A Hierarchical Compositional Model for Face Representation and Sketching

Zijian Xu[1], Hong Chen[1], Song-Chun Zhu[1], Jiebo Luo[2]

[1]Department of Statistics, University of California at Los Angeles, Los Angeles, CA 90095

{zjxu,hchen,sczhu}@stat.ucla.edu

[2]Kodak Research Laboratories, Eastman Kodak Company, Rochester, NY 14650-1816

jiebo.luo@kodak.com

Abstract

This paper presents a hierarchical-compositional model of human faces, as a three-layer And-Or graph to account for the structural variabilities over multiple resolutions. In the And-Or graph, an And-node represents a decomposition of certain graphical structure which expands to a set of Or-nodes with associated relations; an Or-node serves as a switch variable pointing to alternative And-nodes. Faces are then represented hierarchically: the first layer treats each face as a whole; the second layer refines the local facial parts jointly as a set of individual templates; the third layer further divides face into 15 zones and models detail facial features such as eye corners, marks or wrinkles. Transitions between the layers are realized by measuring the *minimum description length*(MDL) given the complexity of an input face image. Diverse face representations are formed by drawing from dictionaries of global faces, parts and skin detail features. A sketch captures the most informative part of a face in a much more concise and potentially robust representation. However, generating good facial sketches is extremely challenging because of the rich facial details and large structural variations, especially in the high-resolution images. The representing power of our generative model is demostrated by reconstructing high-resolution face images and generating the cartoon facial sketches. Our model is useful for a wide variety of applications, including recognition, non-photorealisitc rendering, super-resolution, and low-bit rate face coding.

I. Introduction

A. Motivation

Human faces have been extensively studied in vision and graphics for a wide range of tasks from detection[30], [35], recognition[12], [14], [22], [38], [27], tracking[32], expression[26], [34], animation[13], [29], to non-photorealistic rendering [3], [15], [25], [33], with both the discriminative[3], [13], [28] and generative models[6], [11], [13], [22], [27]. Most existing models were designed only for certain image scale and mainly aimed at faces of small or medium resolutions. These models, though successful in their own problem domains, unfortunately do not capture the rich facial details that appear on the high-resolution or highly-detail (especially aged) faces. These details are very useful for identification and extremely important for generating vivid facial sketches. Furthermore, in addition to the *geometric* and *photometric* variabilities, the *structural* variations are also widely observed for human faces across different expressions, genders, ages races (see Figure 1(a)) and over multi-scales (see Figure 1(b)) but rarely addressed comprehensively by the existing methods. Such variations include the structure transforms of facial parts in extreme expressions (e.g., scream or wink), and the appearance of new facial features (e.g., wrinkles and marks) due to aging and scale transition. To overcome the limitations of existing models, we find it necessary to introduce a flexible multi-resolution representation of human faces, which can capture fine facial details and account for large structural variations.

Fig. 1. Face over different (a) expressions, genders, ages, races, and (b) scales.

B. Overview of a layered, composite, deformable model

Faces may experience abrupt structural transforms during continuous changes of image scales or resolutions. Imagine a person walking towards the camera from a distance: at first the face image is so small and blurry that the whole face can be merely recognized; as the person approaches, the image becomes bigger and clearer so that the individual facial parts can be recognized; when the person is very close, the image is clear enough that all fine facial details such as the marks or wrinkles are visible. We thus built a three-layer representation for faces of *low*, *medium*, and *high* resolutions respectively as shown in Figure 2.

1) *face layer*, where faces are represented as a whole by PCA models[22], [27].
2) *part layer*, where the elements are templates of local facial parts plus the rest skin region. Each part is represented individually and constrained by other parts.
3) *sketch layer*, where the elements are image primitives. A face is divided into 16 zones. Six zones further decompose the local parts into sub-graphs of patches — transformed image primitives. Another ten zones, shaped by the local parts, also represent the discovered skin features (e.g., marks or wrinkles) as sub-graphs of patches.

According to the scale/resolution transition of input face images, elements of coarser layers expand to a sub-graph of elements in the finer layers and thus leads to structural changes. For example, a face expands to facial parts during transition from low to medium resolution, while a facial part expands to image patches during transition from medium to high resolution. On the other hand, the state transitions of facial parts can also cause structural changes like opening or closing eyes, which are widely observed in facial motions. To account for these structural variations, we formulate our representation as a three-layer And-Or graph shown in Figure 2. An And-node represents a decomposition with the constituents as a set of Or-nodes, on which the constraints of node attributes and spatial relations are defined as in a *Markov random field*

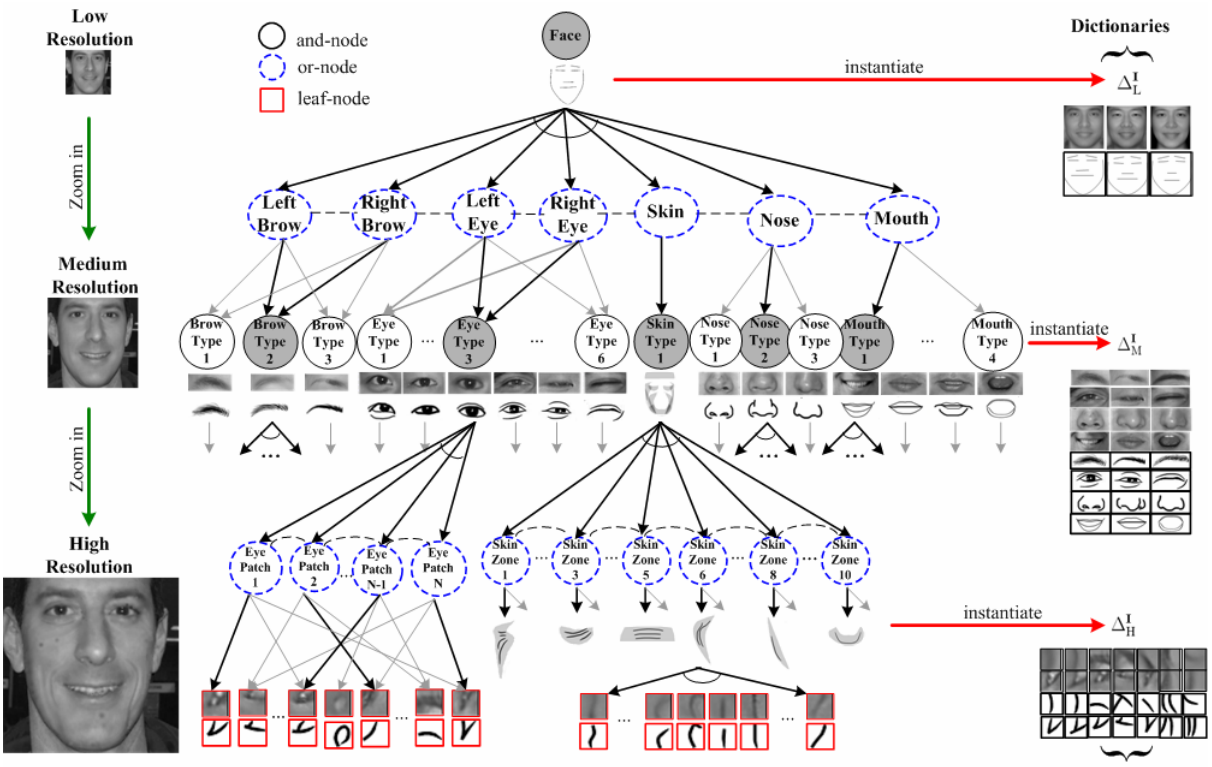


Fig. 2. An illustration of the three-layer face And-Or graph representation. The dark arrows and shadow nodes represent a composition of seven leaf-nodes $\langle BrowType2(L/R), EyeType3(L/R), SkinType1, NoseType2, MouthType1\rangle$, each being a *sub-template* at the medium resolution layer. This generates a *composite graphical template* (at the bottom) representing the specific face *configuration* with the spatial relations (context) inherited from the And-Or graph.

model. An Or-node functions as a switch variable in the *decision trees*, pointing to alternative composite deformable templates that are And-nodes. The selection/transition is then realized by applying a set of *stochastic grammars* and assigning values to the switch variables. A leaf-node is an instantiation of the corresponding And-node, which is associated with an *active appearance model* (AAM) to allow geometric and photometric variations.

In our model, parsing a face image is equivalent to finding a valid traversal from the root node of the And-Or graph. Following the thick arrows to select appropriate templates in Figure 2, we parse the input face image and arrive in a configuration as in Figure 3. In essence, an And-Or graph is essentially a set of multi-scale faces of all structural, geometric and photometric

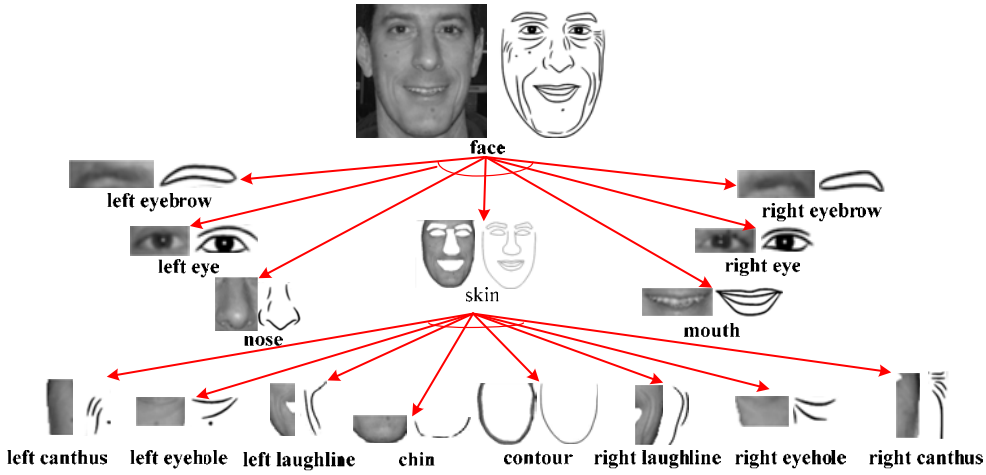


Fig. 3. A face is parsed into the configuration of the local parts and skin zones, of which both the images and symbolic representations are shown. Parts and skin zones can be further parsed into sub-graphs of image primitives.

variations. We construct the And-Or graph by maximizing the likelihood of parameters given a set of annotated face parsing graphs. The parsing of a new face image is then conducted in a coarse to fine fashion using *maximum a-posteriori* (MAP) formulation. To balance the representation power and model complexity, we adopt *minimum description length* (MDL) as the criterion to decide transitions between the graph layers. These transitions are based on not only the scales/resolutions of input face images, but also the accuracy requirement of specific tasks, e.g., low-resolution for detection, medium-resolution for recognition and high-resolution for non-photorealistic rendering.

C. Related work

In computer vision, numerous methods had been proposed to model human faces. Zhao et al suggested [38] that following the psychology study of how human use holistic and local features, existing methods can be categorized as (1) *global* [5], [6], [11], [13], [27], [29], (2) *feature-based (structural)* [8], [14], [28], [30], [31], [37], and (3) *hybrid* [12], [22] methods. Early holistic approaches[11], [27] used intensity pattern of the whole face as input and modeled

the photometric variation by linear combination of the *eigenfaces*. These *PCA models* cannot efficiently account for the geometric deformation and require images to be well aligned. Some later work separately modeled the shape and texture components of faces, e.g., the *Active Appearance Models*(AAM)[6], [32] and *Morphable Models*[13], [29]. Although these well-known methods captured some geometric and photometric variations, they are limited from handling large-scale structural variations due to the linear assumption and fixed topology. To relax the global constraint, some component-based/structural models were presented, including the *Pictorial Model*[8], *Deformable Templates*[37], *Constellation Model*[31], and *Fragment-based Model*[28]. These models first decompose faces into parts in supervised or unsupervised manners, then the intensity patterns of parts are modeled individually and the spatial relations among parts were mdoeled jointly. In addition, there are some hybrid methods [12], [22], which corporate the global and local information to achieve better results. However, in spite of the greater structural flexibility over the global methods, these models have their own limitations: (1) in contrast to the hierarchical transforms that we observed during the scale/resolution changes of face images, the structures of these models are flat and without scale transitions to account for the emergence of new features (e.g.,marks or wrinkles); (2) the topologies of these models are fixed and cannot account for structural changes caused by state transitions of the parts (e.g.,opening or closing eyes); and (3) the relations among parts are usually modeled by global Gaussian or pair-wise Gaussians and therefore the flexibilities are limited.

