. Figure 5(c) shows samples drawn from the distribution with 17 features. These sampled faces have versatile expressions, types and poses, suggesting that we can learn some statistical properties hidden in non-Gaussian distribution from only 73 examples. In our experiments, we have determined that 17 features

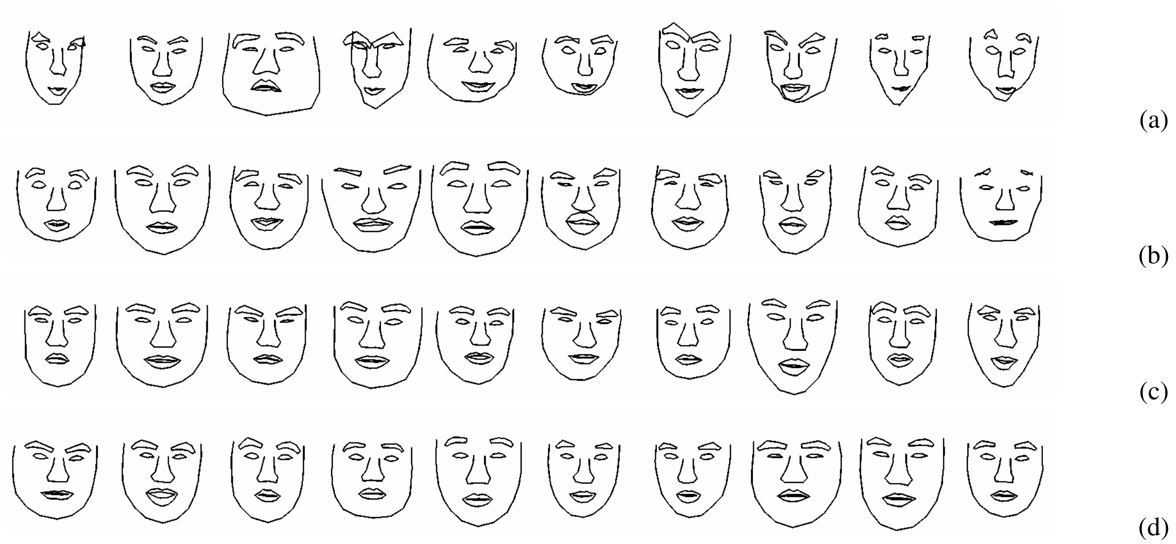

Figure 5. Synthesized face samples (a) without any features; (b) with 4 features; (c) with 17 features; and (d) observed samples from the training set.

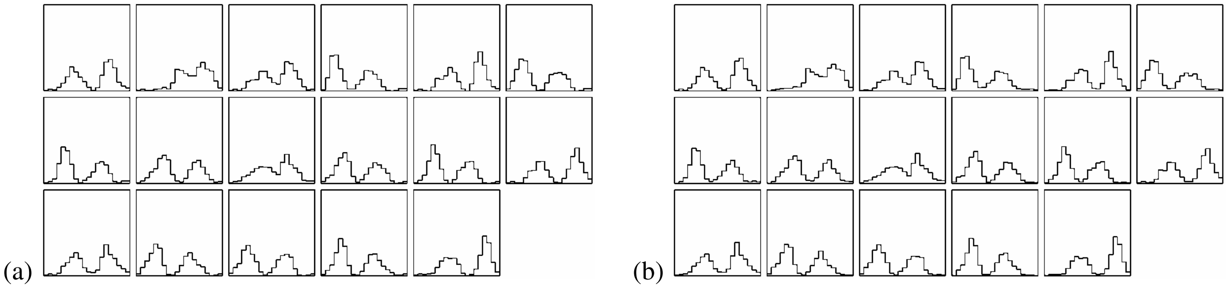

Figure 6. Comparison of histograms of (a) observed samples and (b) synthesized samples with 17 features.

are sufficient to model the distribution because adding more features in the learning process no longer changes the KL divergence significantly. This whole learning process of face (with 17 features) took nearly a day to complete because of the high dimensionality and complexity.

Comparing the synthesized Figure 5(c) and observed samples Figure 5(d), we may find it hard to discriminate them visually. Indeed, the histograms of observed and synthesized samples displayed in Figures 6 suggest that they match each other very well. The mean square error between the synthesized and the observed samples is less than 7%. This precise matching demonstrates that we can learn very complex distributions in high dimensional space. Note that most of these histograms are not simply Gaussians, thus it is inappropriate to assume a Gaussian model for faces.

6. Discussion

a)Why the Maximum Entropy Principle?

The maximum entropy principle generalizes the statistical properties in the observed samples, and makes the learnt model present information no more than what is available. This principle naturally leads to a Gibbs distribution.

b) Why the MCMC Method?

The parameters of Gibbs model are learnt by calculating the expected features. This expectation can be approximated by Monte Carlo integration. An efficient Markov chain is driven by conditional density from the Gibbs model to obtain the independent samples. This method is globally optimal if these samples can represent the underlying distribution.

c) Why the Minimum Entropy Principle?

The minimum entropy principle is to make the learnt distribution close to the observed one. A feature is selected to maximally decrease the entropy of current Gibbs model. The decrease of entropy, or the information gain by the new feature is measured by the KL divergence.

d) Why KL Divergence?

