A High Resolution Grammatical Model for Face Representation and Sketching

Zijian Xu, Hong Chen and Song-Chun Zhu
Departments of Statistics and Computer Science
University of California, Los Angeles
{zjxu, hchen, sczhu}@stat.ucla.edu

Abstract

In this paper we present a generative, high resolution face representation which extends the well-known active appearance model (AAM)[5, 6, 7] with two additional layers. (i) One layer refines the global AAM (PCA) model with a dictionary of learned face components to account for the shape and intensity variabilities of eyes, eyebrows, nose and mouth. (ii) The other layer divides the face skin into 9 zones with a learned dictionary of sketch primitives to represent skin marks and wrinkles. This model is no longer of fixed dimensions and is flexible for it can select the diverse representations in the dictionaries of face components and skin features depending on the complexity of the face. The selection is modulated by the grammatical rules through hidden ”switch” variables. Our comparison experiments demonstrate that this model can achieve nearly lossless coding of face at high resolution (256×256 pixels) with low bits. We also show that the generative model can easily generate cartoon sketches by changing the rendering dictionary. Our face model is aimed at a number of applications including cartoon sketch in non-photorealistic rendering, super-resolution in image processing, and low bit face communication in wireless platforms.

1. Introduction

Human faces have been extensively studied in vision and graphics for a wide range of tasks from detection[18, 16], classification[19, 10], tracking[6], expression[13], animation[11, 15], to non-photorealistic rendering (portrait and sketch)[3, 4], with both discriminative[17, 10, 3] and generative models[9, 6, 11] developed in the literature. The selection of a representation and model depends on two factors: (i) the objectives of the task and its precision request, and (ii) the resolution of the observable face images.

In this paper we present a *generative, grammatical, high resolution* face representation which extends the well-known active appearance model (AAM)[5, 6, 7] with two additional layers (see Fig.1 and Fig.8).

(i) *A face component layer*, which refines the global AAM (PCA) model with more detailed representations in 6-zones for the six facial components: two eyebrows, two eyes, nose and mouth. Each component has a set of diverse representations for the various types of eyes, mouths, noses and their topological configurations, such as open and close states. The representation for each component within its zone is a local AAM model with a various number of landmarks. The selection of the representation is modulated by the grammatical rules[1] through hidden ”switch” variables.

(ii) *A face skin layer*, which further refines the 6 component zones with sketch curves for the subtle differences in eye-lid, eye-shade, nostril, lips etc. In this layer, it divides the face skin into 9 zones (See Fig.6) with a learned dictionary of sketch primitives to represent possible skin marks and wrinkles. We adopt various prior models for sketches in these 15 zones and the number of sketch curves changes depending on the complexity of the faces.

As Fig.1 illustrates, our model achieves nearly lossless representation of high resolution images (256×256 pixels), at the same time it generates a face sketch useful for cartoon rendering. The computation is performed coarse-to-fine: we first infer the global AAM model and register the whole face. Then we refine the face components whose landmarks define the 9 skin zones. Thus we extract the skin sketches under such context with prior models.

Our model is aimed at a number of applications, such as low bit face communication in wireless platforms, cartoon sketch in non-photorealistic rendering, face editing and make-up in an interactive system, and super-resolution in image processing.

It is worth mentioning that one may not achieve such high resolution reconstruction by merely increasing the number of landmarks in the original AAM model, since the gloabl AAM model represents all human faces with the same number of landmarks and PCs, and is not sufficient for the vast variabilities exhibited in different ages, races, and expressions. Our comparison experiments (see Fig.9) confirms that our three layered representation is more ef-