To model the scale variabilities, some researchers construct a Gaussian/Laplacian pyramid from the input image [17] and encode images at multiple resolutions. Others model each object as one point in the high-dimensional feature space, and increase the dimension to match the augmented complexity[18]. Both methods are inefficient and inadequate for human faces where dramatic variabilities exhibited, due to the absence of feature semantics and lack of

structural flexibility. We thus call for meaningful features that are specially designed for different scales/resolutions. In any case, constraints and relations on these features shall be enforced to form valid configurations while still maintaining considerable (structural/geometric/photometric) flexibilities. Ullman et al proposed *Intermediate Complexity*[28] as a criterion for selecting the most informative features. Their learned image fragments of various sizes and resolutions incidentally support our use of the three-layer dictionary: *faces, parts, primitives*. Similar to the AAM models, each element in our dictionary is governed by a number of landmark points to allow more geometric and photometric variabilities, where the landmark number is determined by complexity of the element. For each part (e.g., mouth), we allow selecting from a mixture of elements (e.g., open or closed mouth) and enforce the structural flexibility during state transitions. In addition, a coarse element expands to a sub-graph of finer elements and accounts for the structural change during scale transitions. The selections and expansions are then implemented using the And-Or graph model. While the original And-Or graph was introduced by Pearl as an AI search algorithm[20](1984), our model is more similar to some recent works by Chen et al[4] and Zhu et al[40]. The And-Or graph that we use is shown to be equivalent to an *Context Sensitive Grammar*(CSG)[24], which integrates the *Stochastic Context Free Grammar*(SCFG)[9] and *Markov Random Field*(MRF)[39] models.

With the ability to represent large structural variations and capture rich facial details, our model facilitates the generation of facial sketches for face recognition[34] and non-photorealistic rendering[15], [33]. Supported by psychology studies[2], it is known that sketch captures the most informative part of an object, in a much more concise and potentially robust representation (e.g., for face caricaturing, recognition or editing). Related work includes [25] and [3]. The former renders facial sketches similar to high-pass filtered images by combining linear *eigen-sketches*, and does not provide any high-level description of the face. Constrained on an *Active*

Shape Model(ASM)[5], the latter generates facial sketches by collecting local evidences from artistic drawings in the training set, and lack of structural variations and facial details.

D. Our contributions and organization

We present a hierarchical compositional graph model for representing faces at multiple resolutions (low, medium, and high) and large variations (structural,geometric,photometric). Our model parses the input face images of given resolutions by traversing the constructed And-Or graph and drawing from the multi-resolution template dictionaries. The traversals are guided by the *stochastic grammars* (SG) and *minimum description length* (MDL) criterion. Our hierarchical-compositional model, powered by the stochastic grammars, has been shown to help reconstruct diverse high resolution face images with rich details, and facilitate the generation of meaningful sketches for cartoon rendering. This model is useful for other applications, including recognition, non-photorealisitc rendering, super-resolution, and low-bit face coding.

In the remainder of the paper, we first formulate the face modeling problem as constructing a three-layer And-Or graph model in Section II. In Section III, we define the probabilities on the And-Or graph model and learn the model parameters. Section IV introduces the Bayesian inference algorithm and the scale transition process. Finally the experimental results on reconstructing and sketching are reported in Section V.

II. Composite Template Model for Representing Face Variability

In the following section, we first introduce the And-Or graph with a three-layer face representation as example. Then we follow with the details of each layer.

A. Introduction to Face And-Or Graph

And-Or graph was originally introduced in [20] and revisited in some recent work[4], [40]. In this paper, we adapted it to represent the composite deformable templates of human faces

over multiple scales, as showed in Figure 2. The And-Or graph is formalized as a 5-tuple.

$$\mathcal{G}_{\text{and}-\text{or}} =< \mathcal{S}, V_N, V_T, \mathcal{R}, \mathcal{P} > \quad (1)$$

1 *Root node* $\mathcal{S}$ denotes the human face category, the *Face* node at the top of Figure 2, from which the face instances of all variations are derived.

2 *Non-terminal nodes* $V_N = V^{\text{and}} \cup V^{\text{or}}$ include a set of And-nodes and a set of Or-nodes. The And-nodes $\{u : u \in V^{\text{and}}\}$ are shown by solid circles in Figure 2. Each And-node is a composite template, which expands to a set of Or-nodes according to the image complexity of input faces. The Or-nodes $\{v : v \in V^{Or}\}$ are indicated by dash ellipses in Figure 2. Each Or-node is a switch variable pointing to a number of alternative composite templates known as And-nodes. The dark arrows pointing from Or-nodes indicate the templates that were actually selected in parsing. Both the expansions of And-nodes and selections on Or-nodes are guided by a set of defined *Stochastic Context Sensitive Grammars* (SCSG).

3 *Terminal nodes*, known as Leaf-nodes, are a set of multi-resolution deformable templates governed by various number of landmark points to allow geometric and photometric variations, while the topologies are fixed as traditional deformable templates. Leaf-nodes are essentially the instantiations of And-nodes where no further expansions available. Examples of the Leaf-nodes are shown in Figure 2, which are templates of faces, parts and image primitives (e.g., edgelets, junctions or blobs) in low, medium and high resolutions respectively. For each template, both its intensity and symbolic representations are kept in the dictionaries, where the latter is essentially strokes linked by landmark points.

4 $\mathcal{R} = \{r_1, r_2, ..., r_{N(R)}\}$ represents a set of pairwise relations defined on the edge between two graph nodes $\{(v_i, v_j) : v_i, v_j \in V_T \cup V_N\}$. Each relation is a function of the attributes on two nodes $\{r_a = \psi^a(v_i, v_j) : a = 1, .., N(R)\}$, serving as a statistical constraint. Our defined relations include *center distance*, *size ratio*, *relative angle*, *closeness of bonding*

points and *appearance similarity*. Based on the nodes on which they are defined, relations are categorized into two types. One type is vertically defined on the And-nodes and the Or-nodes that they expand to (black arrows in Figure 2), maintaining the geometric and photometric consistency between parent and children nodes. For example, the appearance of a medium-resolution template shall resemble the composition of its high-resolution sub-templates. Another type is defined horizontally on the Or-nodes of the same layer (dash curves in Figure 2), keeping the spatial configurations valid. For example, the two eyes shall be symmetric and the nose shall be placed above the mouth. The horizontal relations are inheritable through the vertical relations. In other words, the Or-nodes expanded from one And-node are implicitly correlated to the Or-nodes derived from another And-node, through their parents — the And-nodes. We thus avoided designing explicit relations between every two graph nodes in the same layer, which usually leads to over-complicated model and computational inefficiency. In fact, we tend to assume that most of the parallel nodes are conditionally independent given their parents.

5 $\mathcal{P}$ is the probability model defined on the graph structure. As the And-Or graph embeds the MRF in a SCFG, the probabilities from both formulations are adopted.

Traversing from the root node of an And-Or graph to leaf-nodes, a finite set of all valid face configurations *Configurations* $\Sigma = \{g_1, g_2, .., g_M\}$ can be generated. Each of these valid traversals are called *parsing graphs*. Essentially, the And-Or graph stands for a set of multi-resolution face instances with all possible structural, geometric and photometric variations. A parsed example/configuration of the input face image is shown in Figure 3.

B. Three-Layer Face Representation

Given an input face image, the parsing process is trigged at the root node and continue in coarse-to-fine fashion, until the best (sufficient yet compact according to the resolution)

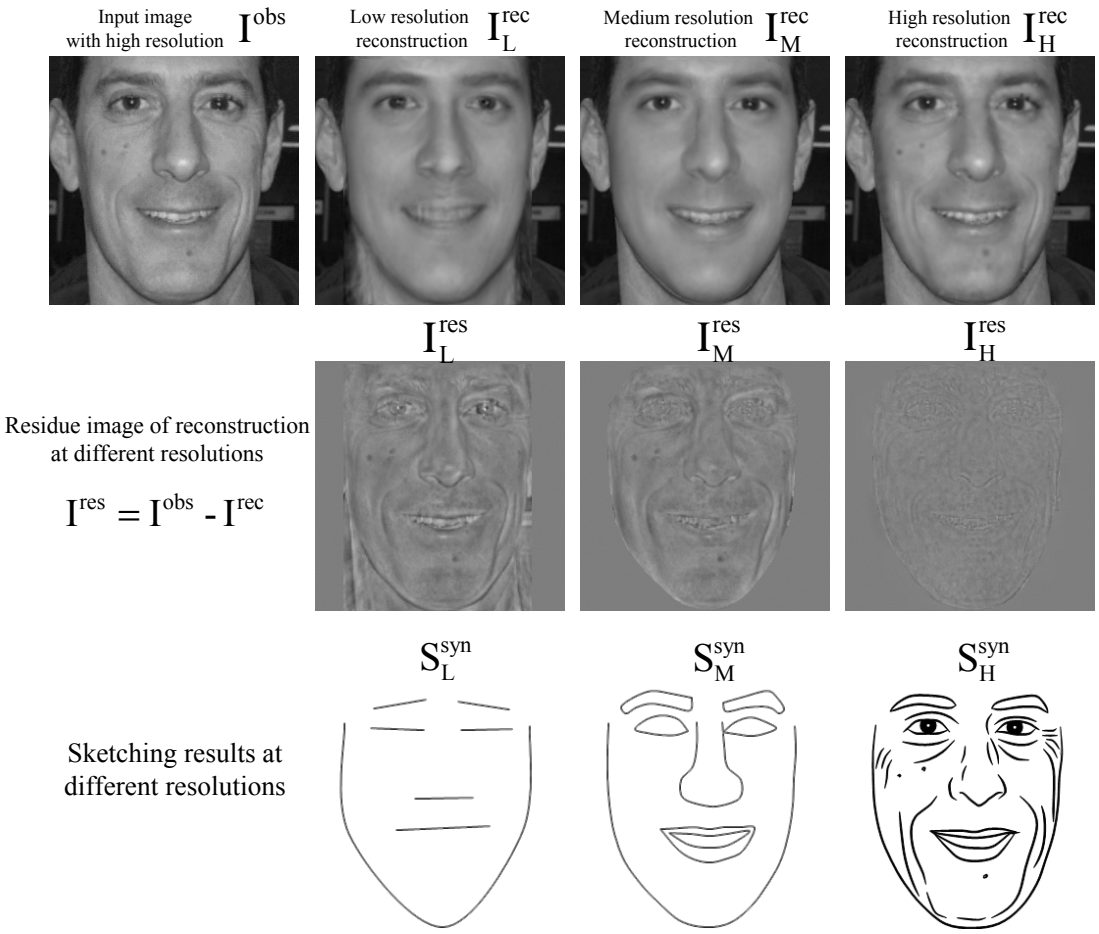
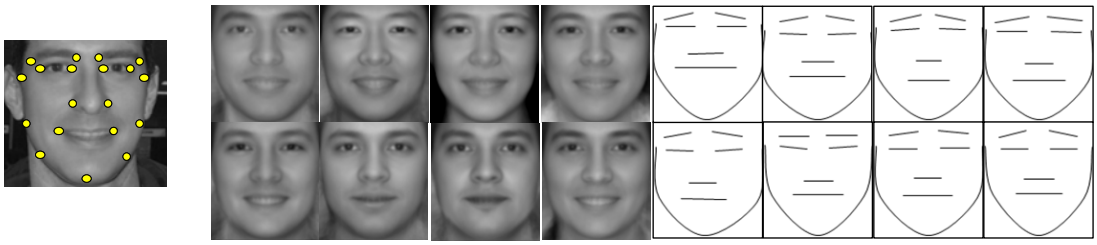


Fig. 4. Face high resolution image $\mathbf{I}^{\mathrm{obs}}$ of 256×256 pixels is reconstructed by the And-Or graph model in coarse-to-fine. The first row shows three reconstructed images $\mathbf{I}^{\mathrm{rec}}_{\mathrm{L}}$, $\mathbf{I}^{\mathrm{rec}}_{\mathrm{M}}$, $\mathbf{I}^{\mathrm{rec}}_{\mathrm{H}}$ in low, medium and high resolution respectively. $\mathbf{I}^{\mathrm{rec}}_{\mathrm{L}}$ is reconstructed by the low-resolution layer, and the facial components like eyes, nose and mouth are refined in $\mathbf{I}_{\mathrm{M}}$ with medium-resolution layer. The skin marks and wrinkles appear in $\mathbf{I}^{\mathrm{rec}}_{\mathrm{H}}$ after adding the high-resolution layer. The residue images are shown in the second row. The third row shows the sketch representations of the face with increasing complexity.

reconstruction is achieved. Figure 4 showed the input face images as well as the reconstructions at various resolution levels. In the transitions from *low resolution* to *medium resolution* and from *medium resolution* to *high resolution*, we see that more and more facial details being captured and the residue being diminished. In designing the type of representing features for certain layers, we resorted to the human intuition and decided on holistic face templates for low-resolution layer, facial component templates (eyes, nose, mouth, etc.) for medium-resolution layer, and



(a) (b)

Fig. 5. (a) Face template with 17 landmark points. (b) The first 8 PCs (plus mean) in the dictionary $\Delta_{\mathrm{L}}^{\mathbf{I}}$.

image primitives like edgelets, junctions or blobs for high-resolution layer. The *Intermediate Complexity* fragments proposed in [28] is probably regarded as the circumstantial evidence.

