Title
Multiscale Generative Model of Human Faces

Authors
Zijian Xu
Hong Chen
Song-Chun Zhu

Publication Date
2011-10-25

Multiscale Generative Model of Human Faces

Zijian Xu, Hong Chen, and Song-Chun Zhu

Department of Statistics
University of California, Los Angeles, Los Angeles, CA 90095
email: {zjxu,hchen,sczhu}@stat.ucla.edu
August, 2004

Abstract

In this paper, we propose a framework for modelling human faces over scales. As a person walks towards the camera, more details of the face will be revealed and thus more random variables and parameters have to be introduced. Accordingly, a series of existing generative models are organized as five regimes, which form nested probabilistic families. The generative model in higher regime is augmented by (1) adding more latent variables, features extractors, and (2) enlarging the dictionary of description, e.g. PCA bases, local parts or sketch patches. The minimum description length *(MDL) is used as a criterion for the model selection and transition. As observed in our experiment, the optimal model switches among the different regimes when the scale changes. A sequence of tasks, such as face detection, recognition, sketching and super-resolution etc. can be accomplished based on the models in the different regimes.*

1. Introduction

Human faces appear in a wide range of scales in images and videos. When a person walks towards the camera, his/her face can grow from a 3×3 pixels to 300×300 pixels in size or even more. While the changes over scales may appear continuous in raw images (at retina), it evokes abrupt "quantum jumps" in our high level perception (at visual cortex). In a mathematical term, we must be augmenting our perception with more and more detailed representations/descriptions and switching the generative models over scales.

Figure 1 demonstrates that while the image size is enlarged, the complexities of graph structures increase and the dictionaries of description become more sparse to account for the variety of features revealed. A series of existing generative models are used to form the perceptual/model space. Based on the experience of human perception, we roughly divide this space into five regimes. (1) *Texture regime*. When faces are viewed at far distance, such as the crowd image in Fig.4, we cannot see reliably the individual faces and thus perceive a texture impression. This is modelled by the FRAME model[17] on pixels. Such model can be used for segmenting crowds from big scenes, (2) *PCA regime*. The PCA model[12, 11], AAM model[3, 4] and morphable model[7, 8, 15] have been proven to be sufficient for characterizing the images at middle scales. Therefore some tasks for low resolution faces can be accomplished in this regime, such as face detection[16]. (3) *Parts regime*. With higher resolutions, the eyes, mouth and nose etc. can be clearly characterized by a bigger dictionary that consists of the local facial component with iconic changes, eg. open/closed mouth or eyes, which is useful for classification and recognition[18, 19]. (4) *Sketch regime*. With higher resolutions, more details are revealed, such as the eyelid, eyebrows and wrinkles. We use a much larger dictionary of small patches to account for the variety of these local structures. The model in this regime can be used to automatically generate artistic face sketches[1]. (5) *Super-resolution regime*. This regime is similar to the sketch regime with more sparse dictionary. To describe even the facial marks of extremely high resolutions, more detailed (usually smaller) structures are learned. The model can be used for super-resolution.

In section 2 we talked about the description, learning and transition of the models over scale. Section 3 described in details the five regimes models and their training. In section 4 we briefly introduced the sampling and inference algorithm. And in section 5 is a sequence of experiments we conducted. Some discussions are given in section 6.

2. Learning Model over Scales

2.1. Model Description

Let $\mathbf{I}$ be an image on lattice Λ with an unknown number of faces at various locations and scales. We thus pose face modelling as a statistical learning problem whose objective is to seek a probability model $p(I)$. It has become clear recently that the p must integrate[6] generative (graphical) models with descriptive (Markov random field) models. The former have multiple graph layers with each layer generating the layer below by a dictionary of image elements. The latter specify the graph structure and arrangement by feature statistics. In our framework, the probability models are

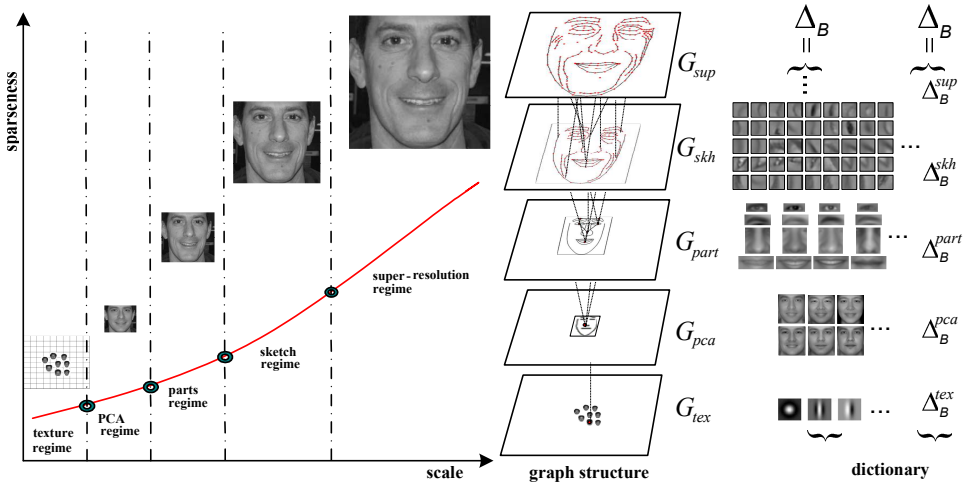


Figure 1: Faces viewed at various distance (scales) fall in different regimes of mathematical models. From far to near, the representation is augmented with more complex graph structures and large dictionary sizes to account for details.

unified as the following form:

$$p(\mathbf{I}; G, \mathbf{B}, \boldsymbol{\alpha}, \mathbf{F}, \boldsymbol{\beta})$$