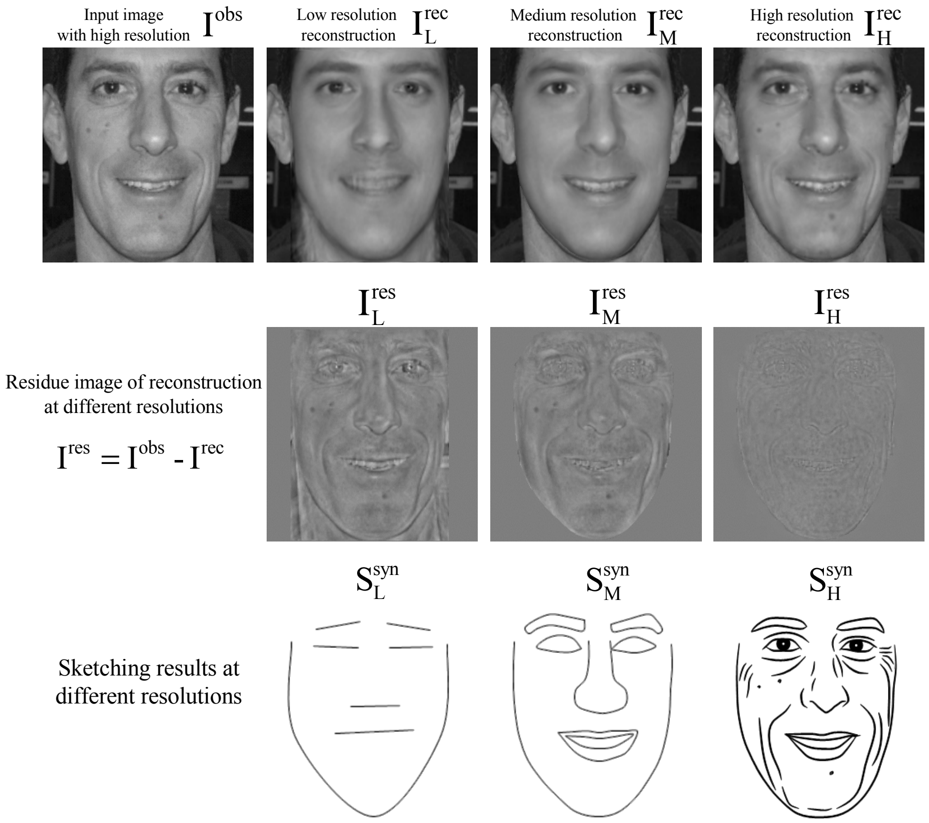

Input image with high resolution $\mathrm{I}^{\mathrm{obs}}$
Low resolution reconstruction $\mathrm{I}_{\mathrm{L}}^{\mathrm{rec}}$
Medium resolution reconstruction $\mathrm{I}_{\mathrm{M}}^{\mathrm{rec}}$
High resolution reconstruction $\mathrm{I}_{\mathrm{H}}^{\mathrm{rec}}$
$\mathrm{I}_{\mathrm{L}}^{\mathrm{res}}$
$\mathrm{I}_{\mathrm{M}}^{\mathrm{res}}$
$\mathrm{I}_{\mathrm{H}}^{\mathrm{res}}$
Residue image of reconstruction at different resolutions
$\mathrm{I}^{\mathrm{res}} = \mathrm{I}^{\mathrm{obs}} - \mathrm{I}^{\mathrm{rec}}$
$\mathrm{S}_{\mathrm{L}}^{\mathrm{syn}}$
$\mathrm{S}_{\mathrm{M}}^{\mathrm{syn}}$
$\mathrm{S}_{\mathrm{H}}^{\mathrm{syn}}$
Sketching results at different resolutions

Figure 1. Face high resolution image $\mathbf{I}^{\mathrm{obs}}$ of 256×256 pixels is reconstructed by the model in coarse-to-fine. The first row shows three reconstructed images $\mathbf{I}_{\mathrm{L}}^{\mathrm{rec}}$, $\mathbf{I}_{\mathrm{M}}^{\mathrm{rec}}$, $\mathbf{I}_{\mathrm{H}}^{\mathrm{rec}}$ in low, medium and high resolution respectively. $\mathbf{I}_{\mathrm{L}}^{\mathrm{rec}}$ is reconstructed by the AAM model, and the eyes, nose and mouth are refined in $\mathbf{I}_{\mathrm{M}}$ after adding the component AAM layer. The skins marks and wrinkles appear in $\mathbf{I}_{\mathrm{H}}^{\mathrm{rec}}$ after adding the sketch layer. The residue images are shown in the second row. The third row shows the sketch representation of the face with increasing complexity.