In the *Low-resolution layer*, we adopted the well-known Active Appearance Model (AAM) [6] on modeling the holistic face templates. A number of landmark points are defined to describe the shape/geometric deformation, while the normalized (according to mean shape computed from training set) image is used to describe the texture/phtometric pattern. The idea is to model the geometric and photometric information separately to allow more variations. Since the structures of low resolution faces are generally simple, only 17 landmark points are (manually) labelled at eye corners, nose wings, mouth corners and on face contour, as shown in Figure 5(a). Another convenient assumption was made that all (frontal) face templates in low resolution layer share the same (fixed) structure. From the training set (face images of 64×64 pixels), a set of shape vectors (landmark point coordinates) $\{x_1, x_2, ..., x_M\}$ and the corresponding texture vectors (normalized image pixels) $\{g_1, g_2, ..., g_M\}$ are collected to build PCA models separately. The principal components of the shape PCA and the texture PCA then form a dictionary in low resolution layer as shown in Figure 5(b)

$$\Delta_{\mathrm{L}}^{\mathbf{I}} = \{\mathbf{B}_{\mathrm{L}}^{\mathrm{geo}},\ \mathbf{B}_{\mathrm{L}}^{\mathrm{pht}}\} \tag{2}$$

Let x and g denote the normalized shape and texture vectors of an input low resolution face image g_{im}, we have $x = \overline{x} + Q_x c_x$ and $g = \overline{g} + Q_g c_g$. Here, $\overline{x}$, $\overline{g}$ are the mean shape and mean

texture, Q_x, Q_g are matrices with columns as the orthogonal bases from $\mathbf{B}_{\mathrm{L}}^{\mathrm{geo}}$, $\mathbf{B}_{\mathrm{L}}^{\mathrm{pht}}$, and c_x, c_g are the PCA coefficients. The final shape is then generated by a similarity transformation $X = f_x(x)$, where f_x has parameters of *rotation* θ, *translation* t_x, t_y and *scale* s_x, s_y. Similarly, the final texture is generated by $g_m = (u_1 + 1)g + u_2\mathbf{1}$, where u_1 and u_2 stand for the *contrast* and *brightness*. To reconstruct the input image g_{im}, we transform the final texture g_m by a warping function $f_w(g_m)$, where f_w has parameters of the mean shape $\overline{x}$ (source) and the final shape X (target). We thus have the hidden variables in the low-resolution layer.

$$W_{\mathrm{L}} = (c_x, c_g, \theta, t_x, t_y, s_x, s_y, u_1, u_2) \tag{3}$$

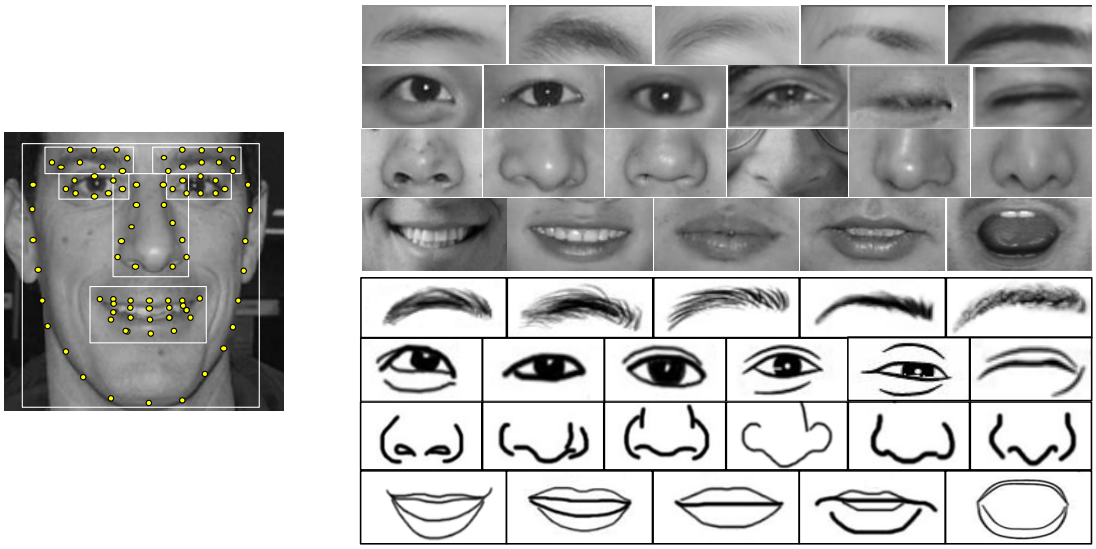
An input low resolution face image $\mathbf{I}_{\mathrm{L}}^{\mathrm{obs}}$ of 64×64 pixels is then reconstructed as in Figure 4

$$\mathbf{I}_{\mathrm{L}}^{\mathrm{obs}} = \mathbf{I}_{\mathrm{L}}^{\mathrm{rec}}(W_{\mathrm{L}}; \Delta_{\mathrm{L}}^{\mathbf{I}}) + \mathbf{I}_{\mathrm{L}}^{\mathrm{res}} \tag{4}$$

In the *Medium-resolution layer*, a face is composed of six local facial components (eyes, eyebrows, nose and mouth) and the rest skin part, which are expanded from the face node in low-resolution layer as in Figure 2. Figure 6(a) shows the partition of a medium resolution face and the landmark points defined on its local parts. Let a medium size lattice Λ_{M} denote a face of medium resolution, and $\Lambda_i^{\mathrm{cp}}, i = 1, ..., 6$ denote the six facial components, then

$$\cup_{i=1}^{6} \Lambda_i^{\mathrm{cp}} = \Lambda_{\mathrm{cp}} \subset \Lambda_{\mathrm{M}} \tag{5}$$

Each Δ_{cp}^i is an Or-node in the And-Or graph, pointing to a number of alternative deformable templates that represent various modes/types, such as closed, open or wide-open mouths. By examining our training data (AR[19],FERET[23],LHI[36] and other collections), we subjectively categorized the local facial components into three types of eyebrows, five types of eyes, three types of nose and four types of mouth. Each one type of the facial components itself is an And-node, which is implemented as a constrained AAM model [6]. Therefore a total number



(a) (b)

Fig. 6. (a) The locations of facial components and the control points defined on them. (b) Dictionary $\Delta_{\mathrm{M}}^{\mathbf{I}}$ of facial components and their artistic sketches drawn according to the control points. The examples in the same row are of same type but different modes, and selected by the Or-nodes according to grammar rules.

of $3+5+3+4=15$ AAM models are trained from the manually labelled medium resolution face images. The dictionary of these models is shown in Figure 6(b).

$$\Delta_{\mathrm{M}}^{\mathbf{I}} = \{\mathbf{B}_{\mathrm{cp},j}^{\mathrm{geo}},\ \mathbf{B}_{\mathrm{cp},j}^{\mathrm{pht}}, j = 1, .., 15\} \tag{6}$$

where $\mathbf{B}_{\mathrm{cp},j}^{\mathrm{geo}}$ and $\mathbf{B}_{\mathrm{cp},j}^{\mathrm{pht}}$ are the geometric and photometric bases of the j^{th} model. The hidden variables in this layer are the union of variables from the local AAM models.

$$W_{\mathrm{M}} = \{(\ell_i, c_x^{\ell_i}, c_g^{\ell_i}, \theta^{\ell_i}, t_x^{\ell_i}, t_y^{\ell_i}, s_x^{\ell_i}, s_y^{\ell_i}, u_1^{\ell_i}, u_2^{\ell_i})\}_{i=1}^{6} \tag{7}$$

where $\ell_i = \{1, ..., 15\}$ is the index of the selected AAM model — switch variable for the i^{th} Or-node. The Λ_{cp} is then reconstructed as the union of reconstruction of $\Lambda_i^{\mathrm{cp}}, i = 1, ..., 6$.

$$\mathbf{I}_{\mathrm{cp}}^{\mathrm{rec}}(W_{\mathrm{M}}; \Delta_{\mathrm{M}}^{\mathbf{I}}) = \cup_{i=1}^{6} \mathbf{I}_{\mathrm{cp},j}^{\mathrm{rec}}$$

An input medium resolution face image $\mathbf{I}_{\mathrm{M}}^{\mathrm{obs}}$ of 128×128 pixels is then reconstructed as in Figure 4. The rest skin pixels $\Lambda_{\mathrm{ncp}} = \Lambda_{\mathrm{M}} - \Lambda_{\mathrm{cp}}$ are up-sampled from $\mathbf{I}_{\mathrm{L}}^{\mathrm{rec}}$ with boundary

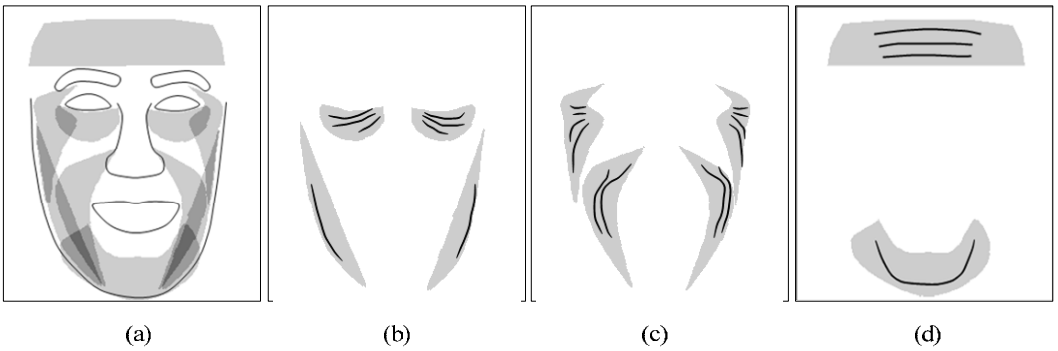


Fig. 7. (a). 16 facial zones for high-resolution face features. Six zones, indicated by solid shapes, are to refine the eyebrows, eyes, nose and mouth. Another ten zones, indicated by shaded regions, are where the skin features like marks or wrinkles occur. These zones are localized by shapes of the facial parts computed in the medium-resolution layer. (b-c-d) typical wrinkles (curves) patterns of the ten skin zones. To reliably detect these subtle features needs strong prior models and global context.

conditions of $\Lambda_{\rm cp}$.

$$\mathbf{I}_{\rm M}^{\rm rec}(x,y) = \begin{cases} \mathbf{I}_{\rm cp}^{\rm rec}(x,y) & \text{if } (x,y) \in \Lambda_{\rm cp} \\ \mathbf{I}_{\rm L}^{\rm rec}(x/2,y/2) & \text{if } (x,y) \in \Lambda_{\rm ncp} \end{cases} \tag{8}$$

In the *High-resolution layer*, much more subtle features are exposed as we can see from Figure 4. Thus the medium-resolution layer representations is further decomposed into sub-graphs of sketchable [10] image primitives (edgelets, junctions, blobs, etc.), to capture the high resolution details such as eye-corners, nose-tip, wrinkles and marks. Intuitively, an input face was divided into 16 facial zones, shown in Figure 7, according to the shapes of facial components and face contour reconstructed in medium-resolution layer. The first six zones refine the local facial components inherited from medium-resolution layer, and the 10 new zones are introduced to cover the features that appear on rest of the skin (forehead, canthus, eyehole, laughline, cheek and chin). We called the former *structural* zones since they are very much dependent on the existing medium-resolution layer facial components, while we called the latter *free* zones since the occurrence and pattern of features within them are rather random. Examples

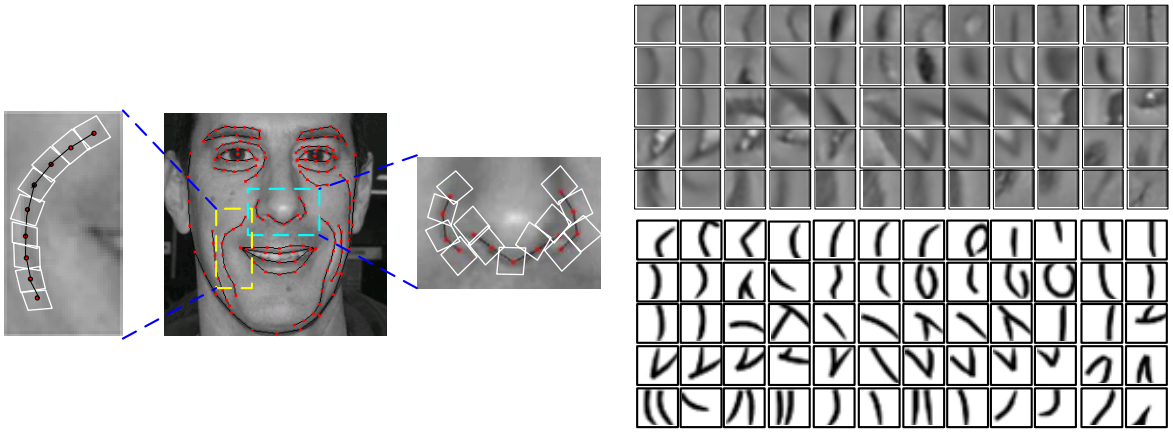


Fig. 8. (a) Refinement of the nose and a "smile fold" by sketch primitives, which are represented by small rectangles. (b) Dictionary $\Delta_{\mathrm{H}}^{\mathbf{I}}$ of sketch primitives and their corresponding sketch strokes.