,where $W = (G, \mathbf{B}, \boldsymbol{\alpha}, \mathbf{F}, \boldsymbol{\beta})$ are the latent variables. G is a graph representation of multiple layers, where the vertices can be pixels, a single face, facial components or local patches according to the layer they rest in. While the image scale goes up, the graph structure becomes more complex by adding new layers on the previous ones. The vertices of previous layers are expanded as subgraphs in the new layer, e.g. in Figure 7 the vertex denoting nose expanding as a subgraph of local patches. Figure 1 roughly shows how graphs evolve over the scale. $\mathbf{B}$ and $\boldsymbol{\alpha}$ denote the geometric and photometric properties on the graph vertices, e.g. the PCA bases and their coefficients. In Figure 1, we can see that dictionaries of $\mathbf{B}$ become more sparse over scale to allow more varieties and describe more details. On an attributed graph G, an inhomogeneous Gibbs (MRF) model is used to characterize the spatial arrangement together with the couplings of their attributes. $\mathbf{F}$ is a set of filters extracting features on G. $\boldsymbol{\beta}$ is the parameters for the Gibbs potentials.

The triple $(\Delta_G, \Delta_{\mathbf{B}}, \Delta_{\mathbf{F}})$ represents the dictionaries of the model description. Δ_G is the set of all valid graph configurations. $\Delta_{\mathbf{B}}$ is a dictionary for basic representation units on the graph vertices. $\Delta_{\mathbf{F}}$ is the filter bank, in which many feature extractor on a valid graph vertex and its neighbors are defined.

2.2. Model Learning and Transition

The probability model p is pursued in a series of probability families which at the end should be sufficient to approximate natural frequency $f(\mathbf{I})$ to any precision.

$$\Omega_0 \subset \Omega_1 \subset \cdots \subset \Omega_K \rightarrow \Omega_f$$

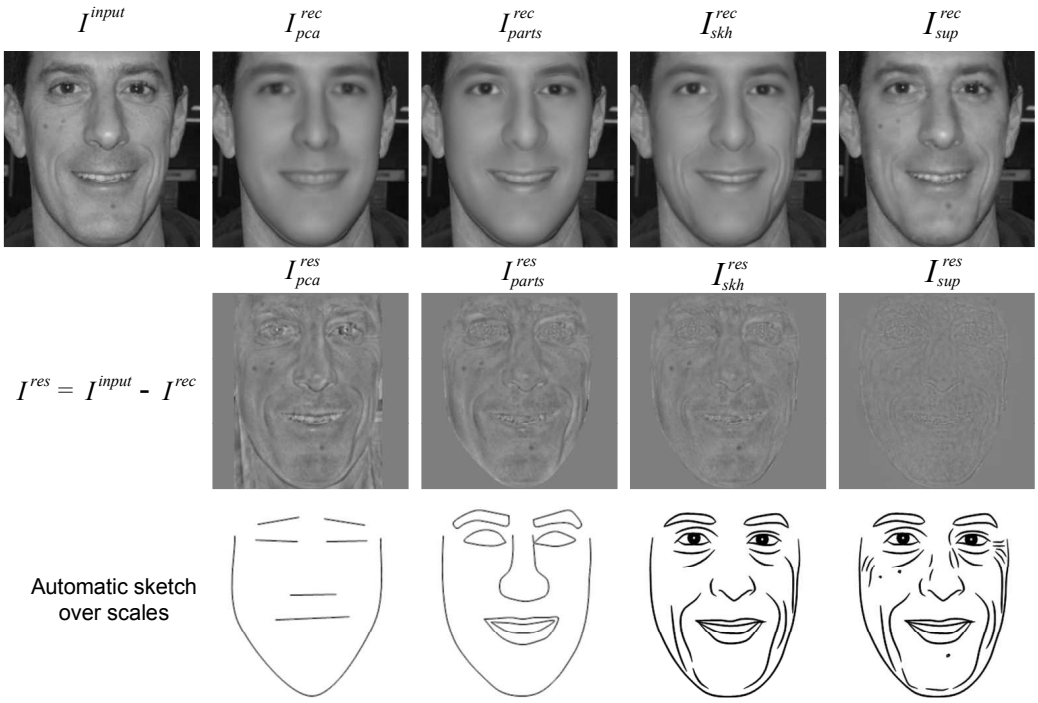


Figure 2: Face image of 256×256 was reconstructed by the four models with dictionary where they reach the *minimum description length*.

The model space is divided into a number of regimes. At the beginning, a series of generative models in the nested probabilistic families are learned in a supervised way, which means the latent variables W are partially or fully given. In our model, the landmarks on face images are manually labelled to give the locations of the local features, such as eyes, nose and face contour, etc.. Later we boostrap the learning procedure by the unsupervised data inferred by the model.

A specific face population at certain scale is modelled by model from one of the regimes. How to select the "appropriate" model is therefore a key problem. In Fig. 8, we compared the absolute value of the *per-pixel* reconstructed error (residua) of testing images for a number of models. It shows that for the same model usually the residua is reduced when the bigger dictionary is applied. However, if the image size keeps increasing, performance of the simples model might level off. Then we have to switch to more complex models. Although enlarging the dictionary size or switching to more complex model can help us reduce the residua, they also increase the uncertainty of estimating the parameters and model complexity. As discussed in [5], we would like a model that is capable of representing any valid face instance and as compact as possible. By posing this problem as one of minimizing the description length (MDL) of the model, we develop a criterion to select the most "sufficient and compact" model by comparing the minimum coding length.

Let $\Omega_I = \{I_1, ..., I_M\}$ be the sample set. In general, the coding length DL required

for a model to describe $\Omega_{\mathbf{I}}$ using the dictionary Δ is:

$$DL = L(\Omega_I; \Delta) + L(\Delta) \tag{1}$$

,where the first term is the expected coding length of $\Omega_{\mathbf{I}}$ given Δ and the second term is the coding length of Δ.

