A Hierarchical Compositional Model for Face Representation and Sketching

Zijian Xu, Hong Chen, Song-Chun Zhu, and Jiebo Luo, *Senior Member*, *IEEE*

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, and the third layer further divides the 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 demonstrated by reconstructing high-resolution face images and generating the cartoon facial sketches. Our model is useful for a wide variety of applications, including recognition, nonphotorealisitc rendering, superresolution, and low-bit rate face coding.

Index Terms—Face sketch, hierarchical, grammar model.

1 Introduction

1.1 Motivation

Human faces have been extensively studied in vision and graphics for a wide range of tasks from detection [33], [38], recognition [14], [17], [26], [41], [31], tracking [35], expression [30], [37], animation [16], [2], superresolution [3], [22] to nonphotorealistic rendering [5], [18], [29], [36], with both the discriminative [5], [16], [32] and generative models [8], [13], [16], [26], [31]. 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 specific problem domains, do not account for rich facial details that appear on the high-resolution or 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 Fig. 1a), and over multiscales (see Fig. 1b) 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 multiresolution representation of human faces, which can capture fine facial details and account for large structural variations.

- *Z. Xu is with Moody's Corporation, Wall Street Analytic, One Front Street, Suite 1900, San Francisco, CA 94111. E-mail: zjxu@stat.ucla.edu.*
- *H. Chen is with Brion Technologies Incorporated, 4211 Burton Drive, Santa Clara, CA 95054. E-mail: chenh@brion.com.*
- *S.-C. Zhu is with the Statistics Department, University of California, Los Angeles, 8125 Math Science Building, Box 951554, Los Angeles, CA 90095. E-mail: sczhu@stat.ucla.edu.*
- *J. Luo is with Kodak Research Laboratories, Eastman Kodak Company, Rochester, NY 14650-1816. E-mail: jiebo.luo@kodak.com.*

Manuscript received 4 June 2007; revised 4 Nov. 2007; accepted 26 Dec. 2007; published online 27 Feb. 2008.
Recommended for acceptance by Y. Ma.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-2007-06-0328.
Digital Object Identifier no. 10.1109/TPAMI.2008.50.

1.2 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 toward 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 Fig. 2:

1. *Face layer*, where faces are represented as a whole by PCA models [26], [31].
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 subgraphs of patches —transformed image primitives. Another 10 zones, shaped by the local parts, also represent the discovered skin features (e.g., marks or wrinkles) as subgraphs of patches.

According to the scale/resolution transition of input face images, elements of coarser layers expand to a subgraph of elements in the finer layers and thus leads to structural

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

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 Fig. 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* 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 Fig. 2, we parse the input face image and arrive in a configuration, as in Fig. 3. In essence, an AND-OR graph is essentially a set of multiscale faces of all structural, geometric, and photometric variations. We construct the AND-OR graph by maximizing the likelihood of parameters given a set of annotated face parse 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 of transitions between layers. These transitions are based on both the scales/resolutions of input face images and the accuracy required by specific tasks, e.g., low resolution for detection, medium resolution for recognition, and high resolution for nonphotorealistic rendering.

1.3 Related Work

In computer vision, numerous methods had been proposed to model human faces. Zhao et al. suggested [41] that following the psychology study of how human use holistic and local features, existing methods can be categorized as 1) *global* [7], [8], [13], [16], [31], [2], 2) *feature-based (structural)* [10], [17], [32], [33], [34], [40], and 3) *hybrid* [14], [26] methods. Early holistic approaches [13], [31] used intensity pattern of the whole face as input and modeled the photometric variation by linear combination of the *eigenfaces*. These *PCA models*

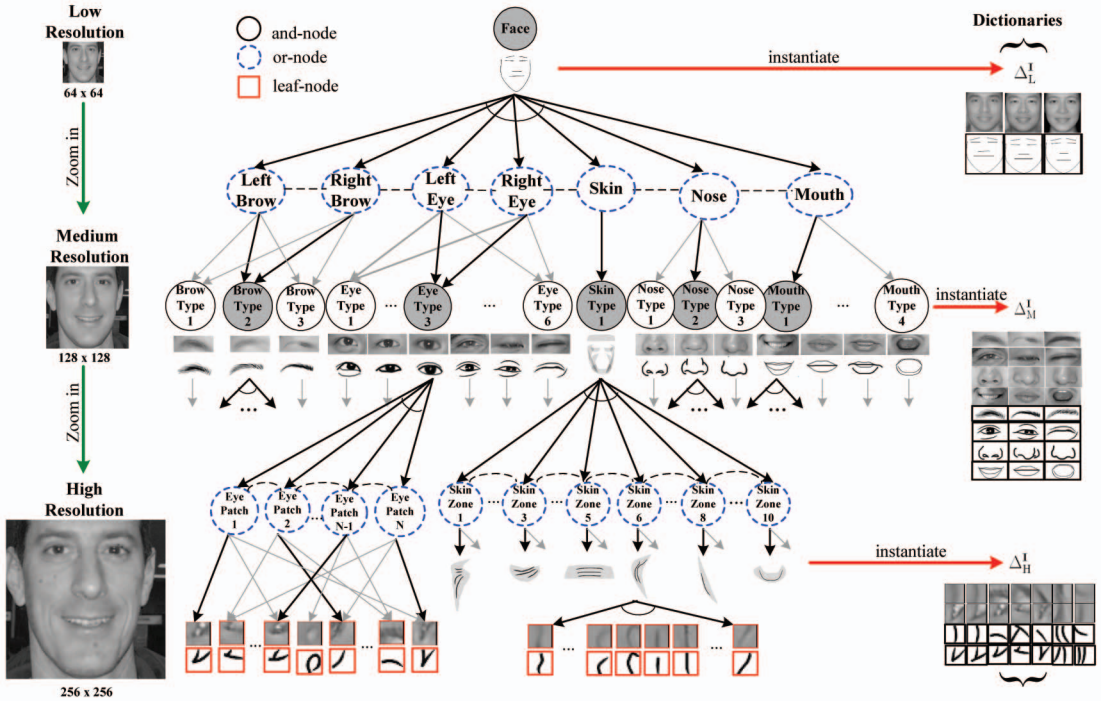


Fig. 2. An illustration of the three-layer face AND-OR graph representation. The dark arrows and shadow nodes represent a composition of seven leaf nodes $\langle BrowType2(L/R), EyeType3(L/R), SkinType1, NoseType2, MouthType1\rangle$, each being a *subtemplate* at the medium-resolution layer. This generates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

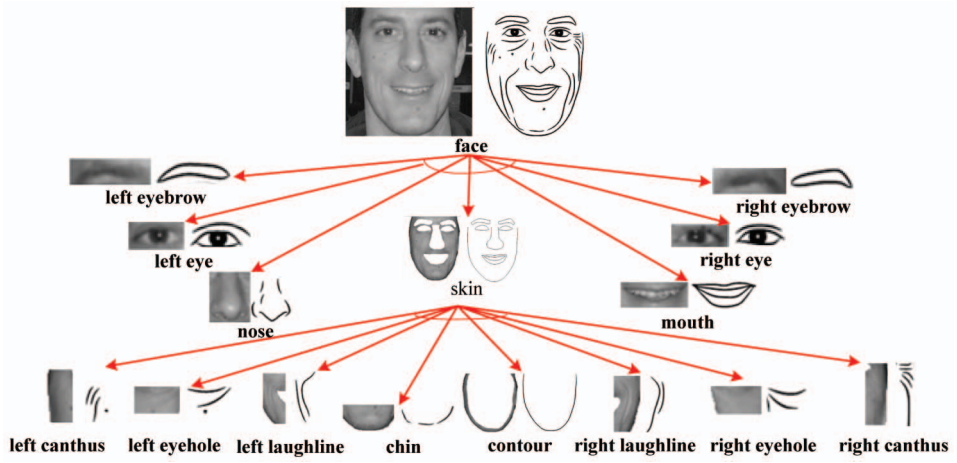


Fig. 3. A face is parsed into the configuration of the local parts and skin zones, of which both the images and symbolic representations are shown. Parts and skin zones can be further parsed into subgraphs of image primitives.

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 AAM [8], [35] and *Morphable Models* [16], [2]. 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* [10], *Deformable Templates* [40], *Constellation Model* [34], and *Fragment-based Model* [32]. 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 modeled jointly. In addition, there are some hybrid methods [14], [26], 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 pairwise Gaussians and, therefore, the flexibilities are limited.

To model the scale variabilities, some researchers construct a Gaussian/Laplacian pyramid from the input image [20] 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 [21]. Both methods are inefficient and inadequate for human faces, where dramatic variabilities are 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* [32] 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, and 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 subgraph 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 [24] (1984), our model is more similar to some recent works by Chen et al. [6] and Zhu and Mumford [43]. The AND-OR graph that we use is shown to be equivalent to a *Context Sensitive Grammar* (CSG) [28], which integrates the *Stochastic Context Free Grammar* (SCFG) [11] and *Markov Random Field* (MRF) [42] 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 [37] and nonphotorealistic rendering [18], [36]. Supported by psychology studies [4], 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 [29] and [5]. The former renders facial sketches similar to high-pass filtered images by combining linear *eigensketches* and does not provide any high-level description of the face. Constrained on an *Active Shape Model* (ASM) [7], the latter generates facial sketches by collecting local evidences from artistic drawings in the training set and lack of structural variations and facial details.