of a *structural* zone (nose) and a *free* zone (laughline) are shown in Figure 8(a). Each of the rectangles represents an image primitive with (small) geometric and photometric deformations. In training stage, both the *structural* and *free* zones of the high resolution face images are manually sketched, then a huge number of image patches of certain size (e.g. 11×11 pixels) are collected along the sketches, from which the image primitives are learned through clustering. Figure 8(b) shows the dictionary of the learned image primitives and their corresponding sketch representations. Note that we defined a small number ($2 \sim 4$) of control points for each sketch patch, to connect with neighboring patches properly and generate smooth face sketches.

$$\Delta_{\mathrm{H}}^{\mathbf{I}} = \{\mathbf{B}_{\mathrm{H},i}^{\mathrm{geo}},\ \mathbf{B}_{\mathrm{H},i}^{\mathrm{pht}}, i = 1, .., N\} \tag{9}$$

where N is the number of different image primitives, which was decided empirically. The hidden variables of this layer are

$$W_{\mathrm{H}} = (K, \{(\ell_k, \theta^{\ell_k}, t_x^{\ell_k}, t_y^{\ell_k}, s_x^{\ell_k}, s_y^{\ell_k}, u_1^{\ell_k}, u_2^{\ell_k})\}_{k=1}^{K}) \tag{10}$$

where K is the total number of image patches, ℓ_k is the primitive type, and θ^{ℓ_k}, $(t_x^{\ell_k}, t_y^{\ell_k})$, $(s_x^{\ell_k}, s_y^{\ell_k})$, $u_1^{\ell_k}$, $u_2^{\ell_k}$ are respectively the *rotation*, *translation*, *scale*, *contrast* and *brightness*. Let

$\Lambda_{\rm H}$ be an input high resolution face image of 256×256 pixels, its sketchable part $\Lambda_{\rm sk}$ is covered by transformed image primitives and form $\mathbf{I}^{\rm rec}_{\rm sk}(W_{\rm H}; \Delta^{\mathbf{I}}_{\rm H})$. The rest non-sketchable part $\Lambda_{\rm nsk} = \Lambda_{\rm H} - \Lambda_{\rm sk}$ is up-sampled from $\mathbf{I}^{\rm rec}_{\rm M}$ with boundary conditions of $\Lambda_{\rm sk}$.

$$\mathbf{I}^{\rm rec}_{\rm H}(x,y) = \begin{cases} \mathbf{I}^{\rm rec}_{\rm sk}(x,y) & \text{if } (x,y) \in \Lambda_{\rm sk} \\ \mathbf{I}^{\rm rec}_{\rm M}(x/2, y/2) & \text{if } (x,y) \in \Lambda_{\rm nsk} \end{cases} \tag{11}$$

Our sketch representation capture more prolific facial details than the state-of-art face sketch method [3] and expression classification method [26].

III. Learning Probabilistic Models on the And-Or Graph

A. Defining the Probabilities

Let $\mathcal{P}$ be the probability model defined over the And-Or graph (see Section II(A)), we argue that $\mathcal{P}$ corresponds to a *probabilistic context-sensitive grammar* (PCSG), which embeds an *Markov random fields* model (MRF) in a *stochastic context-free grammar* tree (SCFG). To show this, we first define a *parsing graph* g as a valid traversal of an And-Or graph $\mathcal{G}$. It consists of a set of traversed nodes $V = \{v_1, v_2, ..., v_{N(v)}\} \in V_N \cup V_T$ and a set of observed relations $R \in \mathcal{R}$. The probability of a graph is then denoted as $p(g; \Theta)$.

As one component of $p(g; \Theta)$, the SCFG (parsing tree) can be expressed as the product of probabilities of all switch variables $T = \{\omega_1, \omega_2, ..., \omega_{N(\omega)}\}$ on the visited Or-nodes.

$$p(T) = \prod_{\omega_i \in T} p_i(\omega_i) \tag{12}$$

Another component, the MRF is probability on the configuration C of resulting nodes. It is written in terms of pairwise energies on two nodes and constraints on each single node.

$$p(C) = \frac{1}{Z} \exp\{- \sum_{v_i \in V} \alpha_i \phi(v_i) - \sum_{<v_i, v_j> \in E} \beta_{ij} \psi(v_i, v_j)\} \tag{13}$$

where E is the set of node pairs on which relations are defined, and ϕ and ψ are respectively the functions of single nodes and node pairs. Given that T is the parsing tree of g, we would

like to derive $p(g)$ by minimizing the KL divergence of $p(g)$ and $p(T)$, subject to constraints that expectations of the energy functions shall match what we observed from training data.

$$p^* = \arg\min \textstyle\sum_g p(g) \log \frac{p(g)}{p(T)}$$
$$\text{subject to} \begin{cases} \mathbf{E}_{p(g)}[\phi^{(a)}(v_i)] = \mu_i, a = 1, 2, ..., N(\phi) \\ \mathbf{E}_{p(g)}[\psi^{(b)}(v_i, v_j)] = \mu_{ij}, b = 1, 2, .., N(\psi) \end{cases} \tag{14}$$

where $N(\phi)$ and $N(\psi)$ are respectively the number of singleton constraints and pairwise constraints. Solving this constrained optimization by Lagrange multipliers yields:

$$p(g; \Theta) = \frac{1}{Z(\Theta)} p(T) \exp\{- \sum_{v_i \in V} \sum_{a=1}^{N(\phi)} \alpha_i^{(a)} \phi^{(a)}(v_i) - \sum_{<v_i, v_j> \in E} \sum_{b=1}^{N(\psi)} \beta_{ij}^{(b)} \psi^{(b)}(v_i, v_j)\} \tag{15}$$

where $\Theta = (\theta, \alpha, \beta)$, θ is the parameters in $p(T)$ while α and β are Lagrange multipliers.

B. Estimating the Model Parameters

Given a set of observed parsing graphs $\hat{G} = \{g_1, g_2, ..., g_N\}$ from the training set, we can estimate parameters Θ by maximizing the log-likelihood $L(\Theta; \hat{G}) = \sum_{g_i} \log p(g_i; \Theta)$.

$$\Theta^* = \arg\max \sum_{i=1}^{N} \log p(g_i; \Theta) \tag{16}$$

Let $p(\omega_i)$ be the probability over the switch variable at an Or-node, the values that ω_i takes depend on the grammar rules we defined on the Or-node. Examples of such grammar rules in medium-layer Or-nodes are shown in Figure 9, which set a specific mode for the facial parts, such as to open an eye or to shut a mouth. Let θ_{ij} be the probability that ω_i takes value j — the jth rule, and n_{ij} be the number of times that we observed this rule, $p(T)$ is rewritten as

$$p(T) = \prod_{\omega_i \in T} \prod_{j=1}^{N(\omega_i)} \theta_{ij}^{n_{ij}} \tag{17}$$

Plug it back into $p(g; \Theta)$ and the MLE for θ is now rewritten as

$$\frac{\partial L(\Theta; \hat{G})}{\partial \theta} = -N \frac{\partial \log Z(\Theta)}{\partial \theta} - \sum_{k=1}^{N} \sum_{\omega_i \in T} \sum_{j}^{N(\omega_i)} \frac{n_{ij}^{(k)}}{\theta_{ij}} = 0$$
$$\text{subject to} \quad \sum_{j=1}^{N(\omega_i)} \theta_{ij} = 1, \ \text{ for all } \omega_i \in T \tag{18}$$

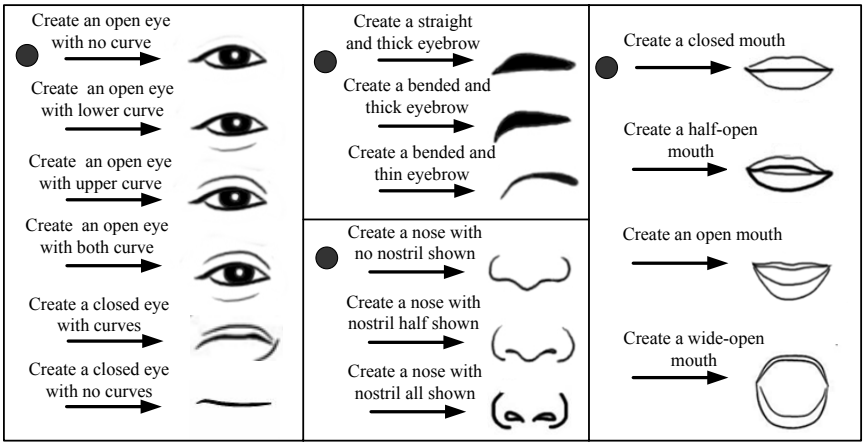


Fig. 9. Grammars defined on Or-nodes of medium-resolution layer, for switching among various composite templates.

where $n_{ij}^{(k)}$ is the n_{ij} for a specific graph g_k. Solve this with Lagrange multiplier yields

$$\hat{\theta}_{ij} = \frac{\sum_{k=1}^{N} n_{ij}^{(k)}}{N_{\omega_i} - N\frac{\partial \log Z(\Theta)}{\partial \theta} + N\frac{\partial \log Z(\Theta)}{\partial \theta}} = \frac{N_{ij}}{N_{\omega_i}} \tag{19}$$

where N_{ω_i} is the total number of times that ω_i was assigned some value in all graphs. Thus $\hat{\theta}_{ij}$ is just the frequency of rule j being applied at Or-node i observed in the training set. Sampling from the $p(T)$ enables us to generate novel parsing trees, e.g. winking and excited, that were not even seen in the training data as shown in Figure 10.

After $p(T)$ is learned, we need to derive α and β to impose the constraints among nodes. Given $\hat{G}$, we define the collection of output values from ϕ and ψ as histograms H_ϕ and H_ψ, then rewrite the energy terms in MRF as $\sum_a < \alpha^a, H_\phi^a >$ and $\sum_b < \beta^b, H_\psi^b >$. Therefore the MLE of α and β is equivalent to maximizing the entropy of $p(g;\Theta)$ subject to the constraint that the expected histograms shall match the observed histograms [39].

$$\begin{aligned} &\frac{\partial L(\Theta;\hat{G})}{\partial \alpha} = -N\frac{\partial \log Z(\Theta)}{\alpha} - \sum_a \sum_{k=1}^{N} H_\phi^{(a)}(g_k) = 0 \\ \text{subject to} \quad &\mathbf{E}_{p(g)}[H_\phi^{(a)}(g)] = \frac{1}{N}\sum_{k=1}^{N} H_\phi^{(a)}(g_k), \ \text{ for all } a; \\ &\frac{\partial L(\Theta;\hat{G})}{\partial \beta} = -N\frac{\partial \log Z(\Theta)}{\beta} - \sum_b \sum_{k=1}^{N} H_\psi^{(b)}(g_k) = 0 \end{aligned} \tag{20}$$

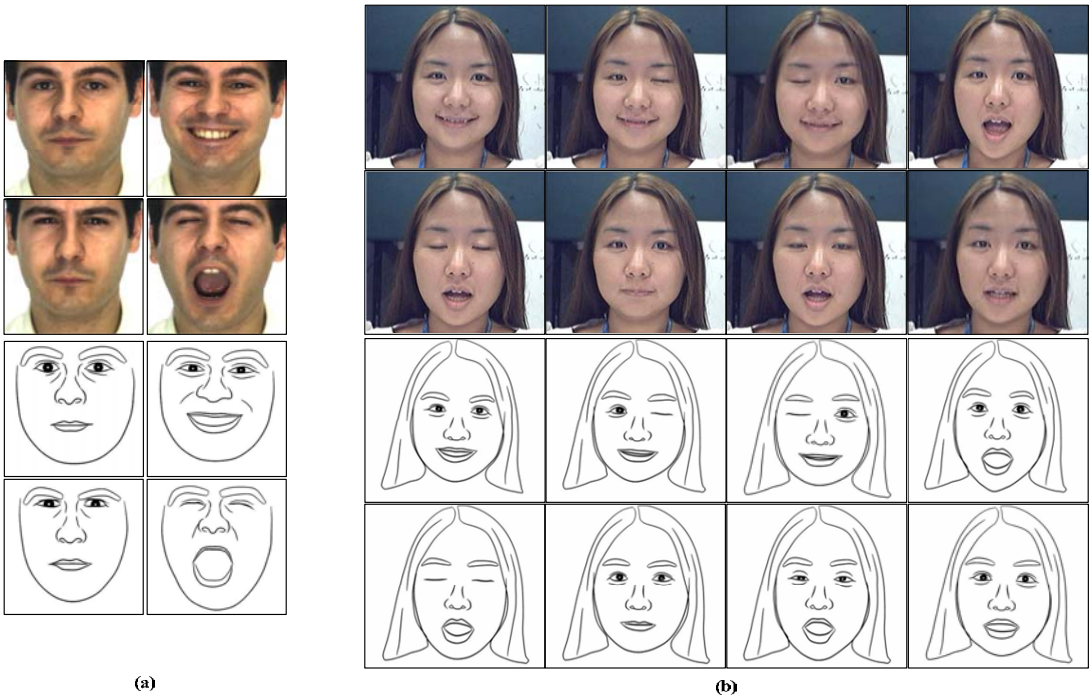


Fig. 10. Different face configurations are composed by various types of local facial components. (a) The four typical face configurations in the AR dataset as *neutral, laughing, angry and screaming*. (b) The eight novel face configurations inferred from the frames in a personal video clip. These configurations correspond to new dramatic expressions, e.g., *winking or excited*.

$$\text{subject to} \qquad \mathbf{E}_{p(g)}[H_{\psi}^{(b)}(g)] = \frac{1}{N}\sum_{k=1}^{N} H_{\psi}^{(b)}(g_k), \ \text{ for all } b \tag{21}$$

Similar to [39], we solve for α and β by iteratively updating them with

$$\frac{d\alpha}{dt} = \mathbf{E}_{p(g)}[H_{\phi}(g)] - \frac{1}{N}\sum_{k=1}^{N} H_{\phi}^{obs}(g_k) = H_{\phi}^{\text{syn}} - H_{\phi}^{\text{obs}} \tag{22}$$

$$\frac{d\beta}{dt} = \mathbf{E}_{p(g)}[H_{\psi}(g)] - \frac{1}{N}\sum_{k=1}^{N} H_{\psi}^{obs}(g_k) = H_{\psi}^{\text{syn}} - H_{\psi}^{\text{obs}} \tag{23}$$

The algorithm of learning α and β proceeds in Figure 11. The sampling results of the learning procedure are shown in Figure 12.