Empirically, we can estimate DL by:

$$\hat{DL} = \sum_{I_i \in \Omega_{\mathbf{I}}} \sum_{w \sim p(w|I_i;\Delta)} (-\log p(I_i|w;\Delta) - \log p(w)) + \frac{|\Delta|}{2} \log M \tag{2}$$

,where $w \sim p(w|I_i;\Delta)$ can be sampled by Markov Chain Monte Carlo (MCMC) inference; M denotes the number of data; $|\Delta|$ is the size of dictionary, e.g. in PCA model it is the pixel number of meanface and eigenfaces used. In Figure 9, we plot how the coding length of the models changes when different dictionary sizes are applied. At small scales, like 32×32 or 64×64, the MDL of PCA model is shorter than parts model or sketch model. And at large scales, like 128×128 and 256×256, parts model and sketch model outperform respectively.

To summarize, Figure 8 and Figure 9 show that: (1) in the sense of reducing the residual, simple models perform as well as the complex ones at small scales, but their performance levels off as scale goes up; (2) compared by using the MDL, simple models are sufficient and more compact for modelling faces at small scales, while better (usually more complex) models are preferred when scales become larger.

3. Five Regimes

In this section, we specify 5 regimes of models and their training to illustrate the nested families of models.

3.1. Texture Regime — FRAME Model

In this regime, each face appears as only a few pixels wide. The model treats faces as a texture phenomenon without having to identify the individual faces. Thus we used FRAME model [17] for the crowd scenes. As Figure 3(a) shows, G is a lattice with each pixel in the image as one vertex, B are the pixels and $\boldsymbol{\alpha}$ are their intensities, F is a number of Gabor filters and Laplacian of Gaussian filters as shown in Figure 3(b). As $W_{tex} = \{\mathbf{I}_{\text{tex}}(x, y), (x, y) \in \Lambda\}$, we can generate it as

$$p(W_{\text{tex}}) \propto \frac{1}{Z} \exp\{-\sum_{\alpha=1}^{K} < \boldsymbol{\beta}^{(\alpha)}, Hist(F^{(\alpha)} * \mathbf{I}) >\} \tag{3}$$

The image is then generated as

$$\mathbf{I}_{\text{tex}} = W_{\text{tex}} \sim p(W_{\text{tex}}) \tag{4}$$

We learned this model from several crowd scenes and drew random samples (Figure 4) from it using Markov chain Monte Carlo simulation. The random images appear like crowd visually.

Figure 3: Texture regime for crowds. (a) Graph G_{tex} is a lattice; (b) Dictionary Δ_{tex} includes filter $\Delta_F = \{\text{LoG}, \text{GaborSine}, \text{GaborCos}\}$.

(a). input image (b). synthesized image

Figure 4: Faces in the texture regime. (a) is the observed crowd image; (b) is the random sampled image from the learned texture model (FRAME) $\mathbf{I} \sim p(\mathbf{I}; \Theta)$.

3.2. PCA Regime — Active Appearance Model

In this regime, G consists of two layers: the lattice from texture regime is inherited as background, on which a number of attributed vertices for individual faces are added as foreground. The connections among these face vertices are very weak, so we assumed that they are independently distributed. AAM[4] was applied for modelling each face. We divided the face patch into shape and intensity, where shape is represented by a set of labelled landmarks (see Figure 5(a)) and intensity is obtained by warping the observed face patch to the mean shape. $B = \{B_{\text{shape}}, B_{\text{inten}}\}$, where the shape bases B_{shape} include the mean shape and the eigen-shapes, and the intensity basess B_{inten} include the mean intensity and eigen-intensities (see Figure 5(b)). To allow more diversity, a mixture model of AAM was trained in our experiment. The training set was clustered into two types: one of them as plain face with the mouth closed and the other one as smiling face with the mouth open. $\boldsymbol{\alpha} = \{\boldsymbol{\alpha}_{shape}, \boldsymbol{\alpha}_{inten}, \ell\}$, where $\boldsymbol{\alpha}_{shape}$ and $\boldsymbol{\alpha}_{inten}$ are weights vectors of shape bases and intensity bases respectively. ℓ is the index of cluster in the mixture model. $F = \emptyset$ for the graph layer of individual faces.

As the iid assumption of individual faces in this regime, $p(W_{\text{pca}}|W_{\text{tex}}) = p(W_{\text{pca}})$. The latent variables W_{pca} can be obtained as

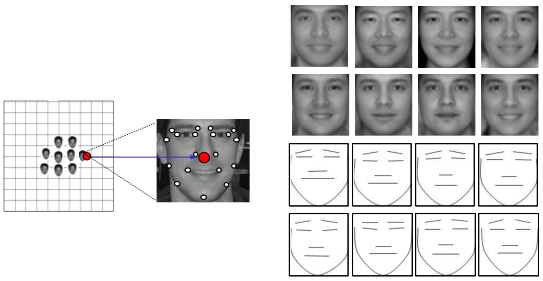
$$p(W_{pca}) = \sum_{\ell=1}^{N_{\text{cluster}}} \lambda_\ell p(W_{pca,\ell}|\ell), \tag{5}$$

,where λ_ℓ are weights of clusters in the mixture model and $\sum_{\ell=1}^{N_{\text{cluster}}} \lambda_\ell = 1$. Then $\mathbf{I}_{pca}$ is generated by $\mathbf{I}_{tex}$ and $\mathbf{J}_{pca}$:

$$\mathbf{I}_{pca}(x,y) = \begin{cases} \mathbf{I}_{tex}(x,y) & \text{if(x, y)} \in \Lambda_{\text{background}} \\ \mathbf{J}_{pca}(x,y) + noise & \text{if(x, y)} \in \Lambda_{\text{face}} \end{cases} \tag{6}$$

$$\mathbf{J}_{pca}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = T(\sum_i \alpha_{\text{i}}^{shape} b_{\text{i}}^{shape}, \sum_j \alpha_{\text{j}}^{inten} b_{\text{j}}^{inten}) \tag{7}$$

,where Λ_{face} is the domain covered by face and $\Lambda_{\text{background}}$ is the background. T is a warping function. Given the input as the reconstructed intensity patch and the reconstructed shape, T gives the output as the reconstructed face patch.