fective (i.e. less reconstruction error) than the AAM model over a test set of 150 face images (256×256 pixels in size), provided that both models use the same size of codebooks.

In comparison to the literature, our face component layer representation is different from the component-based[10] or fragment-based face recognition [17], the latter use local face features for face recognition in a discriminative manner in contrast to our goal of generative reconstruction of the face. Our face skin layer representation is different from the recent face sketching work [3, 4] which are example-based and construct the sketches through a discriminative mapping function using the image analogy technique in graphics. Our sketch rendering is different from graphics interactive system[2].

In the rest of the paper, we present the three-layer representation of the model and coarse-to-fine computation in Section 2, and then we report the experiments in Section 3. Section 4 concludes the paper with some further work.

2. Representation and Computation

The representation and algorithm is illustrated in Fig. 1 and Fig.8. We represent an observed image in low, medium and high three resolutions: $\mathbf{I}_{\mathrm{L}}^{\mathrm{obs}}$ (64×64 pixels), $\mathbf{I}_{\mathrm{M}}^{\mathrm{obs}}$ (128×128 pixels), and $\mathbf{I}_{\mathrm{H}}^{\mathrm{obs}}$ (256×256 pixels), and we compute the three hidden layer representation $W_{\mathrm{L}}, W_{\mathrm{M}}, W_{\mathrm{H}}$ sequentially through Bayesian inference. The dictionaries of PCs and the sketch primitives are treated as parameters of the generative model and learned through fitting the model to a set of 200 training images.

2.1 Layer 1: the low resolution AAM model

In the first AAM layer, all faces share the same number of landmarks. The AAM representation includes a set of principle components (denoted by $\mathrm{PCA}_{\mathrm{geo}}^{\mathrm{aam}}$) for the geometric deformations, and a set of principle components (denoted by $\mathrm{PCA}_{\mathrm{pht}}^{\mathrm{aam}}$ for the intensity (photometric) variabilities after aligning the landmarks. Therefore we have a

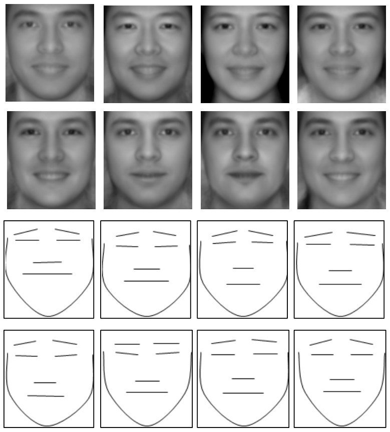


Figure 2. The first 8 PCs (plus mean) for intensity and geometric variations in the learned dictionary $\Delta_{\mathbf{I}}^{\mathrm{aam}}$ with 17 landmarks.

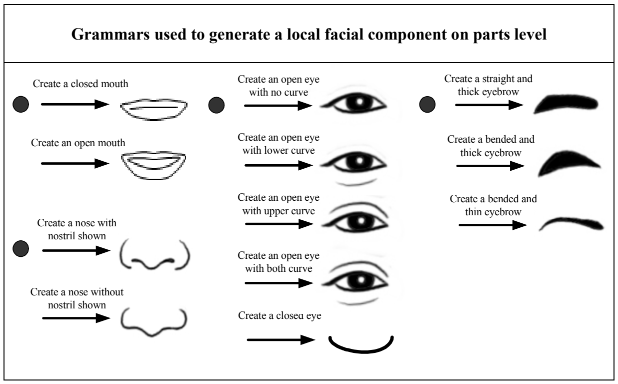

Grammars used to generate a local facial component on parts level
Create a closed mouth
Create an open mouth
Create a nose with nostril shown
Create a nose without nostril shown
Create an open eye with no curve
Create an open eye with lower curve
Create an open eye with upper curve
Create an open eye with both curve
Create a closea eye
Create a straight and thick eyebrow
Create a bended and thick eyebrow
Create a bended and thin eyebrow