1.4 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, and photometric). Our model parses the input face images of given resolutions by traversing the constructed AND-OR graph and drawing from the multiresolution template dictionaries. The traversals are guided by the SG and MDL criterion. Our hierarchical-compositional model,

powered by the SG, 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, nonphotorealisitc 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 2. In Section 3, we define the probabilities on the AND-OR graph model and learn the model parameters. Section 4 introduces the Bayesian inference algorithm and the scale transition process. Finally, the experimental results on reconstructing and sketching are reported in Section 5.

2 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.

2.1 Introduction to Face AND-OR Graph

AND-OR graph was originally introduced in [24] and revisited in some recent work [6], [43]. In this paper, we adapted it to represent the composite deformable templates of human faces over multiple scales, as showed in Fig. 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 in Fig. 2, from which the face instances of all variations are derived.
2. *Nonterminal nodes* $V_N = V^{\text{and}} \cup V^{\text{or}}$ include a set of AND-nodes and a set of OR-nodes.

   Each AND-node in $\{u : u \in V^{\text{and}}\}$ is a composite template, which expands to a set of OR-nodes according to the image complexity of input faces. Each OR-node in $\{v : v \in V^{or}\}$ 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 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 multiresolution deformable templates governed by various numbers of landmark points to allow geometric and photometric variations. Leaf nodes are essentially the instantiations of AND-nodes with no further expansions available. Examples are shown in Fig. 2 as templates of faces, parts, and image primitives (e.g., edgelets, junctions, or blobs) in low, medium, and high resolutions, respectively. Each template has its intensity and symbolic representations kept in the dictionaries, with the latter essentially strokes linked by landmark.
4. $\mathcal{R} = \{r_1, r_2, \ldots, r_{N(R)}\}$ 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 $\{r_a = \psi^a (v_i, v_j) : a = 1, \ldots, N(R)\}$, serving as a statistical constraint. Our defined relations include *center distance*, *size ratio*, *relative angle*, *closeness of bonding points*, and *appearance similarity*. One type of relations are those vertically defined on the AND-nodes and the OR-nodes to which they expand, maintaining the geometric and photometric consistency between parent and children. For example, the appearance of a medium-resolution template shall resemble the composition of its high-resolution subtemplates. Another type is defined horizontally on children of AND-nodes, keeping the spatial configurations valid, e.g., the two eyes are located symmetrically. The horizontal relations of nodes are inheritable from their parents. That is, the expanded OR-nodes from one AND-node are implicitly correlated to the expanded OR-nodes from another AND-node, through the common ancestor of the AND-nodes. We avoided assigning relations between every two graph nodes in the same layer, which leads to overcomplicated 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 an SCFG, the probabilities from both formulations are adopted.

Traversing from the root node, a set of valid face configurations $\Sigma = \{g_1, g_2, \ldots, g_M\}$ of Leaf nodes are generated. Each of these traversals are called *parse graphs*. Essentially, an AND-OR graph is equivalent to a set of multiresolution faces with possible structural, geometric, and photometric variations. A parsed configuration is shown in Fig. 3.

2.2 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) reconstruction is achieved. Fig. 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 image primitives like edgelets, junctions, or blobs for high-resolution layer. The *Intermediate Complexity* fragments proposed in [32] is probably regarded as the circumstantial evidence.

In the *Low-resolution layer*, we adopted the well-known AAM [8] 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/photometric 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) labeled at eye corners, nose wings, mouth corners, and on

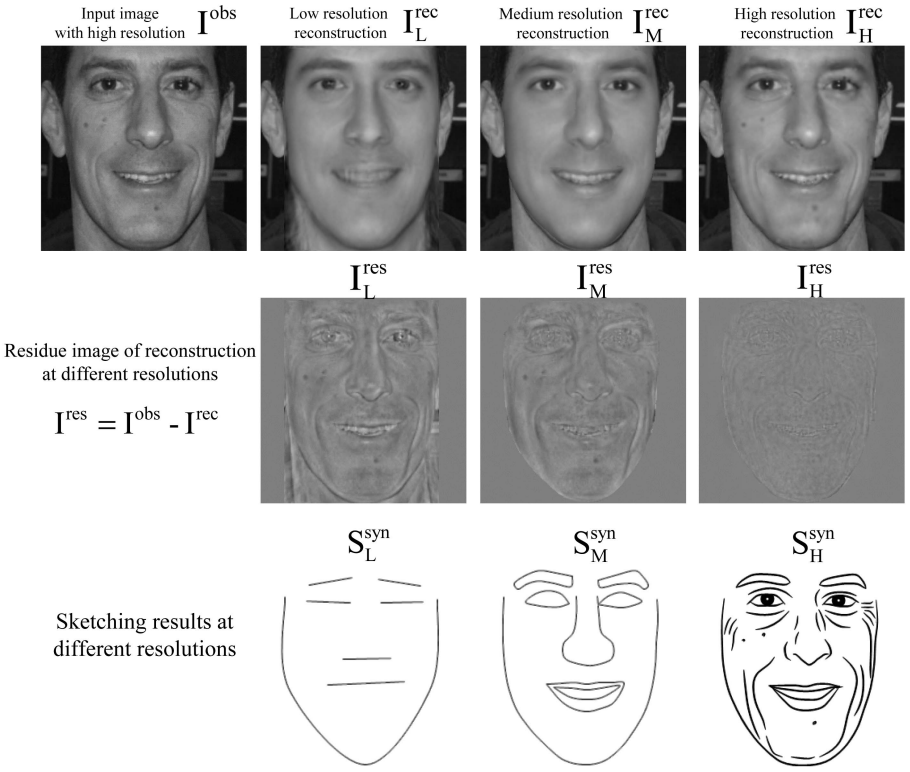
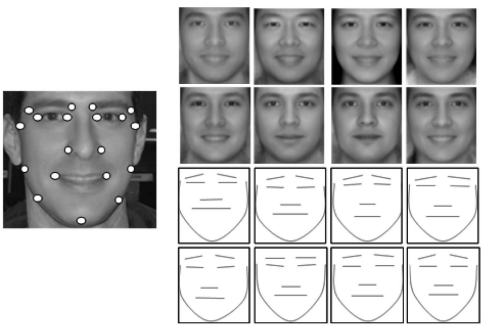


Fig. 4. Face high-resolution image $\mathbf{I}^{\text{obs}}$ of 256×256 pixels is reconstructed by the AND-OR graph model in a coarse-to-fine fashion. The first row shows three reconstructed images $\mathbf{I}^{\text{rec}}_{\text{L}}$, $\mathbf{I}^{\text{rec}}_{\text{M}}$, and $\mathbf{I}^{\text{rec}}_{\text{H}}$ in low, medium, and high resolution, respectively. $\mathbf{I}^{\text{rec}}_{\text{L}}$ is reconstructed by the low-resolution layer, and the facial components like eyes, nose, and mouth are refined in $\mathbf{I}_{\text{M}}$ with medium-resolution layer. The skin marks and wrinkles appear in $\mathbf{I}^{\text{rec}}_{\text{H}}$ 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.

face contour, as shown in Fig. 5a. Another convenient assumption was made that all (frontal) face templates in the 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) $\{x_1, x_2, \ldots, x_M\}$ and the corresponding texture vectors (normalized image pixels) $\{g_1, g_2, \ldots, g_M\}$ 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 Fig. 5b:

$$\Delta^{\mathbf{I}}_{\text{L}} = \left\{\mathbf{B}^{\text{geo}}_{\text{L}}, \mathbf{B}^{\text{pht}}_{\text{L}}\right\}. \tag{2}$$



(a) (b)

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

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 $\mathbf{B}^{\text{geo}}_{\text{L}}$, $\mathbf{B}^{\text{pht}}_{\text{L}}$, 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:

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

An input low-resolution face image $\mathbf{I}^{\text{obs}}_{\text{L}}$ of 64×64 pixels is then reconstructed, as in Fig. 4:

$$\mathbf{I}^{\text{obs}}_{\text{L}} = \mathbf{I}^{\text{rec}}_{\text{L}}\left(W_{\text{L}}; \Delta^{\mathbf{I}}_{\text{L}}\right) + \mathbf{I}^{\text{res}}_{\text{L}}. \tag{4}$$

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 Fig. 2. Fig. 6a 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 Λ^{cp}_i, $i = 1, \ldots, 6$ denote the six facial components, then

$$\cup_{i=1}^{6} \Lambda^{\text{cp}}_i = \Lambda_{\text{cp}} \subset \Lambda_{\text{M}}. \tag{5}$$

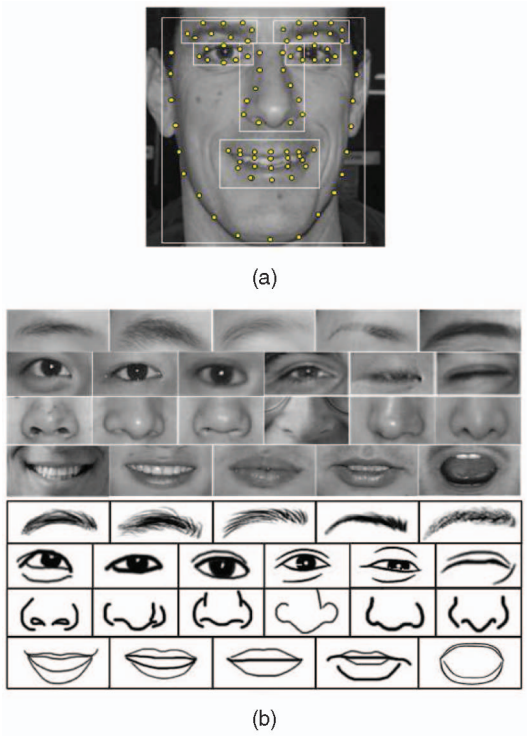


Fig. 6. (a) The locations of facial components and the control points defined on them. (b) Dictionary $\Delta_{\mathrm{M}}^{\mathbf{I}}$ of facial components and their artistic sketches drawn according to the control points. The examples in the same row are of same type but different modes and selected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