In pattern recognition, the Fisher Linear Discriminant [6] is always used to find a linear feature projected on which

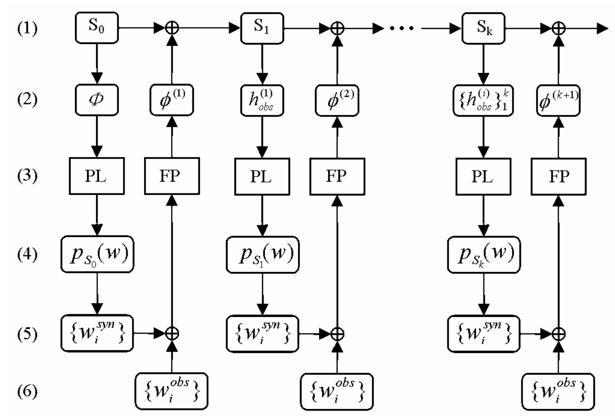

Figure 7. Overview of the learning framework. (1) current feature set; (2) observed histograms with respect to each feature; (3) learning procedure: PL-parameter learning, FP-feature pursuit; 4) the learnt distribution; (5) synthesized samples; (6) observed samples.

two classes can be maximally discriminated. This discriminant is based on the assumption that these two classes are well separated in the feature space. Other discriminants such as L^2 norm [16] merely record the global difference. We use KL divergence to measure the difference of two distributions even if they overlap. In fact, KL divergence emphasizes the tails of the distribution, which are very important to measure the difference of two densities and often have the most interesting characters.

e) Why an Inhomogeneous Model?

In homogeneous models, like texture, all elements (pixels) are unlabelled and thus are treated equal. The histogram of each feature can be directly calculated from them. In inhomogeneous models, such as a face, each element is a landmark that has a label and meaning. Therefore the Gibbs energy function depends on the label and the histogram of each feature must be computed from a set of independent samples.

f) Why Dimension Reduction Using PCA?

Directly applying the inhomogeneous Gibbs model to faces does not work in practice because the dimension of face model is too high. Thus, it is important to find a compact space to simplify the model. PCA provides such compact space constructed by the largest eigenvectors.

g) Why Not Use the Principal Components as the Initial Feature Set?

Instead of using a set of orthogonal bases such as principal components, we use 1D marginal statistics defined in the space formed by the principal components. Therefore, we can construct an over-complete set of bases which are more flexible for capturing the characteristics of the underlying local distribution.

7. Conclusion

We have built an inhomogeneous Gibbs model to learn prior distributions. In particular, this model is applied to obtain a better prior distribution for face modelling. A minimax entropy principle is used to derive an inhomogeneous Gibbs model, and features are selected by minimizing the KL divergence. Such learning framework is illustrated in Figure 7. To deal with the high-dimensional problem in face modelling, PCA is employed to reduce the feature space. The good results on face data demonstrate our model is effective and efficient to capture precise and subtle face patterns.

References

[1] Y. Amit, D. Geman, and K. Wilder. Joint induction of shape features and tree classifications. *IEEE Trans. on PAMI*, 19(11):1300–1305, 1997.

[2] Y. Amit, U. Grenander, and M. Piccioni. Structural image restoration through deformable templates. *J. Am. Stat. Assoc.*, pages 376–387, 1991.

[3] S. Baker and T. Kanade. Hallucinating faces. *Fourth International Conference on Automatic Face and Gesture Recognition*, March 2000.

[4] C. Bishop. *Neural Networks for Pattern Recognition.* Oxford Univ. Press, 1995.

[5] F. Bookstein. Size and shape spaces for landmark data in two dimensions. *Statistical Science*, 1(2):181–242, 1986.

[6] B.Ripley. *Pattern Recognition and Neural Networks.* Cambridge Univ. Press, 1995.

[7] T. Cootes and C. Taylor. Statistical models of appearance for computer vision. Technical report, University of Manchester, 2000.

[8] T. Cootes, C. Taylor, D. Cooper, and J. Graham. Active shape models-their training and application. *Computer Vision and Image Understanding*, 61(1):38–59, 1995.

[9] I. Dryden and K. Mardia. *Statistical Shape Analysis.* Wiley Publisher, 1998.

[10] W. Freeman and E. Pasztor. Learning low-level vision. *7-th International Conference on Computer Vision*, pages 1182–1189, 1999.

[11] W. Freeman and E. Pasztor. Learning to estimate scenes from images. *Neural Information Processing Systems*, 11, 1999.

[12] H. Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. *IEEE Trans. on PAMI*, 20(1):23–38, January 1998.

[13] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neurosciences*, 3:71–86, 1991.

[14] A. Yuille. Deformable templates for face recognition. *Journal of Cognitive Neurosciences*, 3(1), 1991.

[15] S. Zhu. Embedding gestalt laws in markov random fields. *IEEE Trans. on PAMI*, November 1999.

[16] S. Zhu, Y. Wu, and D. Mumford. Minimax entropy principle and its application to texture modeling. *Neural Computation*, 9(8), November 1997.