Figure 3. Grammars used to generate local facial components with different templates.

dictionary of PCs[14] learned for the first layer,

$$\Delta_{\mathbf{I}}^{\mathrm{aam}} = \{\mathrm{PCA}_{\mathrm{geo}}^{\mathrm{aam}},\ \mathrm{PCA}_{\mathrm{pht}}^{\mathrm{aam}}\}$$

Fig.2 shows the first 8 components in $\mathrm{PCA}_{\mathrm{geo}}^{\mathrm{aam}}$ and $\mathrm{PCA}_{\mathrm{pht}}^{\mathrm{aam}}$ learned from 200 training images. We choose 17 landmarks for $\mathbf{I}_{\mathrm{L}}^{\mathrm{obs}}$ as the structures will be represented in other layers. Connecting the 17 landmarks properly, we obtain the low-resolution sketch representation. We will discuss and compare the number of landmarks and principal component in section of the model complexity experiment.

The hidden variable W_{L} includes variables for the global similarity transform T^{aam} and the coefficients $\alpha_{\mathrm{geo}}^{\mathrm{aam}}$ and $\beta_{\mathrm{pht}}^{\mathrm{aam}}$ for geometric and photometric PCs respectively.

$$W_{\mathrm{L}} = (T^{\mathrm{aam}}, \alpha_{\mathrm{geo}}^{\mathrm{aam}}, \beta_{\mathrm{pht}}^{\mathrm{aam}}).$$

Therefore, W_{L} can reconstruct an image $\mathbf{J}_{\mathrm{L}}^{\mathrm{rec}} = \mathbf{J}_{\mathrm{L}}^{\mathrm{rec}}(W_{\mathrm{L}}; \Delta_{\mathbf{I}}^{\mathrm{aam}})$ with the geometric and photometric PCs through the AAM model[6, 5]. The residue image is denoted by $\mathbf{I}_{\mathrm{L}}^{\mathrm{res}}$. Thus we have the first layer generative model,

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

$$W_{\mathrm{L}} = \arg\max p(\mathbf{I}_{\mathrm{L}}^{\mathrm{obs}} | W_{\mathrm{L}};\ \Delta_{\mathbf{I}}^{\mathrm{aam}}) p(W_{\mathrm{L}}).$$

The likelihood is a Gaussian probability following a noise assumption for the residue. The prior model is also Gaussian following the PCA assumptions[5]. The dictionary $\Delta_{\mathbf{I}}^{\mathrm{aam}}$ is treated as parameters of the generative model and learned through fitting (i.e. PCA) to the data.

2.2 Layer 2: the medium resolution model

In the second layer, we work on a medium size lattice Λ_{M} (128×128 pixels) and focus on six zones for the face components: two eyebrows, two eyes, one nose, and one mouth respectively,

$$\Lambda_1^{\mathrm{cp}}, \Lambda_2^{\mathrm{cp}}, ..., \Lambda_6^{\mathrm{cp}} \subset \Lambda_{\mathrm{M}}.$$