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 [23], FERET [27], LHI [39], 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 [8]. Therefore, a total number of $3+5+3+4=15$ AAM models are trained from the manually labeled medium-resolution face images. The dictionary of these models is shown in Fig. 6b:

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

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

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

where $\ell_i=\{1,\ldots,15\}$ is the index of the selected AAM model—switch variable for the ith OR-node. The Λ_{cp} is then reconstructed as the union of reconstruction of Λ_i^{cp}, $i=1,\ldots,6$:

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

An input medium-resolution face image $\mathbf{I}_{\mathrm{M}}^{\mathrm{obs}}$ of 128×128 pixels is then reconstructed, as in Fig. 4. The rest skin pixels

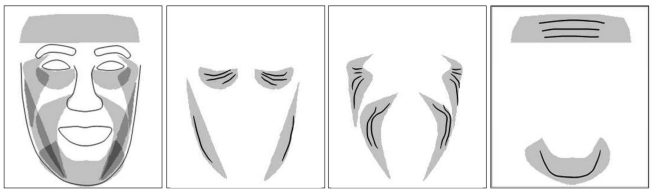


Fig. 7. (a) Sixteen facial zones for high-resolution face features. Six zones, indicated by solid shapes, are to refine the eyebrows, eyes, nose, and mouth. Another 10 zones, indicated by shaded regions, are where the skin features like marks or wrinkles occur. These zones are localized by shapes of the facial parts computed in the medium-resolution layer. (b), (c), and (d) Typical wrinkles (curves) patterns of the 10 skin zones. To reliably detect these subtle features, we need strong prior models and global context.

$\Lambda_{\mathrm{ncp}}=\Lambda_{\mathrm{M}}-\Lambda_{\mathrm{cp}}$ are up-sampled from $\mathbf{I}_{\mathrm{L}}^{\mathrm{rec}}$ with boundary conditions of Λ_{cp}:

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

In the *high-resolution layer*, much more subtle features are exposed, as we can see in Fig. 4. Thus, the medium-resolution layer representations is further decomposed into subgraphs of sketchable [12] 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 Fig. 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 the medium-resolution layer, and the 10 new zones are introduced to cover the features that appear on rest of the skin (forehead, canthus, eyehole, laugh line, 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 of a *structural* zone (nose) and a *free* zone (laugh line) are shown in Fig. 8a. Each of the rectangles represents an image primitive with (small) geometric and photometric deformations. In the 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. Fig. 8b shows the dictionary of the learned image primitives and their corresponding sketch representations. Note that we defined a small number ($2\sim4$) of control points for each sketch patch to connect with neighboring patches properly and generate smooth face sketches. In order to capture marks or specularities, we also include gabor bases of various scales in high-resolution dictionary. Furthermore, the teeth are refined by gabor bases if the mouth is open/wide open. So are the pupils of open eyes. The total number of gabor bases used for reconstruction is limited to less than 100:

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

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

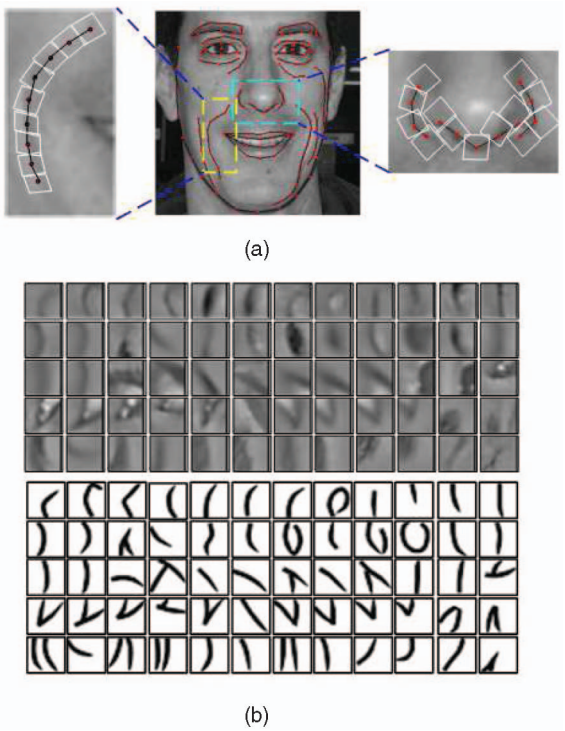


Fig. 8. (a) Refinement of the nose and a "smile fold" by sketch primitives, which are represented by small rectangles. (b) Dictionary $\Delta_{\mathrm{H}}^{\mathbf{I}}$ of sketch primitives and their corresponding sketch strokes.

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

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

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

Our sketch representation capture more prolific facial details than the state-of-art face sketch method [5] and expression classification method [30].

3 Learning Probabilistic Models on the And-Or Graph

3.1 Defining the Probabilities

Let $\mathcal{P}$ be the probability model defined over the AND-OR graph (see Section 2.1), 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 define a *parse graph* g as a valid traversal of $\mathcal{G}_{\text{and-or}}$, which consists of a set of graph nodes $V = \{v_1, v_2, \ldots, v_{N(v)}\} \in V_N \cup V_T$ and a set of relations $R \in \mathcal{R}$. The probability of a graph is then denoted as $p(g; \Theta)$.

We first define $p(g)$ as its parsing tree component g_T, as the product of probabilities of visited OR-nodes $g_T = \{v_1, v_2, \ldots, v_{N(v)}\}$ and the values of their switch variables:

$$\begin{aligned} p(g_T) \propto \prod_{v_i \in g_T} p_i(\omega(v_i)) &= \exp\left\{\sum_{v_i \in g_T} \log p(\omega(v_i))\right\} \\ &= \exp\left\{\sum_{v_i \in g_T} \log \prod_{j=1}^{N(\omega_i)} \theta_{ij}^{\delta(\omega(v_i)-j)}\right\}, \end{aligned} \tag{12}$$

where θ_{ij} is the probability that $\omega(v_i)$ takes value j, and $\delta(.)$ is a *delta* function. Another component, the MRF is a probability on the relations in the parse graph. It is written in terms of energies by relation function ψ on two nodes and by constraint function ϕ on each single node.

Let f be the true probability distribution that produces the faces in the training set. Our goal is now to derive $p(g)$ by minimizing the Kullback-Leibler divergence $KL(f|p)$ subject to relation constraints observed from training set:

$$\begin{aligned} \mathbf{E}_{p(g)}[H(\omega(v_i))] &= \mathbf{E}_f[H(\omega(v_i))], \\ \mathbf{E}_{p(g)}\left[H\left(\phi^{(a)}(v_i)\right)\right] &= \mathbf{E}_f\left[H\left(\phi^{(a)}(v_i)\right)\right], \\ \mathbf{E}_{p(g)}\left[H\left(\psi^{(b)}(v_i, v_j)\right)\right] &= \mathbf{E}_f\left[H\left(\psi^{(b)}(v_i, v_j)\right)\right], \end{aligned}$$

where $v_i \in V^{\mathrm{or}}$, $a = 1, 2, \ldots, N(\phi)$, $v_i \in V$, $b = 1, 2, \ldots, N(\psi)$, $< v_i, v_j > \in E$, $N(\phi)$ and $N(\psi)$ are respectively the number of singleton constraints and pairwise constraints. H are the histograms of output values. Solving this constrained optimization by Lagrange multipliers yields $p(g; \Theta)$ as

$$\begin{aligned} p(g;\Theta) = \frac{1}{Z(\Theta)} \exp\Bigg\{ &- \sum_{v_i \in T} \langle \lambda_i, H(\omega(v_i)) \rangle \\ &- \sum_{v_i \in V} \sum_{a=1}^{N(\phi)} \left\langle \alpha_i^{(a)}, H\left(\phi^{(a)}(v_i)\right) \right\rangle \\ &- \sum_{<v_i, v_j> \in E} \sum_{b=1}^{N(\psi)} \left\langle \beta_{ij}^{(b)}, H\left(\psi^{(b)}(v_i, v_j)\right) \right\rangle \Bigg\}, \end{aligned} \tag{13}$$

where E is the set of node pairs on which relations are defined. $\Theta = (\lambda, \alpha, \beta)$ are the parameters, and $\lambda_{ij} = -\log \theta_{ij}$ in (12).

3.2 Estimating the Model Parameters

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

$$\Theta^* = \arg\max \sum_{i=1}^{N} \log p(g_i; \Theta). \tag{14}$$

The probability over the switch variable at OR-node i depends on the grammar rules we defined on the OR-node. Examples of such grammar rules in medium-layer OR-nodes are shown in Fig. 9, which set a specific mode for the facial parts such as to open an eye or to shut a mouth.