(a) (b)

Figure 5: PCA regime. (a) Expanding graph from $G_{\rm tex}$ to $G_{\rm pca}$; (b) $\Delta_{\rm pca}$, including both the intensity bases and shape bases;

3.3. Parts Regime

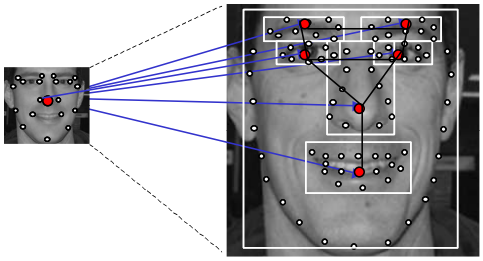
In this regime, the previous graph layers are first inherited to G, then a new layer with the face vertex expanded as a subgraph is added on top of it. Each vertex in the subgraph represents one of the local facial components (See Figure 6(a)). Similar to the PCA regime, the mixture AAM model for each component was trained. Note that the positions of every landmarks are governed by one or more corresponding landmarks in the previous regime. For example, the weighted center of landmarks on nose in parts regime must correspond to weighted center of those denoting nose in PCA regime. The local components are clustered into several types, such as open mouth, closed mouth, double-curved lid, single-curved lid, etc.. $B = \{\{B_{\rm shape,i}\}, \{B_{\rm inten,i}\}, i = 1, .., N_{\rm part}\}$ includes mean shape and eigen-shapes as shape bases, mean intensity and the eigen-intensities as intensity bases for all local components (see Figure 6(b)). $\boldsymbol{\alpha} = \{\{\ell_i, (x_i, y_i), s_i, \theta_i, \boldsymbol{\alpha}_{\ell_i}^{\rm shape}, \boldsymbol{\alpha}_{\ell_i}^{\rm inten}\}, i = 1, .., N_{\rm part}\}$. $N_{\rm part}$ is the number of local facial components. ℓ_i is index of clusters for the mixture model. (x_i, y_i),s_i and θ_i denote the center, size and orientation of the $i^{\rm th}$ component. $\boldsymbol{\alpha}_{\ell_i}^{\rm shape}$ and $\boldsymbol{\alpha}_{\ell_i}^{\rm inten}$ are weights vectors of the shape bases and intensity bases.

Because of the great varieties of facial elements, a compact PCA representation may not be sufficient to model the spacial relationship among the vertices and their neighbors. Therefore, we use a non-parametric Gibbis distribution on the graph G to capture the non-Gaussian properties. Here G consists of three layers, including two layers inherited from previous regimes and one expanded in current regime. We can automatically learn the most effective features by the Minimax Entropy framework [17]. For simplicity we used a set of manually designed features F, such as the distance, the size ratio and the geometrical symmetry of the eyes, the tilting angel of mouth and nose, or the intensity similarity between overlapped domain of the different regimes, etc.. In our experiment these features seemed to be sufficient.

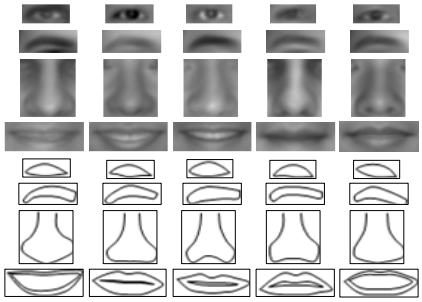
The latent variables in parts regime is obtained as

$$p(W_{\rm part}|W_{\rm pca}) \propto \exp\{-\sum_{<{\rm P_i},{\rm P_j}>\in {\rm E}} \sum_{\ell} \lambda_\ell \psi_\ell(P_i, P_j)\} \tag{8}$$

E denotes the edge set on graph G. ψ_ℓ is the potential function on the attributes of two connected components P_i and P_j, which are governed by W. The face image $\mathbf{I}_{part}$ is



(a)



(b)

Figure 6: Parts regime. (a) Expanding graph from G_{pca} to G_{part}; (b) Dictionary Δ_{part}, including both the intensity bases and shape bases;

generated by updating the local facial components on $\mathbf{I}_{pca}$, i.e. the reconstructed patch in parts regime will occlude the overlapped domain in PCA regime.

$$\mathbf{I}_{part}(x,y) = \begin{cases} \mathbf{I}_{pca}(x,y) & \text{if(x, y)} \in \Lambda_{\text{face}} \\ \mathbf{J}_{part}(x,y) + noise & \text{if(x, y)} \in \Lambda_{\text{part,i}} \end{cases} \tag{9}$$

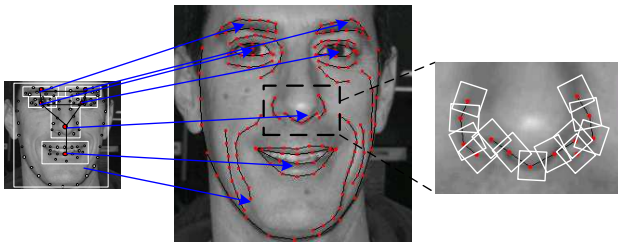
,where $i = 1, .., N_{\text{part}}$ and for the i^{th} component we have:

$$\mathbf{J}_{part,i}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = T(\sum_j \alpha^{\text{shape}}_{\ell_i,j} b^{\text{shape}}_{\ell_i,j}, \sum_k \alpha^{\text{inten}}_{\ell_i,k} b^{\text{inten}}_{\ell_i,k}, x_i, y_i, s_i, \theta_i) \tag{10}$$

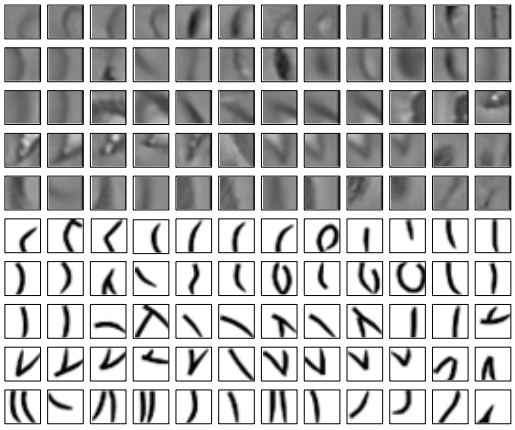
,where T is a function warping the reconstructed intensity to reconstructed shape, while it also performs affine transformation of the reconstructed local patches.