The solid curves show the six zones in Fig.6.(a). Within each zone, we adopt local AAM models for each face component. To count for different types of components and their status (see examples in Fig.4), we adopt 3 sets of AAMs for the eyebrows, and 5 sets for the eyes, 2 sets for the nose, and 2 sets for the mouth. The way to define the types is limited by the training data obtained. More detailed definition and therefore more types may be introduced while more complete dataset is available. The grammars to apply 12 different sets of local models are shown in Fig.3. For example, the different AAM models for eyes may have different number of landmarks and use different PCs for its geometric and photometric variations. Therefore we have a total of 12 pairs of PCs in the dictionary of the second layer representation,

$$\Delta_{\mathbf{I}}^{\mathrm{cp}} = \{\mathrm{PCA}_{\mathrm{geo}}^{\mathrm{cp},j},\ \mathrm{PCA}_{\mathrm{pht}}^{\mathrm{cp},j},\ j = 1, 2, ..., 12\}.$$

The 12 component models are learned in a supervised manner from 200 training face images. The selection of the model for each component is controlled by six switch variables $\ell_i, i = 1, 2, ..., 6$ in a stochastic grammar representation[1]. In fact our grammar is not context free, because the symmetry for the two eyes and eyebrows has to be taken into account. The hidden variable W_{M} in the medium layer includes the switches and the coefficients for the six components,

$$W_{\mathrm{M}} = (\ell^i, \alpha_{\mathrm{geo}}^i, \beta_{\mathrm{pht}}^i)_{i=1}^6.$$

The positions, orientations, and sizes of the components are inherited from the landmarks in layer 1 W_{L}. We denote the six zones by a sub-lattice

$$\Lambda_{\mathrm{cp}} = \cup_{i=1}^6 \Lambda_i^{\mathrm{cp}}.$$

The second layer model generates the ”refined” image

$$\mathbf{J}_{\mathrm{cp}}^{\mathrm{rec}} = \mathbf{J}_{\mathrm{cp}}^{\mathrm{rec}}(W_{\mathrm{M}}; \Delta_{\mathbf{I}}^{\mathrm{cp}}).$$

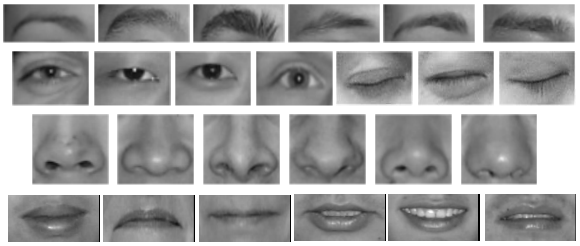


Figure 4. Different types and status of the local facial components, each is modeled by one of the 12 local models defined.

The reconstruction of the medium resolution image on lattice $\Lambda_{\rm M}$ is the following,

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

That is, pixels in the six component zones are generated by the refined models, while other pixels are generated by the low resolution global AAM model upsampled.

In summary, we have the second layer generative model,

$$\mathbf{I}^{\rm obs}_{\rm M} = \mathbf{J}^{\rm rec}_{\rm M}(W_{\rm L}, W_{\rm M};\ \Delta^{\rm aam}_{\mathbf{I}}, \Delta^{\rm cp}_{\mathbf{I}}) + \mathbf{I}^{\rm res}_{\rm M}.$$

$$W_{\rm M} = \arg\max p(\mathbf{I}^{\rm obs}_{\rm M} | W_{\rm L}, W_{\rm M}; \Delta^{\rm aam}_{\mathbf{I}}, \Delta^{\rm cp}_{\mathbf{I}}) p(W_{\rm M}).$$

Fig. 1 (3rd column) shows that the reconstructed face has much more sharpened eyes, nose, and mouth, and the residue image is less structured.