Use MLE to derive θ from $p(g; \Theta)$ with $\frac{\partial L(\Theta; \hat{G})}{\partial \theta} = 0$ yields:

$$-N \frac{\partial \log Z(\Theta)}{\partial \theta} - \sum_{k=1}^{N} \sum_{v_i^{(k)} \in g_T^{(k)}} \sum_{j}^{N(\omega_i)} \frac{\delta\left(\omega\left(v_i^{(k)}\right) - j\right)}{\theta_{ij}} = 0 \tag{15}$$

subject to $\sum_{j=1}^{N(\omega_i)} \theta_{ij} = 1$ for all $v_i \in g_T$. Solve this with Lagrange multiplier yields

$$\hat{\theta}_{ij} = \frac{\sum_{k=1}^{N} \delta\left(\omega\left(v_i^{(k)}\right) - j\right)}{N_{\omega_i} - N \frac{\partial \log Z(\Theta)}{\partial \theta} + N \frac{\partial \log Z(\Theta)}{\partial \theta}} = \frac{N_{ij}}{N_{\omega_i}}, \tag{16}$$

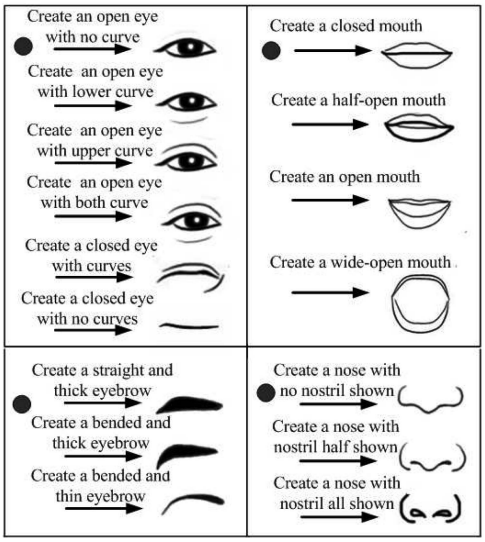


Fig. 9. Grammars defined on OR-nodes of medium-resolution layer for switching among various composite templates.

where N_{ω_i} is the total number of times that v_i is visited in all parse 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 $p(g_T)$ generates novel parsing trees, e.g., winking and excited, that were not even seen in the training data, as shown in Fig. 10.

After $p(g_T)$ is learned, we derive α and β by maximizing the entropy of $p(g;\Theta)$ subject to constraints previously defined—to match the expected histograms with the observed histograms [42]:

$$\frac{\partial L(\Theta;\hat{G})}{\partial \alpha^{(a)}} = -N\frac{\partial \log Z(\Theta)}{\alpha^{(a)}} - \sum_{k=1}^{N} H_{\phi}^{(a)}(g_k) = 0 \tag{17}$$

subject to $\mathbf{E}_{p(g)}[H_{\phi}^{(a)}(g)] = \frac{1}{N}\sum_{k=1}^{N} H_{\phi}^{(a)}(g_k)$ for all a, and

$$\frac{\partial L(\Theta;\hat{G})}{\partial \beta^{(b)}} = -N\frac{\partial \log Z(\Theta)}{\beta^{(b)}} - \sum_{k=1}^{N} H_{\psi}^{(b)}(g_k) = 0 \tag{18}$$

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

Similar to [42], we solve for α and β by iteratively updating them with

$$\begin{aligned}\frac{d\alpha^{(a)}}{dt} &= \mathbf{E}_{p(g)}\left[H_{\phi}^{(a)}(g)\right] - \frac{1}{N}\sum_{k=1}^{N} H_{\phi}^{(a),obs}(g_k) \\ &= H_{\phi}^{(a),\mathrm{syn}} - H_{\phi}^{(a),\mathrm{obs}},\end{aligned} \tag{19}$$

$$\begin{aligned}\frac{d\beta^{(b)}}{dt} &= \mathbf{E}_{p(g)}\left[H_{\psi}^{(b)}(g)\right] - \frac{1}{N}\sum_{k=1}^{N} H_{\psi}^{(b),obs}(g_k) \\ &= H_{\psi}^{(b),\mathrm{syn}} - H_{\psi}^{(b),\mathrm{obs}}.\end{aligned} \tag{20}$$

The algorithm of learning specific $\alpha^{(a)}$ and $\beta^{(b)}$ proceeds in Fig. 12. The sampling results of the learning procedure are shown in Fig. 11.

3.3 Experiment 1: 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.

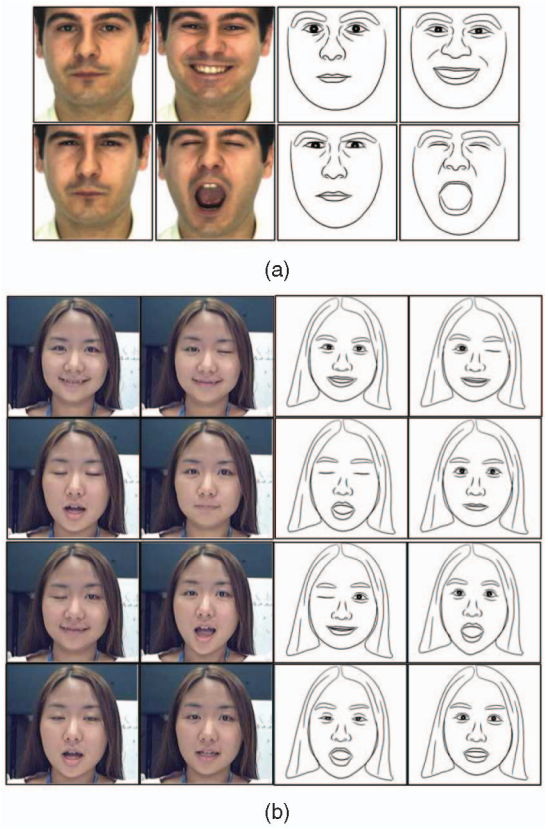


Fig. 10. Different face configurations are composed by various types of local facial components. (a) The four typical face configurations in the AR data set as *neutral*, *laughing*, *angry*, *and screaming*. (b) The eight novel face configurations inferred from the frames in a personal video clip. These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

We learned the $p(T)$ from an AR [23] data set, in which four typical expressions (*neutral*, *smiling*, *angry*, and *screaming*) are observed (Fig. 10a). In a personal video of facial motions, we observed eight facial expressions different from the training data. These novel expressions/configurations unseen in training set such as *winking* and *excited* were successfully sampled from learned AND-OR graph model to match the new observations, as shown in Fig. 10b.

Fig. 11 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.

4 Bayesian Inference and Scale Transition

Given an input face image $\mathbf{I}^{\mathrm{obs}}$, our goal is to determine the $W = (W_{\mathrm{L}}, W_{\mathrm{M}}, W_{\mathrm{H}})$ defined in Section 2.2 by maximizing the Bayesian posterior:

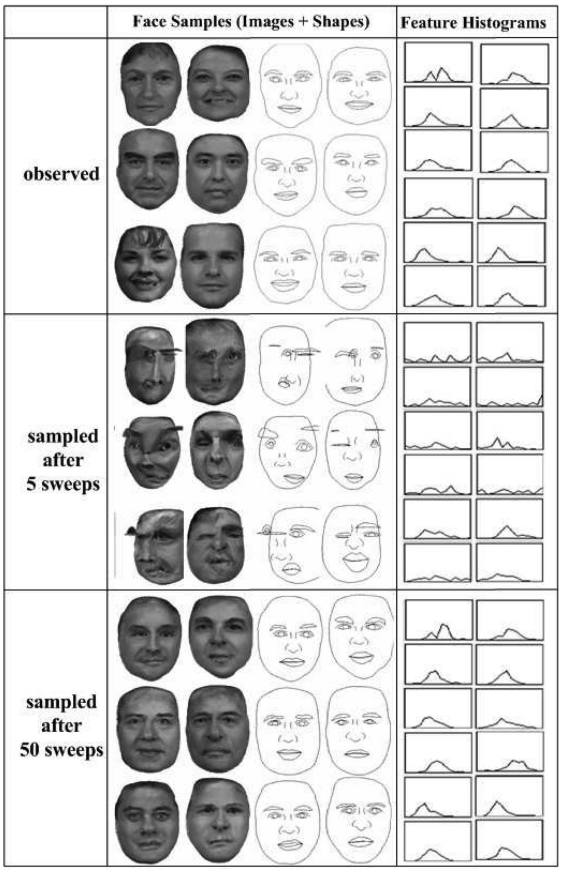


Fig. 11. Examples of observed and synthesized faces images and shapes. Selected feature histograms of observation and synthesis are shown in the right column. As the learning proceeds, synthesis histograms similar to the observations are produced.

$$\begin{aligned}(W_{\mathrm{L}}, W_{\mathrm{M}}, W_{\mathrm{H}})^* &= \arg\max p(W_{\mathrm{L}}, W_{\mathrm{M}}, W_{\mathrm{H}}|\mathbf{I}^{\mathrm{obs}}) \\ &= \arg\max p(\mathbf{I}^{\mathrm{obs}}|W) \cdot p(W) \\ &= \arg\max p(W_{\mathrm{H}}|W_{\mathrm{M}}, W_{\mathrm{L}}, \mathbf{I}^{\mathrm{obs}}) \\ &\quad \cdot p(W_{\mathrm{M}}|W_{\mathrm{L}}, \mathbf{I}^{\mathrm{obs}}) \cdot p(W_{\mathrm{L}}|\mathbf{I}^{\mathrm{obs}}).\end{aligned} \tag{21}$$

We notice that the parse graph g^* for $\mathbf{I}^{\mathrm{obs}}$ 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 3. Given an input image of certain resolution, all Leaf nodes of the resulting parse graph sit in the same layer—of same scale. We first build a three-layer Gaussian pyramid $(\mathbf{I}_{\mathrm{L}}^{\mathrm{obs}}, \mathbf{I}_{\mathrm{M}}^{\mathrm{obs}}, \mathbf{I}_{\mathrm{H}}^{\mathrm{obs}})$ from the input image. Then, $(W_{\mathrm{L}}, W_{\mathrm{M}}, W_{\mathrm{H}})^*$ shall be gradually optimized according to the layers in a coarse-to-fine fashion, as shown in Fig. 13.