3.4. Sketch Regime

In this regime, as shown in Figure 7, G consists of all the previous layers and a new layer with two kind of subgraphs. Each facial component, which is a vertex in the previous regime, is expanded as a subgraph $\{G_{com,i}, i = 1, .., M_{com}\}$, where M_{com} is the number of facial components. There are also several chains to capture the curve features of the face $\{G_{cur,i}, i = 1, .., M_{cur}\}$, where M_{cur} is the number of curves. For each vertex v_i in the graph, we denote the geometry parameters of it as: $T_i = \{(x_i, y_i), s_i, \theta_i\}$, where (x_i, y_i), s_i, θ_i are the center, scale and orientation of the i^{th} local patch respectively. They are conditioned on previous regime, e.g. the center of the small patch is governed by positions of the corresponding landmarks in



(a)



(b)

Figure 7: Sketch regime. (a) Expanding graph from G_{part} to G_{skh}, with subgraph of the nose expanded as an example; (b) A subset of dictionary Δ_{skh}, including both the intensity bases and their sketch correspondingly;

parts regime. Then the image patches can be obtained by using affine transformation T_i. We normalized the image patch of each vertex v_i by photometric transformation A_i. Similar to parts regime, we trained these normalized patches with mixture PCA. $B = \{B_{com}, B_{cur}\}$ includes the mean and PCA bases for both the vertices on subgraphs $G_{\text{com,i}}, i = 1, .., M_{\text{com}}$ and $G_{\text{cur,j}}, j = 1, .., M_{\text{cur}}$. Some examples are shown in figure 7(b). Each image patch can be represented by (ℓ_i, α_i), where ℓ_i is the cluster label and α_i is the weights vector of the PCA bases. Thus $\boldsymbol{\alpha} = \{((x_i, y_i), s_i, \theta_i, A_i, \ell_i, \alpha_i), i = 1, .., N_{\text{skh}}\}$, where N_{skh} is the number of vertices on sketch layer.

For $p(W_{skh}|W_{part})$, if we define a set of features $F = \{\psi_\ell, \ell = 1, .., N_F\}$ on the vertices and the edges that link them, a Gibbs model as in [17] can be built.

$$P(W_{skh}|W_{part}) \propto \exp\{-\sum_l \lambda_l \psi_l(W)\} \tag{11}$$

There are two type of edges in G. One type link the vertices in $G_{\text{com,i}}, i = 1, .., M_{\text{com}}$ with the corresponding vertex in parts regime. The other type connect adjacent vertices in current layer. So we define the following features:

- For the edges between two layers, denote the vertex in G_{skh} as v_i, and the vertex in G_{part} as v_j. The feature set has two parts: geometry features and intensity features.

  Given v_j, we can predict the location of v_i by linear regression as $(\hat{x}_i, \hat{y}_i)$. The geometry features is defined as the distance between (x, y) and $(\hat{x}_i, \hat{y}_i)$.

  Let denote the d_i as the vectorized intensity patches of v_i, and d_j as the corresponding vectorized intensity patches on $\mathbf{I}_{part}$. As in [10], we choose a set of linear features over combined (d_i, d_j) as $\{\psi_{int}^{dif}(v_i, v_j)_\ell, \ell = 1, ..N_F\}$, where N_F is the number of selected features.

- For adjacent nodes v_i and v_j, d_i and d_j are the vector of the overlapped pixels. n_{ij} is the number of overlapped pixels. We define feature as the absolute value of *per-pixel* difference between d_i and d_j.

We can learn the parameters λ_l in Eqn 11 by the gradient descent algorithm in Minmax Entropy framework [17].

Similar to the parts regime, the face image $\mathbf{I}_{skh}$ is generated by occluding the local patches on $\mathbf{I}_{part}$ with domain covered by sketch layer.

$$\mathbf{I}_{skh}(x, y) = \begin{cases} \mathbf{I}_{part}(x, y) & \text{if(x, y)} \in \Lambda_{\text{part}} \\ \mathbf{J}_{skh}(x, y) + noise & \text{if(x, y)} \in \Lambda_{\text{skh,i}} \end{cases} \tag{12}$$

,where $i = 1, .., N_{\text{skh}}$ and for the i^{th} we have:

$$\mathbf{J}_{skh,i}(B, \boldsymbol{\alpha}) = T_i(\sum_j \alpha_{\ell_i j} b_{\ell_i j}, x_i, y_i, s_i, \theta_i, A_i) \tag{13}$$

, where the $b_{\ell_i j}$ and $\alpha_{\ell_i j}$ are the PCA bases and weights respectively. T_i is the affine transformation and A_i is the photometric transformation.