C. Experiment I: Sampling Faces from And-Or Graph

Once the And-Or graph of face is constructed, we can sample the generative model to provide believable human faces of different configurations and large structural variations.

Given a set of observed parsing graph $\hat{G}$ and the initial $\alpha^{(0)} = \mathbf{0}$ and $\beta^{(0)} = \mathbf{0}$.

1) Compute H_{ϕ}^{obs} and H_{ψ}^{obs} from $\hat{G}$.
2) Repeat until $|H^{obs} - H_{(t-1)}^{syn}| - |H^{obs} - H_{(t)}^{syn}| < \epsilon$, where ϵ is the prescribed threshold.
   a) Sample a set of parsing graphs G' from current $p(g;\Theta)$ and compute the synthesized histograms $H_{(t)}^{syn}$ for all defined ϕ and ψ
   b) Update α and β

   $$\alpha^{(t)} = \alpha^{(t-1)} + \eta_{\phi}(H_{\phi,(t)}^{\mathrm{syn}} - H_{\phi}^{\mathrm{obs}})$$

   $$\beta^{(t)} = \beta^{(t-1)} + \eta_{\psi}(H_{\psi,(t)}^{\mathrm{syn}} - H_{\psi}^{\mathrm{obs}})$$

   where η_{ϕ} and η_{ψ} are the step factors that are decided empirically.

Fig. 11. Algorithm for learning parameter of the MRF model.

To sample the configurations, we first learned the $p(T)$ from AR[19] dataset, in which there are four typical configurations that correspond to expressions of *neutral*, *smiling*, *angry* and *screaming* as shown in Figure 10(a). However, eight facial configurations were observed in a personal video of facial motions, which are different from the training data. These novel configurations unseen in training set, such as *winking* and *excited*, were then successfully sampled from our And-Or graph model to match the new observations as shown in Figure 10(b).

Figure 12 visualizes the learning of the MRF model in the medium layer. During this procedure, facial structures which satisfy the learned constraints are synthesized. In the early stage, the synthesized faces appeared rather random and the H^{syn} differed from the H^{obs} significantly. After the algorithm ran for a certain number (e.g., 50) of sweeps, the synthesized faces started to resemble the observed faces as the H^{syn} approximated the H^{obs}. We define ϕ as the constraints on single nodes such as the *shape prior* and *appearance prior* of AAM models, while ψ are the pairwise relations such as *center distance*, *size ratio*, *relative angle*, *closeness of bonding points* and *appearance similarity*. By using these pairwise constraints, the sampled faces accommodate larger structural variations than the global AAM models.

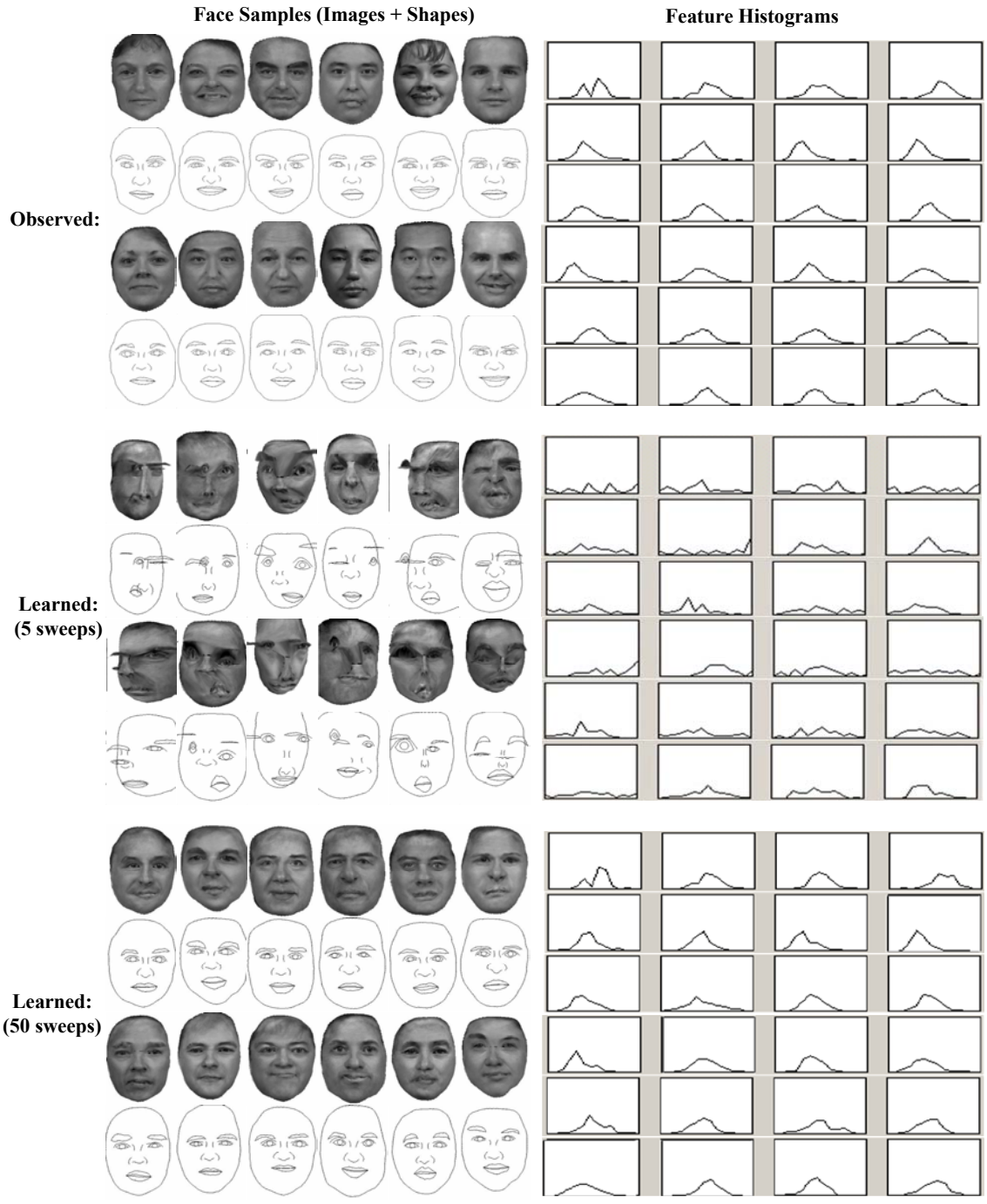


Fig. 12. Examples of observed and synthesized face samples, including images and shapes, and the feature histograms.

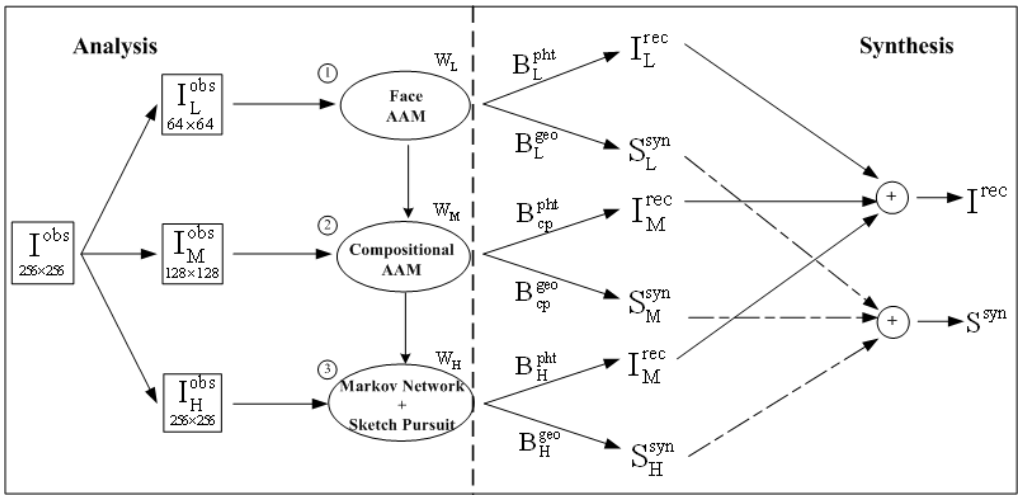


Fig. 13. The diagram of our model and algorithm. The arrows indicate the inference order. Left panel is the three layers. Right panel is the synthesis steps for both image reconstruction and sketching using the generative model.

IV. Bayesian Inference and Scale Transition

Given an input face image $\mathbf{I}^{\text{obs}}$, our goal is to determine the $W = (W_{\text{L}}, W_{\text{M}}, W_{\text{H}})$ defined in Section II(B) by maximizing the Bayesian posterior.

$$\begin{aligned}(W_{\text{L}}, W_{\text{M}}, W_{\text{H}})^* &= \arg\max p(W_{\text{L}}, W_{\text{M}}, W_{\text{H}}|\mathbf{I}^{\text{obs}}) = \arg\max p(\mathbf{I}^{\text{obs}}|W)p(W) \\ &= \arg\max p(W_{\text{H}}|W_{\text{M}}, W_{\text{L}}, \mathbf{I}^{\text{obs}})p(W_{\text{M}}|W_{\text{L}}, \mathbf{I}^{\text{obs}})p(W_{\text{L}}|\mathbf{I}^{\text{obs}}) \qquad (24)\end{aligned}$$

We notice that the parsing graph g^* for $\mathbf{I}^{\text{obs}}$ can be derived from W. For example in the medium-resolution layer, the $\{\ell_i\}$ in W_{M} represent the switch variables $\{\omega_i\}$ on the Or-nodes in g^*, while the $\{(c_x^i, c_g^i, \theta^i, t_x^i, t_y^i, s_x^i, s_y^i, u_1^i, u_2^i)\}$ in W_{M} expand the attributes of the And-nodes $\{v_i\}$ in g^*. The same analogy applies to the other layers and we have $p(W) = p(g; \Theta)$, as defined in Section III. Given an input image of certain resolution, all Leaf-nodes of the resulting parsing graph sit in the same layer — of same scale. We first build a three-layer gaussian pyramid $(\mathbf{I}_{\text{L}}^{\text{obs}}, \mathbf{I}_{\text{M}}^{\text{obs}}, \mathbf{I}_{\text{H}}^{\text{obs}})$ from the input image. Then $(W_{\text{L}}, W_{\text{M}}, W_{\text{H}})^*$ shall be gradually optimized according to the layers in coarse-to-fine as shown in Figure 13.

A. Layer 1: the low resolution AAM model

Only one Leaf-node denoting frontal faces will be derived in the low-resolution layer. We adopted the well-known AAM model [6] in learning and computing $W_{\rm L}$.

$$\begin{aligned} W_{\rm L}^* &= \arg\max p(W_{\rm L}|\mathbf{I}^{\rm obs}) = \arg\max p(\mathbf{I}_{\rm L}^{\rm obs}|W_{\rm L};\Delta_{\rm L}^{\mathbf{I}})p(W_{\rm L}) \\ &= \arg\max \exp\{-|\mathbf{I}_{\rm L}^{\rm obs}-\mathbf{I}_{\rm L}^{\rm rec}|^2/(2\sigma_{\rm L}^2) - \frac{1}{2}W_{\rm L}'(\mathbf{S}_{W_{\rm L}}^{-1})W_{\rm L}\} \end{aligned} \quad (25)$$

The first term of second row denotes the likelihood, where $\mathbf{I}_{\rm L}^{\rm rec}$ is the reconstructed low resolution layer governed by $W_{\rm L}$ and $\sigma_{\rm L}^2$ is the variance of reconstruction error learned from training data. The second term denotes the prior, where $\mathbf{S}_{W_{\rm L}}$ is the covariance matrix of $W_{\rm L}$. The optimized $W_{\rm L}^*$ can be computed efficiently by *stochastic gradient descent* [6].