Given a set of observed parse 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 parse 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. 12. Algorithm for learning parameter of the MRF model.

4.1 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 [8] in learning and computing W_{L}:

$$\begin{aligned}W_{\mathrm{L}}^* &= \arg\max p(W_{\mathrm{L}}|\mathbf{I}^{\mathrm{obs}}) \\ &= \arg\max p\big(\mathbf{I}_{\mathrm{L}}^{\mathrm{obs}}|W_{\mathrm{L}}; \Delta_{\mathrm{L}}^{\mathbf{I}}\big) p(W_{\mathrm{L}}) \\ &= \arg\max \exp\left\{-\big|\mathbf{I}_{\mathrm{L}}^{\mathrm{obs}} - \mathbf{I}_{\mathrm{L}}^{\mathrm{rec}}\big|^2/\big(2\sigma_{\mathrm{L}}^2\big) - \frac{1}{2} W_{\mathrm{L}}' \Big(\mathbf{S}_{W_{\mathrm{L}}}^{-1}\Big) W_{\mathrm{L}}\right\}.\end{aligned} \tag{22}$$

The first term of the second row denotes the likelihood, where $\mathbf{I}_{\mathrm{L}}^{\mathrm{rec}}$ is the reconstructed low-resolution layer governed by W_{L}, and σ_{L}^2 is the variance of reconstruction error learned from training data. The second term denotes the prior, where

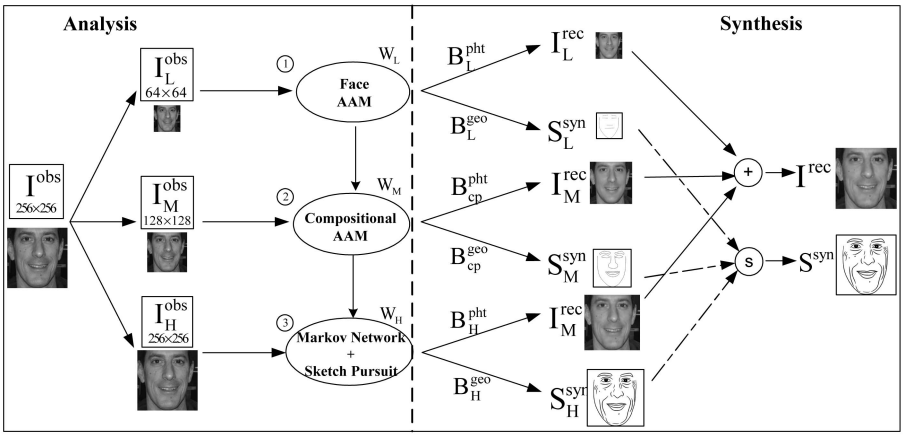


Fig. 13. The diagram of our model and algorithm. The arrows indicate the inference order. Left panel is the three layers. Right panel is the synthesis steps for both image reconstruction and sketching using the generative model.

$\mathbf{S}_{W_{\mathrm{L}}}$ is the covariance matrix of W_{L}. The optimized W_{L}^* can be computed efficiently by a *stochastic gradient descent* [8].

4.2 Layer 2: The Medium-Resolution Compositional AAM Model

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

$$\begin{aligned} W_{\mathrm{M}}^* &= \arg\max p(W_{\mathrm{M}}|W_{\mathrm{L}}, \mathbf{I}^{\mathrm{obs}}) \\ &= \arg\max p\big(\mathbf{I}_{\mathrm{M}}^{\mathrm{obs}}|W_{\mathrm{M}}, W_{\mathrm{L}}; \Delta_{\mathrm{M}}^{\mathbf{I}}, \Delta_{\mathrm{L}}^{\mathbf{I}}\big) p(W_{\mathrm{M}}|W_{\mathrm{L}}). \end{aligned} \tag{23}$$

The first term indicates the likelihood probability:

$$\begin{aligned} & p\big(\mathbf{I}_{\mathrm{M}}^{\mathrm{obs}}|W_{\mathrm{L}}, W_{\mathrm{M}}; \Delta_{\mathrm{M}}^{\mathbf{I}}, \Delta_{\mathrm{L}}^{\mathbf{I}}\big) \\ & \propto \exp\left\{-\frac{1}{2}\big(\mathbf{I}_{\mathrm{M}}^{\mathrm{obs}} - \mathbf{I}_{\mathrm{M}}^{\mathrm{rec}}\big)' \Sigma_r^{-1} \big(\mathbf{I}_{\mathrm{M}}^{\mathrm{obs}} - \mathbf{I}_{\mathrm{M}}^{\mathrm{rec}}\big)\right\} \\ & = \exp\left\{-\sum_{i=1}^{6} \frac{|\mathbf{r}_{\mathrm{cp},i}|^2}{2\sigma_{\mathrm{cp},i}^2} - \frac{|\mathbf{r}_{\mathrm{L}}|^2}{2\sigma_{\mathrm{L}}^2}\right\}, \end{aligned} \tag{24}$$

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

$$\begin{aligned} p(W_{\mathrm{M}}|W_{\mathrm{L}}) \propto & \prod_{i=1}^{6} p(\ell_i) \cdot \prod_{i=1}^{6} p\Big(W_{\mathrm{cp}}^i|W_{\mathrm{L}}\Big) \\ & \cdot \prod_{<v_k,v_l>\in E_{\mathrm{cp}}} p\Big(W_{\mathrm{cp}}^k, W_{\mathrm{cp}}^l\Big). \end{aligned} \tag{25}$$

The first component denotes the prior probability of the parsing tree, as defined in Section 3:

$$\prod_{i=1}^{6} p(\ell_i) \propto \prod_{i=1}^{6} \prod_{j=1}^{N(\omega_i)} \theta_{ij}^{\delta(\ell_i - j)} = \prod_{i=1}^{6} \theta_{i\ell_i}, \tag{26}$$

where $\delta(.)$ is a *Delta* function, and $\theta_{i\ell_i}$ is simply the frequency of that the ith 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 [8]:

$$\prod_{i=1}^{6} p\Big(W_{\mathrm{cp}}^i|W_{\mathrm{L}}\Big) \propto \prod_{i=1}^{6} \exp\Big\{-{W_{\mathrm{cp}}^i}' \mathbf{S}_{W_{\mathrm{cp}}^i}^{-1} W_{\mathrm{cp}}^i - \mathbf{d}_{\mathrm{cp,L}}^i{}' \mathbf{S}_{d_i}^{-1} \mathbf{d}_{\mathrm{cp,L}}^i\Big\}, \tag{27}$$

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 only computed the geometric displacement and ignored the photometric displacement, although the photometric displacement could be critical for other applications like *superresolution* [3], [22]. 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 ratios* 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 *similarity* (ψ_{sm}):

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

1) In step k, each medium-resolution layer Or-nodes proposal in $\{W_{cp,k}^{i,(n)}\}_{n=1}^N$ is associated with a sequence of messages $\omega_{j,k}^{i,(n)}, \langle i,j\rangle \in E_{\mathrm{M}}$ from the connected template and its weight $\pi_k^{i,(n)}$.
2) Propose a set of templates (only the geometric part) of all possible types $\{W_{cp,k+1}^{i,(n)}\}_{n=1}^N \sim p(W_{cp}^i|W_{\mathrm{L}})$. Then diffuse (limiting to a few update steps) them using the constrained AAM models and record the reconstruction errors as likelihoods.
3) For each $W_{cp,k+1}^{i,(n)}$, compute and normalize $\omega_{j,k+1}^{i,(n)}, \langle i,j\rangle \in E_{cp}$ as in *SBP*.
4) For each $W_{cp,k+1}^{i,(n)}$, compute and normalize $\pi_{k+1}^{i,(n)}$ by likelihoods and updated messages.
5) Repeat from 2) to 4) until convergence.
6) Select the proposals with largest weights as approximations.

Fig. 14. Algorithm to infer the medium-resolution layer hidden variables.

$$\begin{aligned} & \prod_{<v_k,v_l>\in E_{\mathrm{cp}}} p\Big(W_{\mathrm{cp}}^k, W_{\mathrm{cp}}^l\Big) \\ & \propto \exp\left\{-\sum_{<v_k,v_l>\in E_{\mathrm{cp}}} \sum_{\psi^{(b)}\in\Psi_{kl}} \Big\langle \beta_{kl}^{(b)}, H\Big(\psi^{(b)}(v_k,v_l)\Big)\Big\rangle\right\}, \end{aligned} \tag{28}$$

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 help 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, the appearance of the nose and mouth of the same person is probably remotely relevant.

For computational simplicity and efficiency, we approximate the optimized W_{M}^* in three steps. First, 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, as shown in Fig. 14. Then, these proposed templates were locally diffused using pretrained constrained AAM models [8]. 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. Each of these proposals are associated with a sequence of messages from every neighbors and its weight. We then introduced the *sequential belief propagation* [25], [15] to update the messages and weights sequentially until convergence or the algorithm exceeds the prescribed maximum iterations. The quality of proposals from $p(W_{\mathrm{M}}|W_{\mathrm{L}})$ affect the inference efficiency and reliability in this layer. If the low-resolution layer AAM result is seriously wrong, we may need to propose excessive number of templates in wider ranges for correction. In this rare case, we can provide a little manual constraints in the low-resolution layer, as discussed in constrained AAM [8].