3.5. Super-resolution Regime

In this regime, we build model for two purpose: (1) to represent vertices in sketch regime graph with more details, such as decomposing the strokes of eyebrow into even smaller patches to represent every single hair; (2) to describe the independent structures such as the beauty spots, the tiny curves and the skin texture. Above the previous layers, the new layer of G consists of: (a) subgraphs expanded from vertices in sketch regime; (b) vertices for independent structures. The vertices in (a) are the same kind of patches in sketch regime but smaller (Figure 7(b)), we learn these patches in the same way as in sketch regime and form the dictionary B_{dep}. For vertices in (b), locally they appear as texture phenomenon, so we may model them by combinations of the bases in texture regime (Figure 3(b)) and form the dictionary B_{ind}. With $B = \{B_{dep}, B_{ind}\}$ and given latent variables, we can generate face image $\mathbf{I}$ in super-resolution regime.

A similar feature bank $F_{dep} = \{\psi_l^{dep}, l = 1, .., N_l^{dep}\}$ can be defined as in sketch regime. Also the filter bank $F_{ind} = \{\psi_k^{\text{ind}}, k = 1, .., N_k^{ind}\}$ as in texture regime are used. Thus the latent variables $W_{sup} = \{W^{ind}, W^{dep}\}$ can be obtained by

$$P(W_{sup}|W_{skh}) \propto \exp\{-\sum_l \lambda_l^{dep} \psi_l^{dep}(W^{dep}) - \sum_k \lambda_k^{ind} \psi_k^{ind}(W^{ind})\} \tag{14}$$

The face image $\mathbf{I}_{sup}$ is generated by occluding the sketch domain by the domain covered by super-resolution layer

$$\mathbf{I}_{sup}(x,y) = \begin{cases} \mathbf{I}_{skh}(x,y) & \text{if(x, y)} \in \Lambda_{\text{skh}} \\ \mathbf{J}_{sup}(x,y) & \text{if(x, y)} \in \Lambda_{\text{sup}} \end{cases} \tag{15}$$

$$\mathbf{J}_{sup} = \mathbf{J}_{sup}^{ind} + \mathbf{J}_{sup}^{dep} \tag{16}$$

4. Sampling and Inference

After the model is learned over scales, we can draw random samples I^{syn} or infer hidden variable W for given image I^{obs}. Therefore, we are able to boostrap the learning procedure by unsupervised data. These basic operations and their combinations can also be applied for many tasks. In this section we briefly discuss how the sampling and inference can be done.

4.1 Sampling Over Scales

Had model learned, we are able to draw random samples $I^{\text{syn}} \sim p(I|W)p(W)$. Because of the hierarchical structure of our model, we can sample it from coarse to fine. For two adjacent regimes, we denote the latent variables as W^- in lower regime and W^+ in higher regime respectively. Given W^-, W^+ can be sampled from $p(W^+|W^-)$. Due to the high dimension of W, MCMC is selected for sampling method. To reduce the burn-in time and mixing rate, we use a mixing MCMC which combines two dynamics. The first dynamic is Metroplis-Hasting algorithm. We estimate a simple model $q(W^+|W^-) = \prod_i q(W_i^+|N_i(W^-))$, where $N_i(W^-)$ is a subset of W^- related to W_i^+. $q(W_i^+|N_i(W^-)$ can be modelled by Gaussian or Mixture Gaussian to simplify the sampling. $q(W^+|W^-)$ is used as the proposal distribution of the Metroplis-Hasting algorithm. This type of dynamic can produce long jumps and can jump between different dimension solution space. The second dynamic is Gibbs sampler. The latent variable $W_i^+ \sim p(W_i^+|W^-, N(W_i^+))$ is sampled iteratively.

4.2 Inference Over Scales

Similar to the sampling algorithm, we can infer $W^* \sim p(W|I^{\text{obs}})$ in a *coarse-to-fine* strategy. Given W^{-*} in the lower regime, we infer W^{+*} in higher regime according to $W^{+*} = \arg\max_{W^+} p(W^+|W^-, I^+; \Delta)$. To improve the efficiency of our algorithm, We can also integrate several existing efficient method such as Adaboosting face detector, AAM and ASM algorithms. To summarize, our inference algorithm consists of following steps:

- Draw several samples from $q(W^+|W^-)$;
- For each node or subgraph, run efficient algorithm with the samples $q(W^+|W^-)$ as the starting point to get a set of candidates $W^{+*}(i)$. For example, we use AAM algorithm for each cluster of PCA Regime and each node in Parts Regime, and ASM algorithm for every modes of each subgraph in Sketch regime.

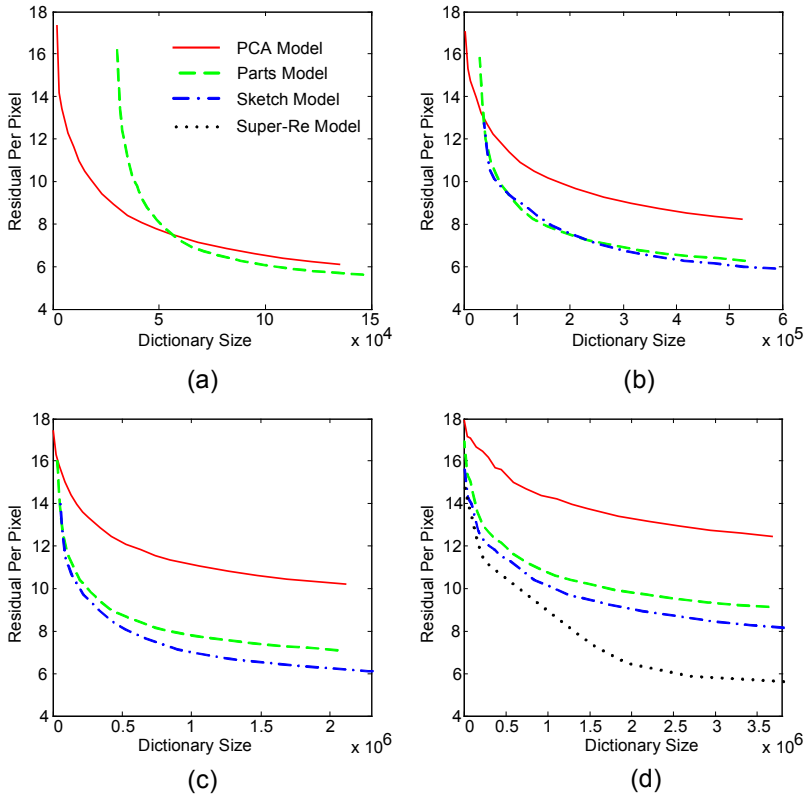


Figure 8: Plot of *per-pixel* reconstructed error v.s. dictionary size $|\Delta|$ at four scales. (a) 32×32; (b) 64×64; (c) 128×128; (d) 256×256