B. Layer 2: the medium resolution compositional AAM model

The medium-resolution layer is inferred by maximizing posterior of $W_{\rm M}$ given $\mathbf{I}_{\rm M}^{\rm obs}$ and $W_{\rm L}^*$.

$$W_{\rm M}^* = \arg\max p(W_{\rm M}|W_{\rm L},\mathbf{I}^{\rm obs}) = \arg\max p(\mathbf{I}_{\rm M}^{\rm obs}|W_{\rm M},W_{\rm L};\Delta_{\rm M}^{\mathbf{I}},\Delta_{\rm L}^{\mathbf{I}})p(W_{\rm M}|W_{\rm L}) \quad (26)$$

The first term indicates the likelihood probability.

$$\begin{aligned} p(\mathbf{I}_{\rm M}^{\rm obs}|W_{\rm L},W_{\rm M};\Delta_{\rm M}^{\mathbf{I}},\Delta_{\rm L}^{\mathbf{I}}) &\propto \exp\{-\frac{1}{2}(\mathbf{I}_{\rm M}^{\rm obs}-\mathbf{I}_{\rm M}^{\rm rec})'\Sigma_r^{-1}(\mathbf{I}_{\rm M}^{\rm obs}-\mathbf{I}_{\rm M}^{\rm rec})\} \\ &= \exp\{-\sum_{i=1}^{6}\frac{|\mathbf{r}_{{\rm cp},i}|^2}{2\sigma_{{\rm cp},i}^2} - \frac{|\mathbf{r}_{\rm L}|^2}{2\sigma_{\rm L}^2}\} \end{aligned} \quad (27)$$

where $\{\mathbf{r}_{{\rm cp},i}\}_{i=1}^6$ denote the reconstructed residue of the pixels covered by the six facial components $\Lambda_{\rm cp}$, $\mathbf{r}_{\rm L}$ is the reconstructed residue of the rest pixels $\Lambda_{\rm ncp}$, $\{\sigma_{{\rm cp},i}^2\}_{i=1}^6$ and $\sigma_{\rm L}^2$ are the variances of errors learned from training data. The second term of the conditional prior can be factorized to three components.

$$p(W_{\rm M}|W_{\rm L}) \propto \prod_{i=1}^{6} p(\ell_i)\cdot\prod_{i=1}^{6} p(W_{\rm cp}^i|W_{\rm L})\cdot\prod_{<v_k,v_l>\in E_{\rm cp}} p(W_{\rm cp}^k,W_{\rm cp}^l) \quad (28)$$

The first component denotes the prior probability of the parsing tree as defined in Section III.

$$\prod_{i=1}^{6} p(\ell_i) \propto \prod_{i=1}^{6} \prod_{j=1}^{N(\omega_i)} \theta_{ij}^{\delta(\ell_i,j)} = \prod_{i=1}^{6} \theta_{i\ell_i} \tag{29}$$

where $\delta(.)$ is a *Delta* function and $\theta_{i\ell_i}$ is simply the frequency of that the i^{th} switch variable was assigned value ℓ_i in the training data. The second component is the singleton prior of W_{M} conditioned on W_{L} in a manner similar to the constrained AAM model[6].

$$\prod_{i=1}^{6} p(W_{\mathrm{cp}}^i | W_{\mathrm{L}}) \propto \prod_{i=1}^{6} \exp\{-{W_{\mathrm{cp}}^i}' \mathbf{S}_{W_{\mathrm{cp}}^i}^{-1} W_{\mathrm{cp}}^i - \mathbf{d}_{\mathrm{cp,L}}^i{}' \mathbf{S}_{d_i}^{-1} \mathbf{d}_{\mathrm{cp,L}}^i\} \tag{30}$$

where $\mathbf{d}_{\mathrm{cp,L}}^i$ denotes the photometric and geometric displacements between current $\hat{W_{\mathrm{cp}}^i}$ and W_{L}^*. In this paper, we actually computed the geometric displacement only and ignored the photometric displacement, although which is critical for other applications like *super-resolution*. Here $\mathbf{d}_{\mathrm{cp,L}}^i = (d_{t_x}^i, d_{t_y}^i, d_\theta^i, d_{s_x}^i, d_{s_y}^i)'$ are respectively the *center displacement*, *relative angle* and *scale ratio* between the global face template and each of the local part templates. $\mathbf{S}_{W_{\mathrm{cp}}^i}$ and $\mathbf{S}_{d_i}$ are the covariance matrix of $W_{\mathrm{cp}}^i = (c_x^{\ell_i}, c_g^{\ell_i}, \theta^{\ell_i}, t_x^{\ell_i}, t_y^{\ell_i}, s_x^{\ell_i}, s_y^{\ell_i}, u_1^{\ell_i}, u_2^{\ell_i})$ and $\mathbf{d}_{\mathrm{cp,L}}^i$. The third component addressed the pairwise constraints defined on each graph node and their neighbors, including *center distance*(ψ_{t_x}, ψ_{t_y}), *size ratio*(ψ_{s_x}, ψ_{s_y}), *relative angle*(ψ_θ), *closeness of bonding points*(ψ_{cl}) and *appearance similarity*(ψ_{sm}).

$$\prod_{<v_k,v_l>\in E_{\mathrm{cp}}} p(W_{\mathrm{cp}}^k, W_{\mathrm{cp}}^l) \propto \exp\{-\sum_{<v_k,v_l>\in E_{\mathrm{cp}}} \sum_{\psi^{(b)}\in\Psi_{kl}} \beta_{kl}^{(b)} \psi^{(b)}(v_k, v_l)\} \tag{31}$$

where E_{cp} is a set of edges that linked the nodes, $\Psi_{kl} \subseteq \{\psi_{t_x}, \psi_{t_y}, \psi_{s_x}, \psi_{s_y}, \psi_\theta, \psi_{sm}\}$ is a set of pairwise constraints defined on $\langle v_k, v_l \rangle$, and $\{\beta_{kl}^{(a)}\}$ are the potential functions. These constraints helps maintain the consistency of our graph configuration. For example, the left eye and right eye tend to be symmetric (both shape and appearance) when they are of the same mode (open/closed). However, to model all possible constraints on every two graph nodes is expensive in computation and usually unnecessary. For example, we can safely assume that the appearance of the nose and mouth of the same person is remotely relevant. In this paper, the

constraints were selected based on *minimax entropy* [39]. Figure 12 showed some examples as histograms of the output values of the chosen constraints functions.

For computational simplicity and efficiency, we approximated W_{M}^{*} in three steps. Firstly from $p(W_{\mathrm{M}}|W_{\mathrm{L}})$ we proposed a set of templates (only the geometric part) with all possible types for every local facial components. Then these proposed templates were locally diffused using pre-trained constrained AAM models [6]. Finally we resulted in a pairwise MRF of the proposed templates. For each of them, we computed the local evidences as the likelihood and parameter priors, while the compatibilities were the pairwise constraints defined above. We then introduced *belief propagation* [21] in finding the optimized W_{M}^{*}. The algorithm proceeds as in Figure 14.

Given W_{L}^{*} computed from the low-resolution layer and the medium-resolution input image $\mathbf{I}_{\mathrm{M}}^{\mathrm{obs}}$.

1) For every medium-resolution layer Or-nodes, propose a set of templates (only the geometric part) of all possible types based on W_{L}.
2) For every proposed templates, diffuse them locally using the corresponding constrained AAM models and record the reconstruction errors as likelihoods.
3) For the proposed and diffused templates, compute the local evidences as likelihoods and parameter priors, while compute the compatibilities as the aforehand defined pairwise constraints. Use belief propagation algorithm to find the optimized configuration W_{M}^{*}.

Fig. 14. Algorithm for inference of the medium-resolution layer hidden variables.

C. Layer 3: the high resolution sketch model

Similarly we made reasonable assumption that W_{H} only depends on $\mathbf{I}_{\mathrm{H}}^{\mathrm{obs}}$ and W_{M}.

$$W_{\mathrm{H}}^{*} = \arg\max p(W_{\mathrm{H}}|W_{\mathrm{M}}, \mathbf{I}_{\mathrm{H}}^{\mathrm{obs}}) = \arg\max p(W_{\mathrm{H}}^{\mathrm{fr}}|W_{\mathrm{H}}^{\mathrm{st}}, \mathbf{I}_{\mathrm{H}}^{\mathrm{obs}}) p(W_{\mathrm{H}}^{\mathrm{st}}|W_{\mathrm{M}}, \mathbf{I}_{\mathrm{H}}^{\mathrm{obs}}) \tag{32}$$

where $W_{\mathrm{H}}^{\mathrm{st}}$ and $W_{\mathrm{H}}^{\mathrm{fr}}$ are respectively the hidden variables of the *structural* and *free* zones defined in Section II(B). They are inferred sequentially in the high-resolution layer.

$W_{\mathrm{H}}^{\mathrm{st}}$ includes six facial zones (Figure 7(a)), in which the eyebrows, eyes, nose and mouth are further decomposed into subgraphs of image primitives, e.g. the nose in Figure 8(a). Once the W_{M}^{*} was computed, the modes of these local facial components are completely determined, e.g. whether the mouth is open or closed. We model the subgraph $W_{\mathrm{H}}^{\mathrm{st},i}$ of zone i as a Markov network of N_i image primitives with fixed structure.

$$p(W_{\mathrm{H}}^{\mathrm{st},i}|W_{\mathrm{cp}}^{i},\mathbf{I}_{\mathrm{H},\Lambda_i}^{\mathrm{obs}}) \propto \exp\{-\sum_{k=1}^{N_i}\frac{|\mathbf{r}_k|^2}{2\sigma_k^2}-\frac{1}{2}\mathbf{d_i}'\Sigma_{c_i}^{-1}\mathbf{d_i}-\sum_{<k,l>}\frac{1}{2}(E_{kl}^{d}(\mathrm{p}_k,\mathrm{p}_l)+E_{kl}^{a}(\mathrm{p}_k,\mathrm{p}_l))\} \quad (33)$$

where $\mathbf{I}_{\mathrm{H},\Lambda_i}^{\mathrm{obs}}$ denotes the pixels in zone i and $\{\mathrm{p}_k\}$ are the image primitives. $\mathbf{r}$ in the likelihood term denotes the reconstructed residue of p_k. $\mathbf{d_i}$ in the prior term is the center distance between $\{\mathrm{p}_k\}$ and the corresponding land mark points in W_{cp}^{i}, which serves as the global shape constraint. $\langle k,l\rangle$ denotes a pair of connected image primitives on which pairwise energies are defined: $E_{kl}^{d}(\mathrm{p}_k,\mathrm{p}_l)=|e_k-e_l|^2/\sigma_{d_{kl}}^2$ for distance between two nearest endpoints, and $E_{kl}^{a}(\mathrm{p}_k,\mathrm{p}_l)=|\sin(\theta_k-\theta_l)-\mu_{kl}^{a}|^2/\sigma_{a_{kl}}^2$ for the relative angle. $\{\sigma_k^2\}$, Σ_{c_i}, $\{\sigma_{d_{kl}}^2\}$, and $\{\mu_{kl}^{a},\sigma_{a_{kl}}^2\}$ are all learned from the training data. We sequentially maximized the posteriors of every facial zones using belief propagation similar to [16]. Experiments showed fast convergence and accurate fitting.

$$W_{\mathrm{H}}^{\mathrm{st}*}=\{W_{\mathrm{H}}^{\mathrm{st},i^*}\}_{i=1}^{6}=\arg\max\prod_{i=1}^{6}p(W_{\mathrm{H}}^{\mathrm{st},i}|W_{\mathrm{cp}}^{i},\mathbf{I}_{\mathrm{H},\Lambda_i}^{\mathrm{obs}}) \quad (34)$$

$W_{\mathrm{H}}^{\mathrm{fr}}$ includes another 10 facial zones, covering the rest of the skin regions. These zones, shown in Figure 7(b, c, d), are determined by landmark points computed from $W_{\mathrm{H}}^{\mathrm{st}}$. Similar to the *structural* zones, skin features such as wrinkles and marks in the *free* zones are also represented by subgraphs of image primitives, e.g. the laugh-line in Figure 8(a). However, the patterns of both the occurrence and distribution of these features are much more random and sometimes locally imperceptible without global context. We manually labelled the skin features in every *free* zones for a set of training images. Some "typical" curves are shown in Figure 7(b, c, d), from which the prior models were learned in favor of certain properties.