4.3 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}:

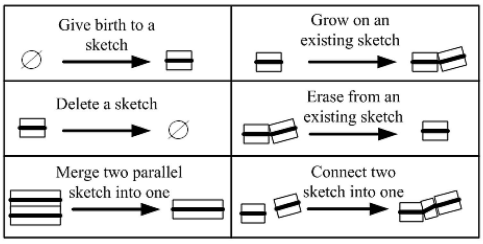


Fig. 15. Grammars used for free curve pursuit in the high-resolution layer, including *birth/death*, *split/merge*, and *connect*.

$$\begin{aligned} W_{\mathrm{H}}^{*} &= \arg\max p\left(W_{\mathrm{H}} | W_{\mathrm{M}}, \mathbf{I}_{\mathrm{H}}^{\mathrm{obs}}\right) \\ &= \arg\max p\left(W_{\mathrm{H}}^{\mathrm{fr}} | W_{\mathrm{H}}^{\mathrm{st}}, \mathbf{I}_{\mathrm{H}}^{\mathrm{obs}}\right) p\left(W_{\mathrm{H}}^{\mathrm{st}} | W_{\mathrm{M}}, \mathbf{I}_{\mathrm{H}}^{\mathrm{obs}}\right), \end{aligned} \tag{29}$$

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 2.2. They are inferred sequentially in the high-resolution layer.

$W_{\mathrm{H}}^{\mathrm{st}}$ includes six facial zones (Fig. 7a), in which the eyebrows, eyes, nose, and mouth are further decomposed into subgraphs of image primitives, e.g., the nose in Fig. 8a. 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:

$$\begin{aligned} p(W_{\mathrm{H}}^{\mathrm{st},i} | W_{\mathrm{cp}}^{i}, \mathbf{I}_{\mathrm{H},\Lambda_i}^{\mathrm{obs}}) \propto \exp\Bigg\{ & -\sum_{k=1}^{N_i} \frac{|\mathbf{r}_k|^2}{2\sigma_k^2} - \frac{1}{2}\mathbf{d}_{\mathbf{i}}' \Sigma_{c_i}^{-1} \mathbf{d}_{\mathbf{i}} \\ & - \sum_{<k,l>} \frac{1}{2}\left(E_{kl}^{d}(\mathrm{p}_k, \mathrm{p}_l) + E_{kl}^{a}(\mathrm{p}_k, \mathrm{p}_l)\right) \Bigg\}, \end{aligned} \tag{30}$$

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 [19]. Experiments showed fast convergence and accurate fitting:

$$W_{\mathrm{H}}^{\mathrm{st}*} = \left\{W_{\mathrm{H}}^{\mathrm{st},i*}\right\}_{i=1}^{6} = \arg\max \prod_{i=1}^{6} p\left(W_{\mathrm{H}}^{\mathrm{st},i} | W_{\mathrm{cp}}^{i}, \mathbf{I}_{\mathrm{H},\Lambda_i}^{\mathrm{obs}}\right). \tag{31}$$

To help detect the iris that are partially occluded, we combined the Hough transform of circles in between the eyelids after the shape of eyes were accurately fitted.

$W_{\mathrm{H}}^{\mathrm{fr}}$ includes another 10 facial zones, covering the rest of the skin regions. These zones, shown in Figs. 7b, 7c, and 7d, 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 Fig. 8a. 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 labeled the skin features in every *free* zones for a set of training images. Some "typical" curves are shown in Figs. 7b, 7c, and 7d, from which the prior models were learned in favor of certain properties:

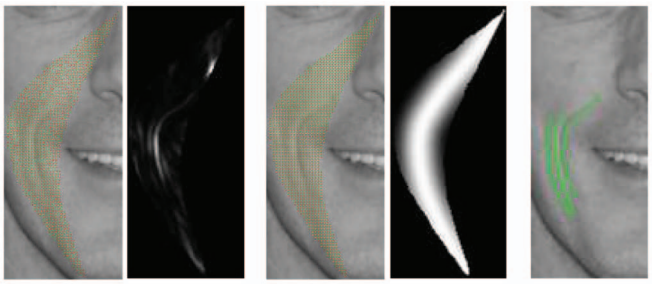


Fig. 16. The process of curve tracking. (a) The bottom-up results of orientation and gradient magnitudes. (b) The prior of orientation field and gradient magnitudes learned from training data. (c) Curve tracking results.

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. $p_{\mathrm{on}}(\mathrm{on}|x, y) = p_{xy}^{\mathrm{on}}$ 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}^{on}, μ_{xy}^{θ}, and σ_{xy}^{θ} by accumulating information from nearby "typical" curves in the normalized training data (Fig. 16b).
4. $p_{sm}(\mathrm{p}_k, \mathrm{p}_l) \propto \exp\{-\frac{1}{2}(E^d + E^\theta + E^s + E^t)\}$ guarantees the *position*, *orientation*, *scale*, and *intensity* consistency of two consecutive primitives p_k and 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 $E^t = |\mathrm{p}_k - \mathrm{p}_l|^2/\sigma_t^2$.

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

$$\begin{aligned} p\left(W_{\mathrm{H}}^{\mathrm{fr},i} | \mathbf{I}_{\mathrm{H},\Lambda_i}^{\mathrm{obs}}\right) \propto\ & p_n(N_i) \cdot \prod_{j=1}^{N_i} p_\ell(L_j) \cdot \prod_{<k,l>} p_{sm}(\mathrm{p}_k, \mathrm{p}_l) \\ & \cdot \prod_{k=1}^{K} p_{\mathrm{on}}(\mathrm{on}|x_k, y_k) p_\theta(\theta_k|x_k, y_k) p_r(\mathbf{r}_k), \end{aligned} \tag{32}$$

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 and steering filters) was taken for initialization (Fig. 16a). In step $t + 1$, we proposed $W_{\mathrm{H},t+1}^{\mathrm{fr},i}$ from $W_{\mathrm{H},t}^{\mathrm{fr},i}$ by selecting from a set of grammars (Fig. 15) and computed the posterior ratio:

$$\frac{p\left(W_{\mathrm{H},t+1}^{\mathrm{fr},i} | \mathbf{I}_{\mathrm{H},\Lambda_i}^{\mathrm{obs}}\right)}{p\left(W_{\mathrm{H},t}^{\mathrm{fr},i} | \mathbf{I}_{\mathrm{H},\Lambda_i}^{\mathrm{obs}}\right)} = \theta. \tag{33}$$

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 Fig. 17, and results are shown in

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

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

Fig. 17. Algorithm for pursuing free curves of facial zone i.

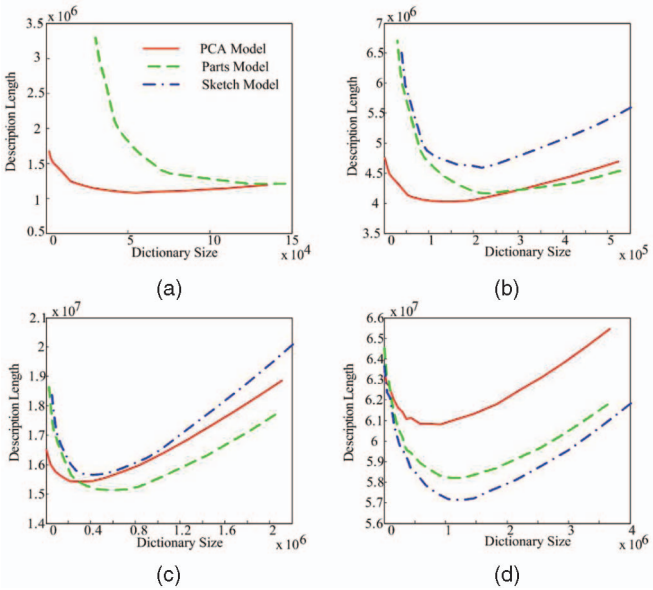


Fig. 18. Plot of coding length $\hat{DL}$ for the ensemble of testing images versus dictionary size $|\Delta|$ at four different scales.

Fig. 20. Gabor filters of various scales are used in capturing other features like marks and specularities.

4.4 Experiment 2: 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. Similar to [9], we formulated this problem as model selection under the MDL principle: $DL = L(\Omega_I; \Delta) + L(\Delta)$, where $\Omega_I = \{I_1, \ldots, I_M\}$ 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

$$\hat{DL} = \sum_{\mathbf{I}_i \in \Omega_{\mathbf{I}}} \sum_{w \sim p(W|I_i;\Delta)} (-\log p(I_i|w;\Delta) - \log p(w)) + \frac{|\Delta|}{2} \log M. \tag{34}$$

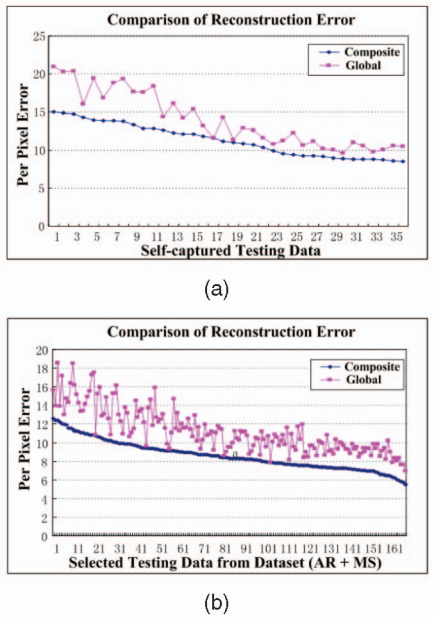


Fig. 19. Comparison of reconstruction errors of our composite model against a global AAM model. The test is conducted on (a) selected testing images from AR and MSRA images and (b) images from self-captured videos.