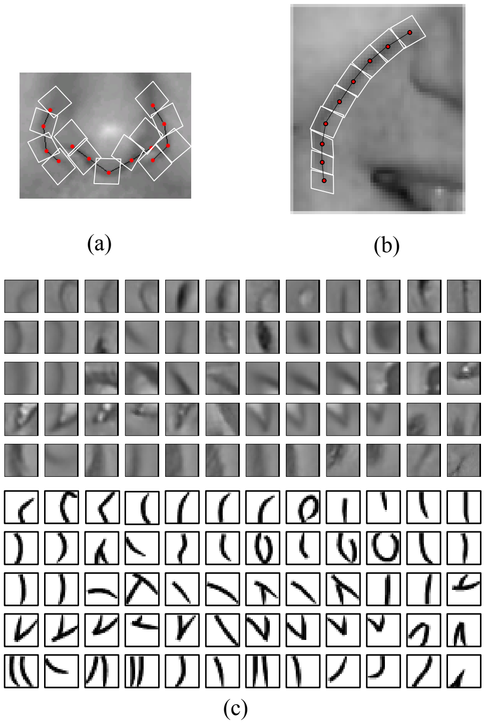
The likelihood is a Gaussian probability following a noise assumption for the residue. The prior model for each component is also Gaussian following the PCA assumptions for the components[5]. The dictionary $\Delta^{\rm cp}_{\mathbf{I}}$ is treated as parameters of the generative model and learned in a supervise way through fitting (i.e. PCA) to the data.

The inference of the "switch" variables $\ell_i, i = 1, 2, ..., 6$ is done through model comparison within each zones. For example, we select the best fitted eye representation among the 5 eye models, with a prior which is favor of the same model for the two eyes or the two eyebrows.

2.3 Layer 3: the high resolution sketch model

In the third layer, we further refine the 6 component with sketch curves for the subtle differences in eye balls, eye twinkles, eye-lid, eye-shade, nostril, wings of nose, lips etc. We also divide the face skin into 9 zones shown in Fig.6. The boundaries of these zones are decided by the landmark points computed in $W_{\rm L}$ and $W_{\rm H}$.

Our sketch representation has much more details than previous example-based face sketch method[4] or the face features used for expression classification[13]. In fact, these details are sometimes so subtle that one may not see them (even with human vision) unless they are viewed in the global context of the face. Fig.7 shows such example of the skin wrinkles which are nearly imperceptible but become quite prominent when they are put in the face image. This argues for the coarse-to-fine computation and model – a method that this paper is taking.



(a) (b) (c)

Figure 5. (a) Refinement on the nose by sketch primitive. (b). the sketch curve for a "smiling fold". Each rectangle in (a-b) represents a sketch primitive. (c) Examples in the dictionary of sketch primitives $\Delta^{\rm sk}_{\mathbf{I}}$ (above) and their corresponding strokes (below) in a cartoon sketch dictionary $\Delta^{\rm sk}_{\mathbf{S}}$.

Following the same notation in the medium resolution layer, we divide the high resolution lattice $\Lambda_{\rm H}$ (e.g. 256×256 pixels) into two parts: the sketch part $\Lambda_{\rm sk}$ where the image will be refined by a number of small image primitives, and the rest of the image $\Lambda_{\rm nsk}$ where there is no sketch is represented by the medium resolution through up-sampling.

The sketch part consists of many image primitives

$\Lambda^k_{\rm sk}, k = 1, 2, ..., K$. They are small rectangular windows (e.g. 7×7 pixels), and the number of primitives K is a variable depending on the medium resolution representation $W_{\rm M}$ and the image $\mathbf{I}^{\rm obs}_{\rm H}$.

Each primitive is an image patch with a small number ($2 \sim 3$) of control points, and thus with both geometric deformation and photometric variations. We collect a large set of image primitives by manually drawing the sketches on the 200 training images, and some examples are shown in Fig. 5. Then a data clustering was done to yield a dictionary of primitives in layer 3. In order to capture more