- Run MCMC with two dynamics. One dynamic using Metroplis-Hasting algorithm with the proposal distribution:

$$q(W_A^+|W_B^+, W^-, I) = \sum G(W^{+*}(i), \sigma^2 \mathrm{I}) \tag{17}$$

  where σ is a small constant. Gibbs sampler is used as another dynamic to jump in the solution space with same dimension.

5. Experiments

To verify the framework we proposed, three experiments were conducted based on 350 frontal face images chosen from different genders, ages and races — 200 for training and 150 for testing. All the images are resized to four different scale levels: 32×32, 64×64, 128×128 and 256×256. The landmarks over scale for the face images are manually labelled. With the inference algorithm, later we can keep including more training data with the landmarks automatically located.

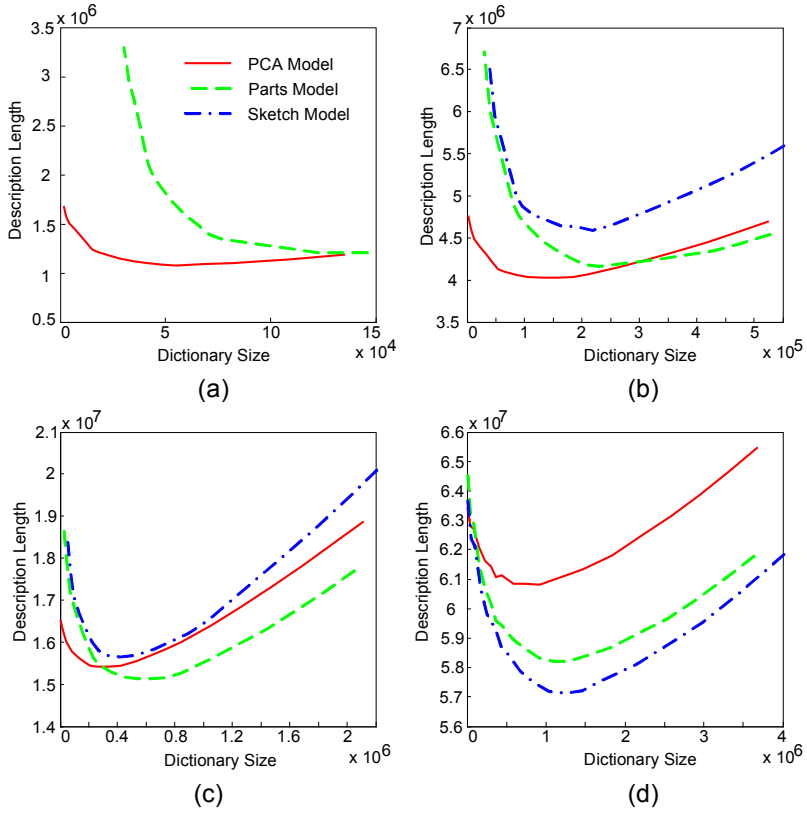


Figure 9: Plot of coding length $\hat{DL}$ for the ensemble of testing images v.s. dictionary size $|\Delta|$ at four scales. (a) 32×32; (b) 64×64; (c) 128×128; (d) 256×256

- *Experiment I*. The four models, PCA, Parts, Sketch and Super-resolution, were trained at four scales respectively. Given the testing images, we first inferred the latent variables as $W \sim p(W|I^{\text{obs}})$. After that the testing images were reconstructed by $I^{rec} \sim p(I|W)$ and the absolute value of reconstructed error (residua) $\|I^{obs} - I^{rec}\|$ were then calculated. Figure 8 plots the *per-pixel* reconstructed error to dictionary size for four scales. It shows that although we can somehow reduce the residua at the beginning by using bigger dictionary and more parameters, the performance of a relatively simple model will finally level off as image scale increases, e.g. adding the high-order PCA bases becomes ineffective. In order to further reduce the residua, we have to switch to a higher regime in the model space, which usually means the use of a more complex model.

- *Experiment II*. To select a most appropriate model for given image size, we compared the *minimum description length* (MDL) for PCA model, Parts model and Sketch model. Figure 9 plots the coding length to dictionary size for four scales. We can see that at small scale like 32×32 simple model such as PCA has the

minimum coding length, while at larger scale like 128×128 the Parts model outperforms, and at the largest scale like 256×256 the Sketch model beats the other two. By applying this criterion we are able to select the most "sufficient and compact" generative model for coding a given set of face data at certain scales.

- *Experiment III*. We tested the performance of our framework in reconstructing and rendering face images of 256×256. Using all the models with corresponding dictionary where they reach the *minimum coding length*, we inferred and reconstructed the new coming testing face images respectively. Figure 2 shows the reconstructed results. As a benefit of our generative models, the sketch of face at each scales are automatically obtained by simply replacing the dictionary of intensity bases with symbolic/sketch bases. More results of reconstruction and face sketch are shown in Figure 10. With the model, we also randomly sampled some new faces and their sketches for different scales, which are not included in this paper due to the page limit.