1. $p_n(N_i = n) = \sum_{i=1}^{M} \alpha_i \delta(n, i)$. N_i is the number of curves in zone i, M is the maximum number of curves, α_i are frequencies of observed curve numbers. $\sum \alpha_i = 1$.

2. $p_\ell(L_j = \ell) = \frac{\lambda_L{}^\ell e^{-\lambda_L}}{\ell!}$. L_j is the length of curve j and λ_L is specified by "typical" curves.

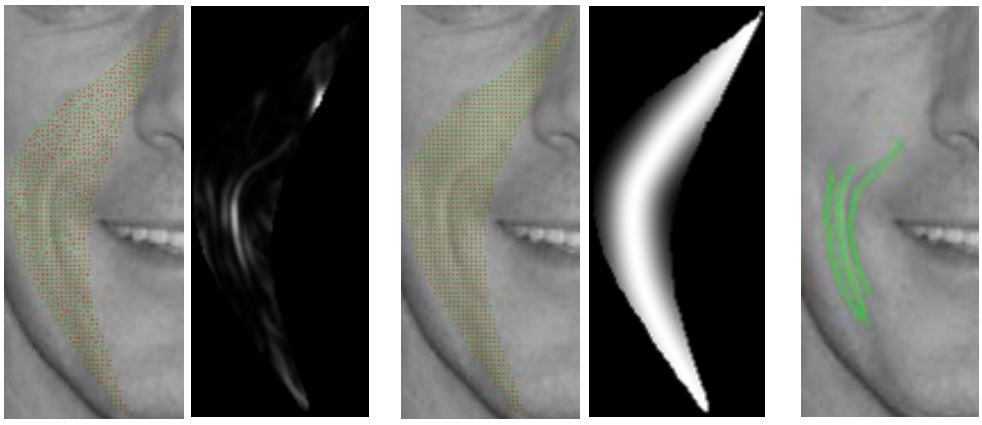
3. $p_{\rm on}({\rm on}|x, y) = p_{xy}^{\rm on}$ is the chance that point (x, y) is on a curve. $p_\theta(\theta_k|x, y) = G(\theta_k; \mu_{xy}^\theta, \sigma_{xy}^\theta)$. θ_k is the orientation of primitive k centered at (x, y). We learned $p_{xy}^{\rm on}$, μ_{xy}^θ and σ_{xy}^θ by accumulating information from nearby "typical" curves in the normalized training data (Figure 15(b)).

4. $p_{sm}({\rm p}_k, {\rm p}_l) \propto \exp\{-\frac{1}{2}(E^d + E^\theta + E^s + E^t)\}$ guarantees the *position*, *orientation*, *scale* and *intensity* consistency of two consecutive primitives ${\rm p}_k$ and ${\rm p}_l$, where $E^d = |e_k - e_l|^2/\sigma_d^2$, $E^\theta = |\sin(\theta_k - \theta_l)|^2/\sigma_\theta^2$, $E^s = |s_k - s_l|^2/\sigma_s^2$, and $E^t = |{\rm p}_k - {\rm p}_l|^2/\sigma_t^2$.

We therefore rewrote the posterior of *free* zone i which was partitioned by $W_{\rm H}^{\rm st}$.

$$p(W_{\rm H}^{{\rm fr},i}|\mathbf{I}_{{\rm H},\Lambda_i}^{\rm obs}) \propto p_n(N_i) \cdot \prod_{j=1}^{N_i} p_\ell(L_j) \cdot \prod_{k=1}^{K} p_{\rm on}({\rm on}|x_k, y_k) p_\theta(\theta_k|x_k, y_k) p_r(\mathbf{r}_k) \cdot \prod_{<k,l>} p_{sm}({\rm p}_k, {\rm p}_l) \quad (35)$$

where K is the number of primitives and $p_r(\mathbf{r}_k) = \frac{1}{Z_r} \exp\{-\frac{|\mathbf{r}_k|^2}{2\sigma_r^2}\}$ is local likelihood of primitive k. Before pursuing curves in zone i, a quick bottom-up step (edge and ridge detection, steering filters) was taken for initialization (Figure 15(a)). In step $t+1$ we proposed $W_{{\rm H},t+1}^{{\rm fr},i}$ from $W_{{\rm H},t}^{{\rm fr},i}$



(a) (b) (c)

Fig. 15. The process of curve tracking. (a) The bottom-up results of orientation and gradient magnitudes; (b) The prior of orientation field and gradient magnitudes learned from training data; (c) Curve tracking results.

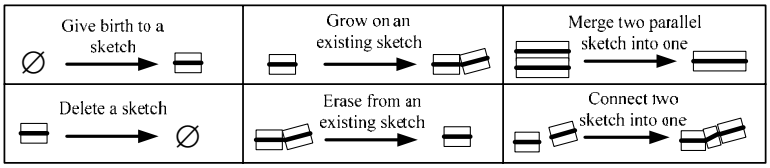


Fig. 16. Grammars used for free curve pursuit in the high-resolution layer, including *birth/death*, *split/merge*, and *connect*.

by selecting from a set of grammars (Figure 16) and computed the posterior ratio.

$$\frac{p(W^{\mathrm{fr},i}_{\mathrm{H},t+1}|\mathbf{I}^{\mathrm{obs}}_{\mathrm{H},\Lambda_i})}{p(W^{\mathrm{fr},i}_{\mathrm{H},t}|\mathbf{I}^{\mathrm{obs}}_{\mathrm{H},\Lambda_i})} = \theta \tag{36}$$

We choose the grammar that gives the greatest $\theta > 1$. If $\theta \leq 1$ for all grammars, the pursuit stops. The algorithm of curve pursuit proceeds in Figure 17, and results are shown in Figure 20. Gabor filters of various scales are used in capturing other features like marks and specularities.

Given the high-resolution input image $\mathbf{I}^{\mathrm{obs}}_{\mathrm{H}}$ and a partitioned *free* facial zone i.

1) Compute bottom-up results and initialize $W^{\mathrm{fr},i}_{\mathrm{H},0} = \emptyset$.
2) In step $t+1$, for every grammars $\{g_j\}$, propose $W^{\mathrm{fr},i}_{\mathrm{H},t+1}$ from $W^{\mathrm{fr},i}_{\mathrm{H},t}$ and calculate the posterior ratio $\frac{p(W^{\mathrm{fr},i}_{\mathrm{H},t+1}|\mathbf{I}^{\mathrm{obs}}_{\mathrm{H},\Lambda_i})}{p(W^{\mathrm{fr},i}_{\mathrm{H},t}|\mathbf{I}^{\mathrm{obs}}_{\mathrm{H},\Lambda_i})} = \theta^j_{t+1}$.
3) Select the greatest $\theta^j_{t+1} > 1$, accept $W^{\mathrm{fr},i}_{\mathrm{H},t+1}$, and repeat step 2. Otherwise if $\theta^j_{t+1} \leq 1$ for all g_j, stop the pursuit.

Fig. 17. Algorithm for pursuing free curves of zone i in the high-resolution layer.

D. Experiment II: Scale Transition and Model Selection

A crucial yet unaddressed issue is the *scale transition*. In previous sections, we showed how to parse an input face image on all three layers of the And-Or graph. However, the layers of representations that we need depends on both the resolution of observed images and the model complexity. It is against our intuition to model a high resolution face with a simple holistic PCA, or to describe a low resolution face with a sophisticated graphical model of image primitives.

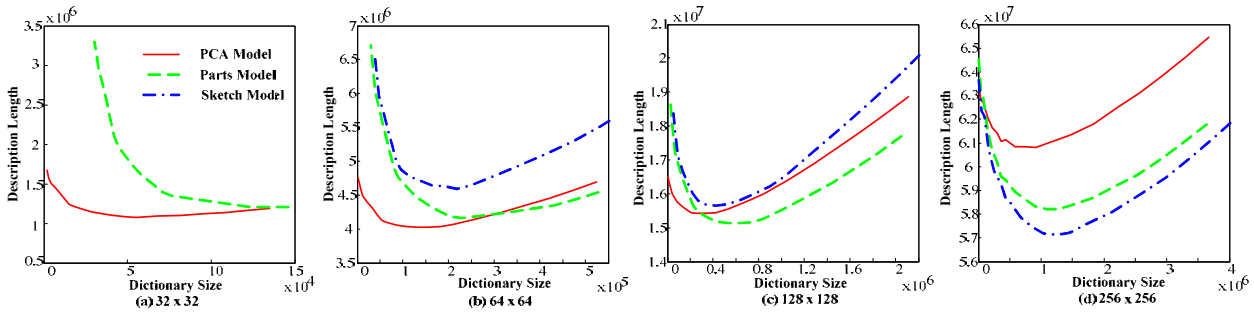


Fig. 18. Plot of coding length $\hat{DL}$ for the ensemble of testing images v.s. dictionary size $|\Delta|$ at four different scales.

Similar to [7], we formulated this problem as model selection under the *minimum description length* (MDL) principle: $DL = L(\Omega_I; \Delta) + L(\Delta)$, where $\Omega_I = \{I_1, ..., I_M\}$ is the sample set. The first term is the expected coding length of $\Omega_{\mathbf{I}}$ given dictionary Δ and the second term is the coding length of Δ. Empirically, we can estimate DL by:

$$\hat{DL} = \sum_{\mathbf{I}_i \in \Omega_{\mathbf{I}}} \sum_{w \sim p(W|I_i;\Delta)} (-\log p(I_i|w;\Delta) - \log p(w)) + \frac{|\Delta|}{2} \log M \tag{37}$$

We randomly partitioned the face images into a training set and a testing set. Training data was used to construct the three-layer And-Or graph model. Then the testing data was resized in four different resolutions: 32×32, 64×64, 128×128 and 256×256. $\hat{DL}$ was computed for every resolution set with different layers of our model. To obtain the minimum description length, we simply variate the size of the dictionaries/codebooks, e.g. increasing the number of principal components or image primitives. In practice, we computed $-\log p(I_i|w;\Delta)$ by the reconstruction error, $-\log p(w)$ by counting bits of the binary file storing the variables, $|\Delta|$ by counting bits of the binary file storing the models, and M was the number of testing data. Figure 18 showed that enlarging the codebook soon reached limit if the resolution continuously increased, thus switching to more sophisticated models (finer layers) became necessary.

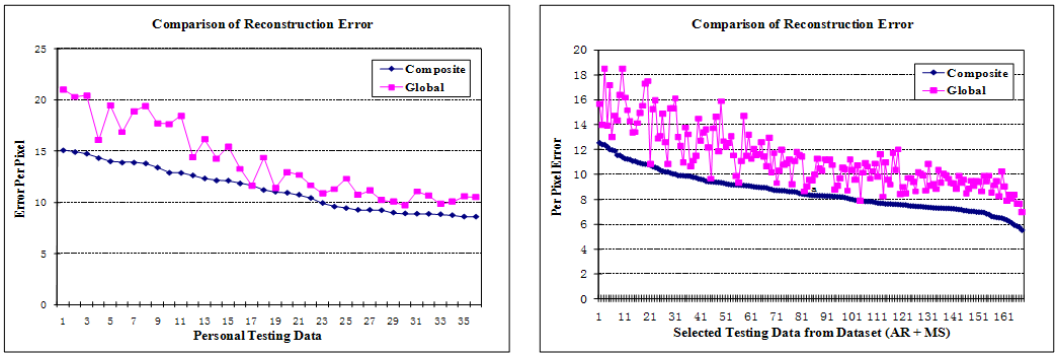


Fig. 19. Comparison of reconstruction errors of our composite model against a global AAM model. The test are conducted on (a) Selected testing images from AR and MSRA images; and (b) images from self-captured videos.

V. Experiment III: Reconstructing Images and Generating Cartoon Sketches

We construct a three-layer And-Or graph model with 811 parsing graphs annotated on face images across different genders, ages, and expressions selected from AR[19], FERET[23], LHI[36] and some MSRA images. Given an input image, the faces are first localized by AdaBoost[30] in OpenCV, on which the parsing proceeds until reaching a valid configuration. Experiments show that our model reconstructs face images with rich details, generates vivid facial sketches (Figure 21), and especially helps where the details (e.g.,wrinkles) are critical for face characterization (e.g., aged people in Figure 20). Quantitative improvement of the reconstruction accuracy on images from both standard databases and personal videos is shown in Figure 19, where our composite model compares favorably in terms of lower error and better consistency (smoother curves) against a global AAM model with codebook of approximately same size. Furthermore, the structural variabilities of our model is illustrated by parsing a video of facial motion in Figure 10(b) with the hair manually labelled.