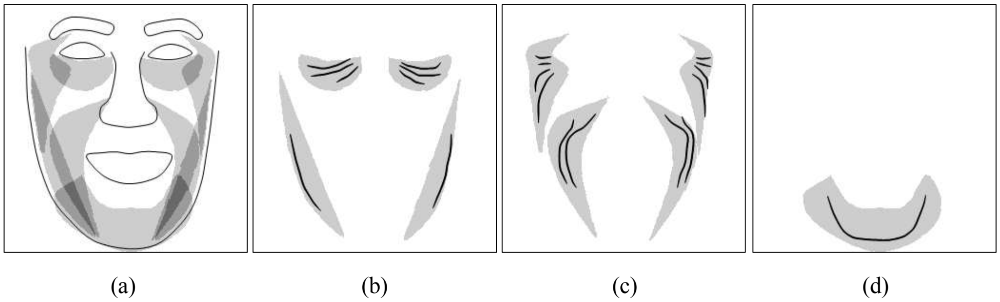


Figure 6. (a). 15 zones for detailed skin features. The 6 zone for the eyebrows, eyes, nose and mouth, and 9-zones for shaded skins areas where the wrinkles occur. The boundaries of these zones are decided by the landmarks computed in the low and medium resolution, and thus are inherited from $W_{\rm L}$ and $W_{\rm H}$. (b-c-d) typical wrinkles (curves) at the 9 skin zones. Strong prior models and global context are needed in order to detect the wrinkles.

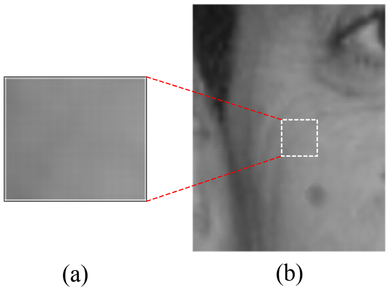


Figure 7. (a). A 15×15 (before zoom in) patch sampled from 256×256 face image; (b). The same local patch viewed in its global context — on a wrinkle.

details on skin, especially for detecting and reconstructing the skin marks (dark) or small secularity spots (highlight), we also labelled and trained a set of blob type of bases for the dictionary.

$$\Delta_{\mathbf{I}}^{\rm sk} = \{B_i : i = 1, 2, ..., N\}.$$

Each B_i is an image patch. Then the hidden variables in the 3rd layer $W_{\rm H}$ include the index ℓ_k for the primitive type, an affine transform t_k for positions, orientations and scales of these patches, and the photometric contrast α_k,

$$W_{\rm H} = (K, \{(\ell_k, t_k, \alpha_k) : k = 1, 2, ..., K\}).$$

Thus we generate the high resolution image in the sketchable part $\Lambda_{\rm sk}$,

$$\mathbf{J}_{\rm sk}^{\rm rec} = \mathbf{J}_{\rm sk}^{\rm rec}(W_{\rm H};\ \Delta_{\mathbf{I}}^{\rm sk}).$$

The final generative model at high resolution is,

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

That is, pixels in the sketch part are generated by the refined models, while other pixels are generated by the medium resolution model upsampled. Therefore,

$$\mathbf{I}_{\rm H}^{\rm obs} = \mathbf{J}_{\rm H}^{\rm rec}(W_{\rm M}, W_{\rm H};\ \Delta_{\mathbf{I}}^{\rm cp}, \Delta_{\mathbf{I}}^{\rm sk}) + \mathbf{I}_{\rm H}^{\rm res}.$$

$$W_{\rm H} = \arg\max p(\mathbf{I}_{\rm H}^{\rm obs} | W_{\rm M}, W_{\rm H}; \Delta_{\mathbf{I}}^{\rm cp}, \Delta_{\mathbf{I}}^{\rm sk}) p(W_{\rm H}).$$

Fig. 1 (4th column) shows that the reconstructed face has much more skin details and the residue is greatly reduced, such that the reconstruction $\mathbf{J}_{\rm H}^{\rm rec}$ is almost lossless.

The likelihood is a Gaussian probability following a noise assumption for the residue. The prior model for each component is also Gaussian following the clustering assumptions.

An ASM model [5] is trained for each of the ”structual” sketches like eye-lid, eye-shape or nostril, etc., which is initialized and constrained by $W_{\rm M}$ from previous layer in the inference process. Experiments shows fast convergence and accurate searching result.