6. Discussion

In the literature, there is a well-known scale space theory which is characterized by the Gassian and Laplacian pyramids (see[9]). This theory is mostly focused on the *image space* with continuous and linear additive representations. We argue that there is a need for developing a new scale space theory on the *perceptual space* or *model space*. This new scale space consists of a series of generative models from nested probabilistic families, each characterizing the face population at a certain scale. The new theory is mostly concerned with the augmentation and switches/jumps of models over the image scale. There are two main axes for this augmentation: (1) adding more latent variables, features extractor on graphs, parameters, and (2) enlarging the vocabulary (dictionary) of representation. This new scale-space theory is needed for an integrated treatment of various vision tasks, such as detection, recognition and super-resolution in a common framework. As scaling is ubiquitous in all natural images, study of the new scale-space theory will be crucial for generic vision modelling with generative models.

References

[1] H. Chen, Y. Q. Xu, H. Y. Shum, S. C. Zhu, and N. N. Zhen, "Example-based facial sketch generation with non-parametric sampling", ICCV, 2001.

[2] C. Choi, T. Okazaki, H. Harashima, and T. Takebe, "A system of analyzing and synthesizing facial images", *Proc. of IEEE*, 2665-2668, 1991.

[3] T.F.Cootes, C.J.Taylor, D. Cooper, and J. Graham, "Active Appearance Models–Their training and applications", *Computer Vision and Image Understanding*, 61(1):38-59, 1995.

[4] T.F.Cootes, G.J. Edwards and C.J.Taylor. "Active Appearance Models", *Proceedings of ECCV*, 1998

[5] R.H. Davies, T.F. Cootes, C.Twining and C.J. Taylor, "An Information Theoretic Approach to Statistical Shape Modelling", *Proc. British Machine Vision Conference*, pp.3-11, 2001

[6] C.E. Guo, S.C. Zhu, and Y.N. Wu, "Modeling visual patterns by integrating descriptive and generative models", *IJCV*, 53(1), 2003.

[7] P.L. Hallinan, G.G. Gordon, A.L. Yuille, and D.B. Mumford, "Two and Three Dimensional Patterns of the Face", *A.K. Peters, Natick*, MA, 1999.

[8] M. J. Jones and T. Poggio, "Multi-dimensional morphable models: a framework for representing and matching object classes", *Int'l J. of Computer Vision*, 2(29), 107-131, 1998.

[9] T. Lindeberg, *"Scale-Space Theory in Computer Vision"*, Kluwer Academic Publishers, Netherlands, 1994.

[10] C. Liu, S.C. Zhu, and H.Y. Shum, "Learning inhomogeneous Gibbs models of faces by minimax entropy", *Proc. 8th Int'l Conf. on Computer Vision*, Vancouver, CA, 2001.

[11] M. Turk and A. Pentland, "Eigenfaces for recognition," *J. of Cognitive Neurosciences*, vol.3, no.1, pp. 71-86, 1991.

[12] L. Sirovich and M. Kirby, "Low-dimensional procedure for the characterization of human faces", *J. of Optical Society of America*, 4:519-524, 1987.

[13] K.K. Sung and T. Poggio, "Example-based learning for view-based human detection", *IEEE trans. on PAMI*, vol.20, no.1, 39-51, 1998.

[14] M. Turk and A. Pentland, "Eigenfaces for recognition," *J. of Cognitive Neurosciences*, vol.3, no.1, pp. 71-86, 1991.

[15] T. Vetter, "Synthesis of novel views from a single face image", *Int'l J. Computer Vision*, 2(28), 103-116, 1998.

[16] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features,", *CVPR*, 2001.

[17] S. C. Zhu, Y. N. Wu and D. B. Mumford, "Minimax entropy principle and its application to texture modeling", *Neural Computation* Vol. 9, no 8, Nov. 1997.

[18] S. Ullman, E. Sali, "Object Classification Using a Fragment-Based Representatio", *First IEEE International Workshop, BMVC*, 2000

[19] B. Heisele, P. Ho, J. Wu and T. Poggio, "Face Recognition: Component-based versus Global Approaches", *Computer Vision and Image Understanding*, Vol. 91, No. 1/2, 6-21 2003.

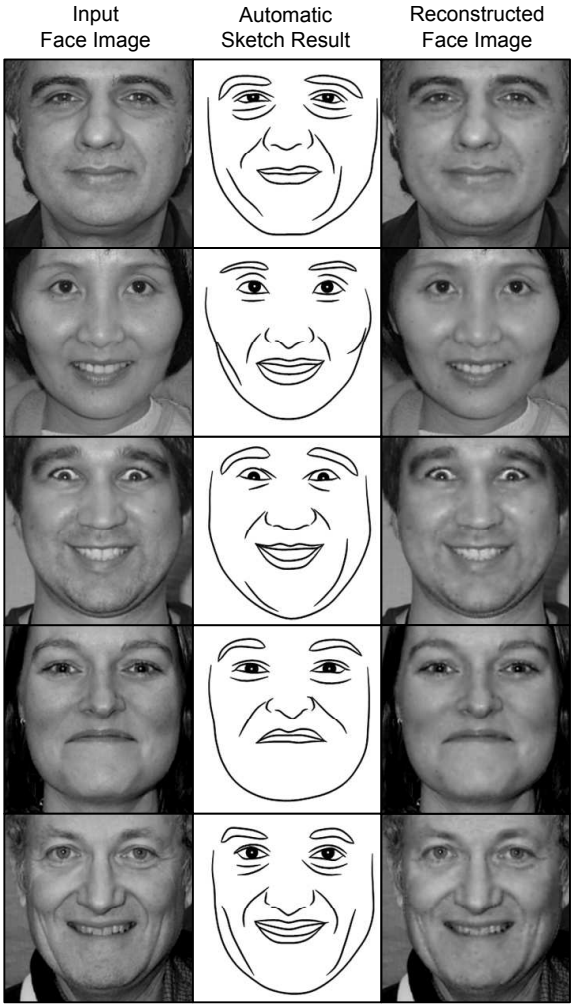


Figure 10: More results of reconstructed image and generated sketch of our model.