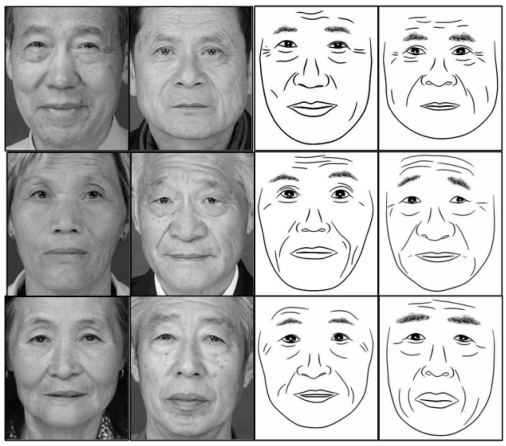


Fig. 20. Sketching results for aged faces, where wrinkles are very important features for perception.

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 MDL, we simply variate the size of the dictionaries/code books, 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. Fig. 18 showed that enlarging the code book soon reached limit if the resolution continuously increased, thus switching to more sophisticated models (finer layers) became necessary.

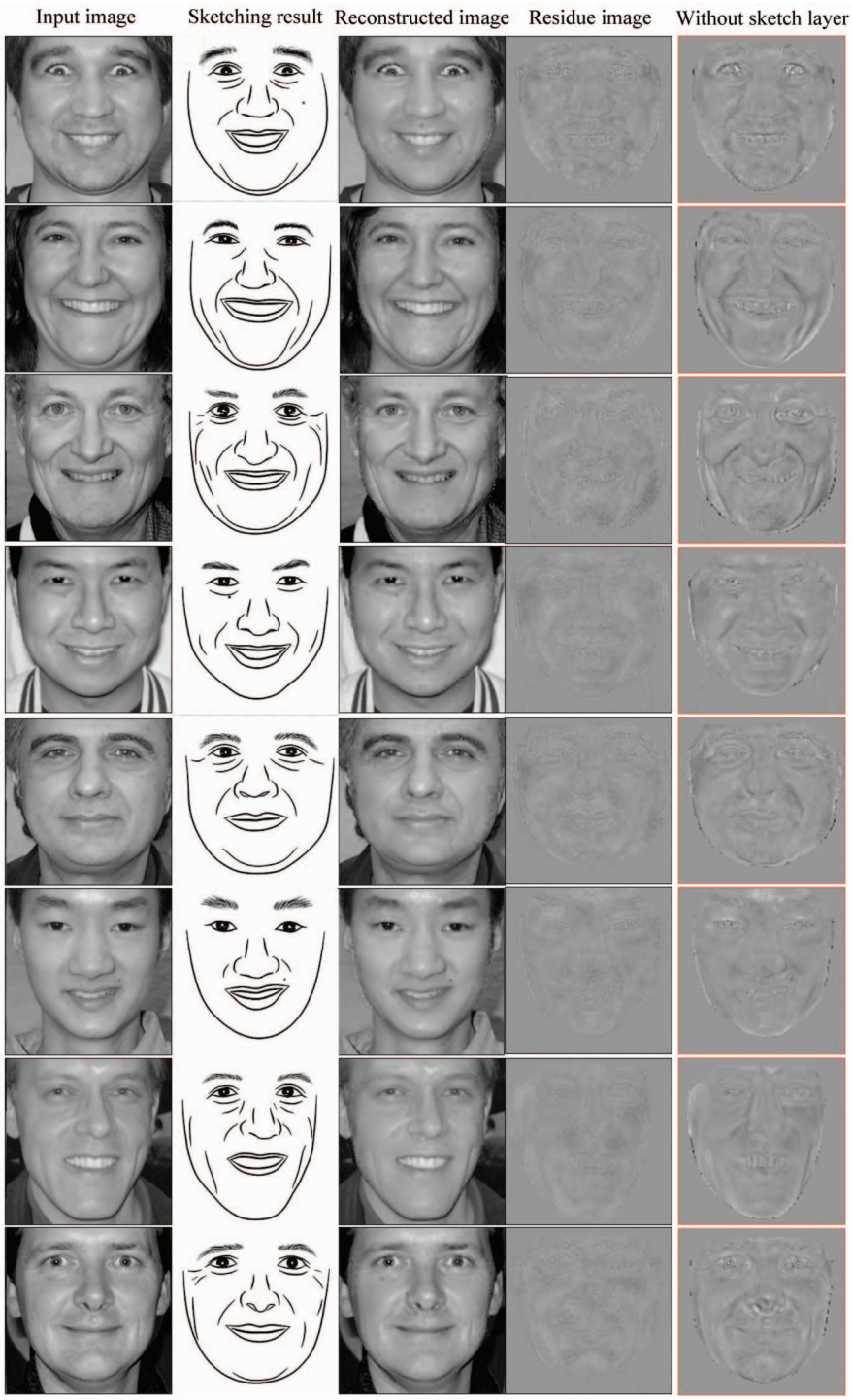


Fig. 21. More results of reconstructed images, automatically generated sketches and residue images of our model. The residue images from reconstruction without sketch layer are also shown for comparison. We easily see that our model helps capture rich details and generate vivid facial sketches. Difference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

5 Experiment 3: Reconstructing Images and Generating Cartoon Sketches

We construct a three-layer AND-OR graph model with 811 parse graphs annotated on face images across different genders, ages, and expressions selected from AR [23], FERET [27], LHI [39], and some MSRA images. In performing the comparison (Figs. 18 and 19), 650 parse graphs were used for training and the other 161 as testing. Given an input image, the faces are first localized by AdaBoost [33] 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 (Fig. 21), and especially helps where the details (e.g., wrinkles) are critical for face characterization. Quantitative improvement of the reconstruction accuracy on images from both standard databases and personal videos is shown in Fig. 19, where our composite model compares favorably in

terms of lower error and better consistency (smoother curves) against a global AAM model with code book of approximately same size. Furthermore, the structural variabilities of our model is illustrated by parsing a video of facial motion in Fig. 10b with the hair manually labeled.

After computing $(W_{\mathrm{L}}^{*}, W_{\mathrm{M}}^{*}, W_{\mathrm{H}}^{*})$, we reconstructed $(\mathbf{I}_{\mathrm{L}}^{\mathrm{rec}}, \mathbf{I}_{\mathrm{M}}^{\mathrm{rec}}, \mathbf{I}_{\mathrm{H}}^{\mathrm{rec}})$ and generated the corresponding sketches $(S_{\mathrm{L}}^{\mathrm{syn}}, S_{\mathrm{M}}^{\mathrm{syn}}, S_{\mathrm{H}}^{\mathrm{syn}})$ by replacing the rendering dictionaries:

$$\left(\mathbf{B}_{\mathrm{L}}^{\mathrm{pht}}, \mathbf{B}_{\mathrm{cp}}^{\mathrm{pht}}, \mathbf{B}_{\mathrm{H}}^{\mathrm{pht}}\right) \longrightarrow \left(\mathbf{B}_{\mathrm{L}}^{\mathrm{geo}}, \mathbf{B}_{\mathrm{cp}}^{\mathrm{geo}}, \mathbf{B}_{\mathrm{H}}^{\mathrm{geo}}\right). \tag{35}$$

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. The objective evaluation of facial sketches is difficult, and we must resort to human perceptions at the beginning. The preliminary studies showed that people are more sensitive to global properties like hair styles, face contour, or shading effect. Being able to correctly capture (even exaggerate) the distinctive features is also crucial. For example, people pay immediate attention to facial components of irregular size/shape/expression or the wrinkle patterns in aged faces.

6 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 [5], or produce lively cartoon animations from video [44]. 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 US National Science Foundation IIS-0222967, IIS-0244763, and a Kodak Fellowship program.