After computing $(W^*_{\rm L}, W^*_{\rm M}, W^*_{\rm H})$, we reconstructed $(\mathbf{I}^{\rm rec}_{\rm L}, \mathbf{I}^{\rm rec}_{\rm M}, \mathbf{I}^{\rm rec}_{\rm H})$ and generated the corresponding sketches $(S^{\rm syn}_{\rm L}, S^{\rm syn}_{\rm M}, S^{\rm syn}_{\rm H})$ by replacing the rendering dictionaries in Figure 4.

$$(\mathbf{B}^{\rm pht}_{\rm L}, \mathbf{B}^{\rm pht}_{\rm cp}, \mathbf{B}^{\rm pht}_{\rm H}) \longrightarrow (\mathbf{B}^{\rm geo}_{\rm L}, \mathbf{B}^{\rm geo}_{\rm cp}, \mathbf{B}^{\rm geo}_{\rm H}) \tag{38}$$

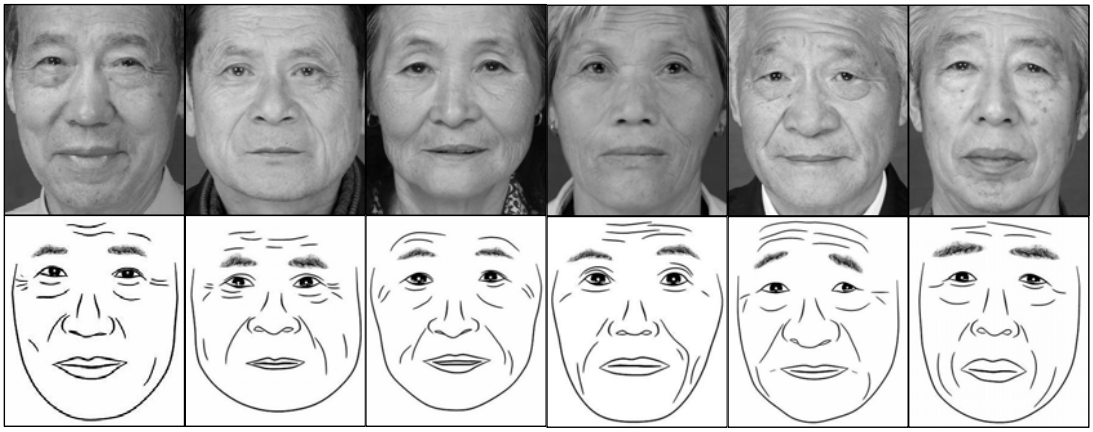


Fig. 20. Sketching results for aged faces where wrinkles are very important features for perception.

We called $S_{\mathrm{L}}^{\mathrm{syn}}, S_{\mathrm{M}}^{\mathrm{syn}}$ the initial sketches not shown since they are formed by linking the landmark points. The final facial sketch $S_{\mathrm{H}}^{\mathrm{syn}}$ assembles the symbolic representations of the image primitives, where smoothness constraints are enforced on their connections. More sketching results are shown in Figure 21.

VI. Conclusion and Future Work

In conclusion, we present a hierarchical-compositional representation for modeling human faces in the form of an And-Or graph model, which simultaneously account for the face regularity and dramatic structural variabilities caused by scale transitions and state transitions. Experiment had shown that our model helps reconstruct face images with great structural variations and rich details, and facilitates the generation of vivid cartoon sketches. We can also generate stylish sketches by learning the dictionaries from artistic drawings[3]. Another interesting future work is to synthesize the images from sketches.

Acknowledgments

The authors would like to thank Microsoft Research Asia for sharing some of the images. This work was supported by NSF IIS-0222967, IIS-0244763 and a Kodak Fellowship program.

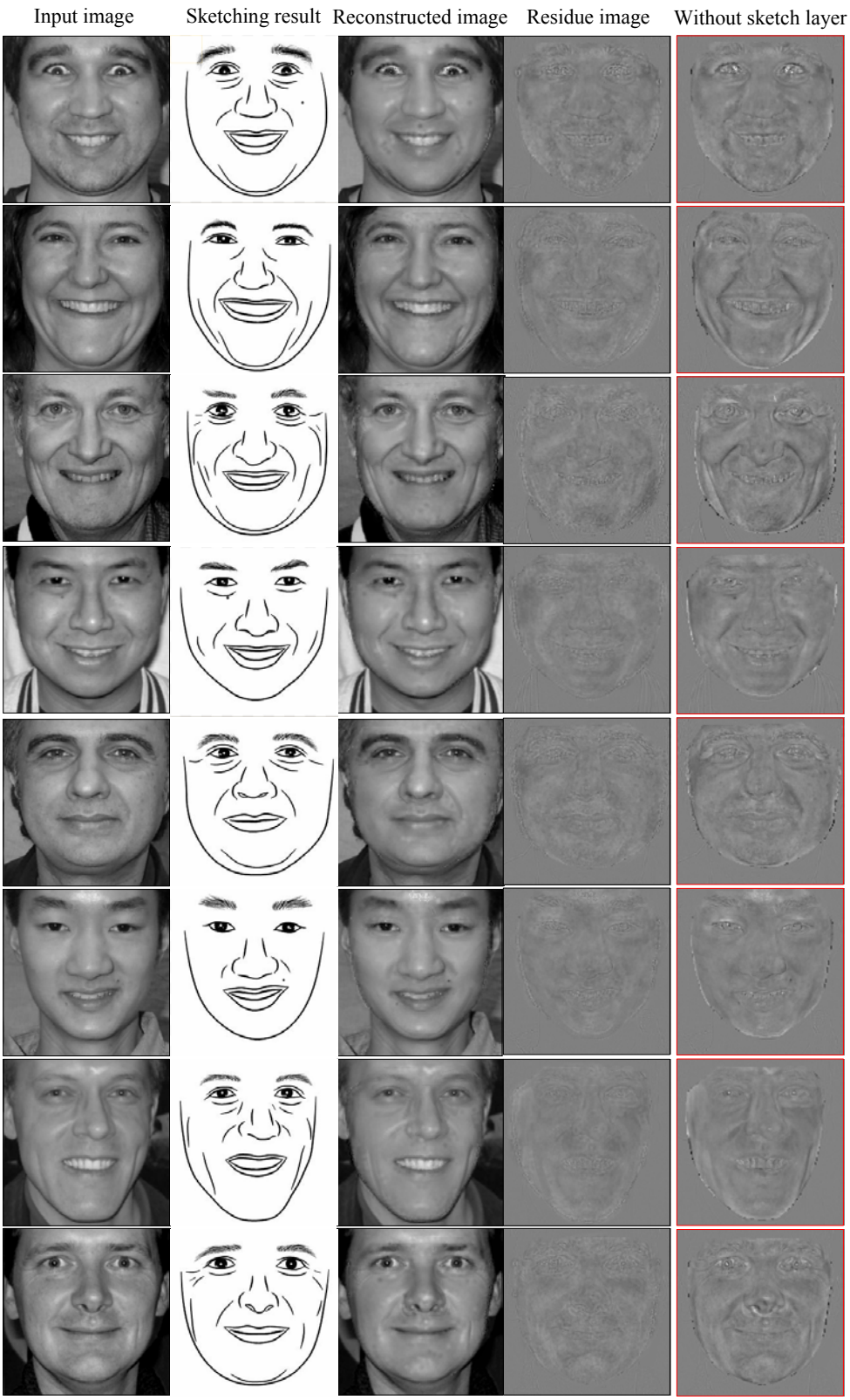


Fig. 21. More results of reconstructed images, automatically generated sketches and residue images of our model. The residue images from reconstruction without sketch layer are also shown for comparison. We easily see that our model helps capture rich details and generate vivid facial sketches. Difference styles can be achieved by replacing the rendering dictionaries.

REFERENCES

[1] S.P. Abney, “Stochastic attribute-value grammars”, *Computational Linguistics*, 23(4), 597-618, 1997.

[2] V. Bruce, E. Hanna, N. Dench, P. Healey and M. Burton, “The importance of “Mass” in line drawings of faces”, *Applied Cognitive. Psychology.*, vol.6, pp.619-628, 1992.

[3] 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.

[4] H.Chen, Z.J.Xu, Z.Q.Liu and S.C.Zhu, “Composite templates for cloth modeling and sketching”, *CVPR*, 2006.

[5] T.F. Cootes, C.J. Taylor, D. Cooper, and J. Graham, “Active shape models–their training and applications”, *CVIU*, 61(1):38-59, 1995.

[6] T.F. Cootes and C.J. Taylor. “Constrained Active Appearance Models” *ICCV*, 2001

[7] R.H. Davies, T.F. Cootes, C.Twining and C.J. Taylor, “An information theoretic approach to statistical shape modelling”, *BMVC*, 2001.

[8] M. Fischler and R. Elschlager, “The representation and matching of pictorial structures”, *IEEE Trans. on Computers*, 22(1):67C92, 1973.

[9] K.S. Fu, “Syntactic Pattern Recognition and Applications”, *Prentice Hall*, 1981.

[10] C. Guo, S.C. Zhu and Y.N. Wu, “Towards a Mathematical Theory of Primal Sketch and Sketchability”, *ICCV*, 2003.

[11] 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.

[12] B. Heisele, P. Ho, J. Wu and T. Poggio, “Face recognition: component-based versus global approaches”, *CVIU*, Vol. 91, No. 1/2, 6-21 2003.

[13] 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.

[14] T. Kanade, “Computer recognition of human faces”, 1973.

[15] H. Koshimizu, M. Tominaga, T. Fujiwara, and K. Murakami, “On Kansei facial processing for computerized caricaturing system Picasso”, *Int’l Conf. Sys. Man, Cyber.*, vol.6, 294-299, 1999.

[16] L. Liang, F. Wen, Y.Q. Xu, X. Tang, H.Y. Shum, “Accurate face alignment using shape constrained Markov network”, *CVPR*, 2006.

[17] T. Lindeberg, “Scale-space Theory in Computer Vision”, *Kluwer Academic Publishers*, 1994.

[18] C. Liu, H.Y. Shum and C.S. Zhang. “Hierarchical shape model for automatic face localization”, *ECCV*, pp. 687-703, 2002.

[19] A.M. Martinez and R. Benavente, “The AR Face Database“, *CVC Technical Report*, no.24, 1998.

[20] J. Pearl, “Heuristics: Intelligent Search Strategies for Computer Problem Solving”, *Addison-Wesley*, 1984.

[21] J. Pearl. “Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference”, *Morgan Kaufmann Publishers*, 1988.

[22] A. Pentland, B. Moghaddam and T. Starner, “View-based and Modular eigenspaces for face recognition”, *CVPR*, 1994

[23] P.J. Phillips, H. Wechsler, J. Huang and P. Rauss, “The FERET database and evaluation procedure for face recognition algorithms”, *Image and Vision Computing J.*, vol.16, no.5, pp 295-306, 1998

[24] J. Rekers and A. Schürr, “A parsing algorithm for context sensitive graph grammars”, *TR*, Leiden Univ. 1995.

[25] X. Tang and X. Wang, “Face sketch synthesis and recognition”, *ICCV*, 2003.

[26] Y. Tian, T. Kanade, and J. Cohn, “Recognizing action units of facial expression analysis”, *IEEE Trans. on PAMI*, vol.23, no.2, 229-234, 2001.

[27] M. Turk and A. Pentland, “Eigenfaces for recognition”, *J. of Cogn. Neurosci.*, vol.3, no.1, pp. 71-86, 1991.

[28] S. Ullman and E. Sali, “Object classification using a fragment-based representation”, *BMVC*, 2000.

[29] T. Vetter, “Synthesis of novel views from a single face image”, *Int’l J. of Comp. Vision* 2(28) 103-116, 1998.

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

[31] M. Weber, M. Welling, and P. Perona, “Towards automatic discovery of object categories”, *CVPR*, 2000.

[32] J. Xiao, S. Baker and T. Kanade, “Real-time combined 2d+3d active appearance models”, *CVPR*, 2004.

[33] Z.J. Xu, H. Chen and S.C. Zhu, “A high resolution gramatical model for face representation and sketching”, *CVPR*, 2005.

[34] Z.J. Xu and J. Luo “Face recognition by expression-driven sketch graph matching”, *ICPR*, 2006.

[35] M.H. Yang, D.J. Kriegman, and N. Ahuja, “Detecting faces in images: a survey”, *IEEE Trans. on PAMI*, vol 24, no.1, pp. 1-25 2002.

[36] Z.Y. Yao, X. Yang, and S.C. Zhu, “Introduction to a large scale general purpose groundtruth dataset: methodology, annotation tool, and benchmarks”, *EMMCVPR*, 2007.

[37] A.L. Yuille, D. Cohen, and P. Hallinan, “Feature extraction from faces using deformable templates”, *Int’l J. of Computer Vision*, vol.8 99-111, 1992.

[38] W. Zhao, R. Chellappa, A. Rosenfeld, and P.J. Phillips, “Face recognition: a literature survey”, *UMD Cfar TR 948*, 2000.

[39] S.C. Zhu, Y.N. Wu and D.B. Mumford, “Filters, random fields and maximum entropy (FRAME)”, *Int’l J. of Computer Vision* 27(2) 1-20, 1998.

[40] S.C. Zhu and D. Mumford, “Quest for a stochastic grammar of images”, *Foundations and Trends in Computer Graphics and Vision*, 2007.