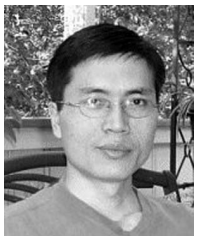
References

[1] S.P. Abney, "Stochastic Attribute-Value Grammars," *Computational Linguistics,* vol. 23, no. 4, pp. 597-618, 1997.
[2] V. Blanz and T. Vetter, "Face Recognition Based on Fitting a 3D Morphable Model," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 25, no. 9, pp. 1063-1074, Sept. 2003.
[3] S. Baker and T. Kanade, "Hallucinating Faces," *Proc. IEEE Int'l Conf. Automatic Face and Gesture Recognition,* 2000.
[4] 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.
[5] H. Chen, Y.Q. Xu, H.Y. Shum, S.C. Zhu, and N.N. Zhen, "Example-Based Facial Sketch Generation with Non-Parametric Sampling," *Proc. IEEE Int'l Conf. Computer Vision,* 2001.
[6] H. Chen, Z.J. Xu, Z.Q. Liu, and S.C. Zhu, "Composite Templates for Cloth Modeling and Sketching," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition,* 2006.
[7] T.F. Cootes, C.J. Taylor, D. Cooper, and J. Graham, "Active Shape Models—Their Training and Applications," *Computer Vision and Image Understanding,* vol. 61, no. 1, pp. 38-59, 1995.
[8] T.F. Cootes and C.J. Taylor, "Constrained Active Appearance Models," *Proc. IEEE Int'l Conf. Computer Vision,* 2001.
[9] R.H. Davies, T.F. Cootes, C. Twining, and C.J. Taylor, "An Information Theoretic Approach to Statistical Shape Modelling," *Proc. British Machine Vision Conf.,* 2001.
[10] M. Fischler and R. Elschlager, "The Representation and Matching of Pictorial Structures," *IEEE Trans. Computers,* vol. 22, no. 1, p. 67C92, Jan. 1973.
[11] K.S. Fu, *Syntactic Pattern Recognition and Applications.* Prentice Hall, 1981.
[12] C. Guo, S.C. Zhu, and Y.N. Wu, "Towards a Mathematical Theory of Primal Sketch and Sketchability," *Proc. IEEE Int'l Conf. Computer Vision,* 2003.
[13] P.L. Hallinan, G.G. Gordon, A.L. Yuille, and D.B. Mumford, *Two and Three Dimensional Patterns of the Face.* A.K. Peters, 1999.
[14] B. Heisele, P. Ho, J. Wu, and T. Poggio, "Face Recognition: Component-Based versus Global Approaches," *Computer Vision and Image Understanding,* vol. 91, nos. 1/2, pp. 6-21, 2003.
[15] G. Hua and Y. Wu, "Multi-Scale Visual Tracking by Sequential Belief Propagation," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition,* 2004.
[16] M.J. Jones and T. Poggio, "Multi-Dimensional Morphable Models: A Framework for Representing and Matching Object Classes," *Int'l J. Computer Vision,* vol. 2, no. 29, pp. 107-131, 1998.
[17] T. Kanade, *Computer Recognition of Human Faces,* 1973.
[18] H. Koshimizu, M. Tominaga, T. Fujiwara, and K. Murakami, "On Kansei Facial Processing for Computerized Caricaturing System Picasso," *Proc. Int'l Conf. Systems, Man, and Cybernetics,* vol. 6, pp. 294-299, 1999.
[19] L. Liang, F. Wen, Y.Q. Xu, X. Tang, and H.Y. Shum, "Accurate Face Alignment Using Shape Constrained Markov Network," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition,* 2006.
[20] T. Lindeberg, *Scale-Space Theory in Computer Vision.* Kluwer Academic, 1994.
[21] C. Liu, H.Y. Shum, and C.S. Zhang, "Hierarchical Shape Model for Automatic Face Localization," *Proc. European Conf. Computer Vision,* pp. 687-703, 2002.
[22] C. Liu, H.Y. Shum, and C.S. Zhang, "Two-Step Approach to Hallucinating Faces: Global Parametric Model and Local Nonparametric Model," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition,* 2001.
[23] A.M. Martinez and R. Benavente, "The AR Face Database," CVC Technical Report 24, 1998.
[24] J. Pearl, *Heuristics: Intelligent Search Strategies for Computer Problem Solving.* Addison-Wesley, 1984.
[25] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann, 1988.
[26] A. Pentland, B. Moghaddam, and T. Starner, "View-Based and Modular Eigenspaces for Face Recognition," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition,* 1994.
[27] 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.
[28] J. Rekers and A. Schürr, "A Parsing Algorithm for Context Sensitive Graph Grammars," technical report, Leiden Univ., 1995.
[29] X. Tang and X. Wang, "Face Sketch Synthesis and Recognition," *Proc. IEEE Int'l Conf. Computer Vision,* 2003.
[30] Y. Tian, T. Kanade, and J. Cohn, "Recognizing Action Units of Facial Expression Analysis," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 23, no. 2, pp. 229-234, Feb. 2001.
[31] M. Turk and A. Pentland, "Eigenfaces for Recognition," *J. Cognitive Neuroscience,* vol. 3, no. 1, pp. 71-86, 1991.
[32] S. Ullman and E. Sali, "Object Classification Using a Fragment-Based Representation," *Proc. British Machine Vision Conf.,* 2000.
[33] P. Viola and M. Jones, "Rapid Object Detection Using a Boosted Cascade of Simple Features," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition,* 2001.
[34] M. Weber, M. Welling, and P. Perona, "Towards Automatic Discovery of Object Categories," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition,* 2000.
[35] J. Xiao, S. Baker, and T. Kanade, "Real-Time Combined 2D+3D Active Appearance Models," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition,* 2004.

[36] Z.J. Xu, H. Chen, and S.C. Zhu, "A High Resolution Gramatical Model for Face Representation and Sketching," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition,* 2005.
[37] Z.J. Xu and J. Luo, "Face Recognition by Expression-Driven Sketch Graph Matching," *Proc. Int'l Conf. Pattern Recognition,* 2006.
[38] M.H. Yang, D.J. Kriegman, and N. Ahuja, "Detecting Faces in Images: A Survey," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 24, no. 1, pp. 1-25, Jan. 2002.
[39] Z.Y. Yao, X. Yang, and S.C. Zhu, "Introduction to a Large Scale General Purpose Groundtruth Dataset: Methodology, Annotation Tool, and Benchmarks," *Proc. Int'l Workshop Energy Minimization Methods in Computer Vision and Pattern Recognition,* 2007.
[40] A.L. Yuille, D. Cohen, and P. Hallinan, "Feature Extraction from Faces Using Deformable Templates," *Int'l J. Computer Vision,* vol. 8, pp. 99-111, 1992.
[41] W. Zhao, R. Chellappa, A. Rosenfeld, and P.J. Phillips, "Face Recognition: A Literature Survey," UMD Cfar Technical Report 948, 2000.
[42] S.C. Zhu, Y.N. Wu, and D.B. Mumford, "Filters, Random Fields and Maximum Entropy (FRAME)," *Int'l J. Computer Vision,* vol. 27, no. 2, pp. 1-20, 1998.
[43] S.C. Zhu and D. Mumford, "Quest for a Stochastic Grammar of Images," *Foundations and Trends in Computer Graphics and Vision,* 2007.
[44] Z. Xu and J. Luo, "Accurate Dynamic Sketching of Faces from Video," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition, Workshop on Semantic Learning Applications in Multimedia,* 2007.

Zijian Xu received the BE degree from the Department of Computer Science and Technology, University of Science and Technology of China in 2001 and the doctorate degree from the Statistics Department, University of California, Los Angeles, in 2007, where he studied and researched on computer vision and was supervised by Professor Song-Chun Zhu. He is currently working with Moody's Corp. His research interests include but not limited to statistical modeling, computer vision, and machine learning.

Hong Chen received the BS degree from the College of Electrical and Communication Engineering, Xi'an Jiaotong University in 1996 and the PhD degree from the Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University in 2002. He worked as a postdoctoral researcher in the Department of Statistics and the Center of Image and Vision Science, University of California, Los Angeles (UCLA), from 2003-2006. He currently works in Brion Technologies, Inc., Santa Clara. His research interests include computer vision and machine learning.



Song-Chun Zhu received the BS degree from the University of Science and Technology of China in 1991, and the MS and PhD degrees from Harvard University in 1994 and 1996, respectively. He is currently a professor jointly with the Department of Statistics and Computer Science, University of California, Los Angeles (UCLA). Before joining UCLA, he worked at Brown University (applied math from 1996-1997), Stanford University (computer science from 1997-1998), and Ohio State University (computer science from 1998-2002). His research is focused on computer vision and learning, statistical modeling, and stochastic computing. He has published more than 90 papers in computer vision and received a number of honors, including the David Marr prize in 2003, the Marr prize honorary nomination in 1999 and 2007, a Sloan fellow in computer science in 2001, a US National Science Foundation Career Award in 2001, and a US ONR Young Investigator Award in 2001. In 2004, he founded, with friends, the Lotus Hill Institute for Computer Vision and Information Science in China as a nonprofit research institute (www.lotushill.org).

Jiebo Luo received the BS and MS degrees in electrical engineering from the University of Science and Technology of China in 1989 and 1992, respectively, and the PhD degree in electrical engineering from the University of Rochester in 1995. He is a senior principal scientist with Kodak Research Laboratories, Rochester, New York. His research interests include image processing, pattern recognition, computer vision, computational photography, medical imaging, and multimedia communication. He is the author of more than 120 technical papers and holds more than 40 granted US patents. He currently serves on the editorial boards of the *IEEE Transactions on Pattern Analysis and Machine Intelligence* (*TPAMI*), the *IEEE Transactions on Multimedia* (*TMM*), *Pattern Recognition* (*PR*), and the *Journal of Electronic Imaging*. He was a guest editor for the special issue on Image Understanding for Digital Photos in *PR* (2005), the special issue on Real-World Image Annotation and Retrieval in *TPAMI* (2008), the special issue on Event Analysis in *TCSVT* (2008), and the special issue on Integration of Content and Context for Multimedia Management in *TMM* in 2009. He is a Kodak Distinguished Inventor and a winner of the 2004 Eastman Innovation Award. He has also been an organizer of numerous technical conferences, most notably the general chair of the 2008 ACM International Conference on Image and Video Retrieval (CIVR), an area chair of the 2008 IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), a program cochair of the 2007 SPIE International Symposium on Visual Communication and Image Processing (VCIP), and a special sessions cochair of the 2006 IEEE International Conference on Multimedia and Expo (ICME). He is a fellow of the SPIE and a senior member of the IEEE.