To infer the sketches in the 9 zones, which have much more flexibility and sometimes locally almost imperceptible, we need to define the prior more carefully. As shown in Fig.6, a group of typical sketches are formed in each of the zone by learning of the labelled sketches, and $p(W_{\rm H})$ is accordingly defined, which favors the following properties.

- *Length L of the sketch is approximated by a poisson distribution $p(L = l) = (\lambda^l / l!)e^{-\lambda}$, where λ is specified by the typical sketches in the zone.*
- *Smoothness in scale, orientation and intensity pattern of two consecutive primitives $\{B_i, B_j\}$.*
- *Orientation and chance to appear for primitive B_i are biased by the neighboring typical sketches. That is, the orientation of B_i shall be more consistent with the closer typical sketch, and the closer B_i is to the typical sketches, the bigger chance is for it to appear.*
- *Spatial relationship between two sketches, e.g. two parallel sketches which are too close will be merged, or two consecutive short sketches which are too close will be connected.*

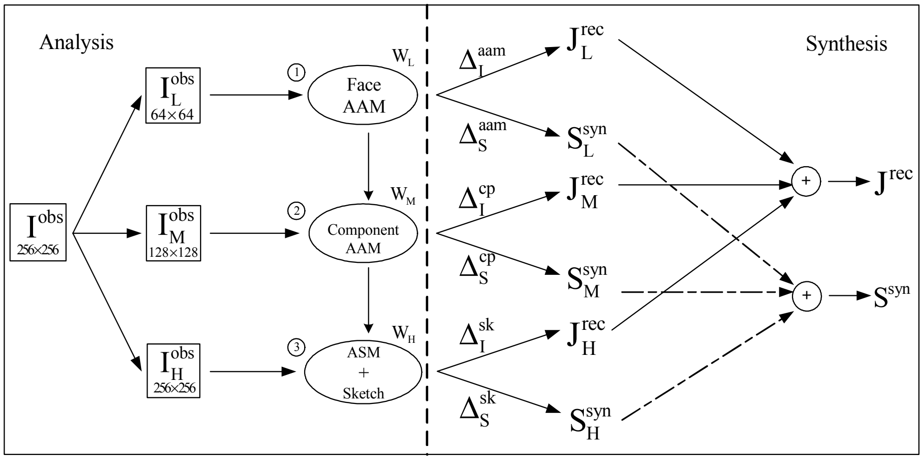


Figure 8. The diagram of our model and algorithm. The arrows represent the order of inference. Left panel is the three layers. Right panel is the synthesis steps for both high-resolution image reconstruction and face cartoon sketch using the generative model.

In each step of the sketch pursuit, a group of primitive candidates are proposed by the bottom-up methods, such as edge detection, and the existing sketches in the same zone. We decide whether to grow new primitive, make change to existing sketches or stop the process according to the posterior defined.

2.4 Generating the cartoon sketch S

Fig. 8 summarizes the generating process for the nearly lossless coding of the image with the code being

$$W = (W_{\mathrm{L}}, W_{\mathrm{M}}, W_{\mathrm{H}})$$

through three layers of occluding representations. The model uses three dictionaries of increasing details

$$\Delta_{\mathbf{I}} = (\Delta^{\mathrm{aam}}_{\mathbf{I}}, \Delta^{\mathrm{cp}}_{\mathbf{I}}, \Delta^{\mathrm{sk}}_{\mathbf{I}})$$

For each elements in these dictionaries, we always have a corresponding graph representation, shown in Fig.2, Fig.4 and Fig. 5. We call them the sketch dictionaries

$$\Delta_{\mathbf{S}} = (\Delta^{\mathrm{aam}}_{\mathbf{S}}, \Delta^{\mathrm{cp}}_{\mathbf{S}}, \Delta^{\mathrm{sk}}_{\mathbf{S}})$$

Thus by replacing the ”photo-realistic” intensity dictionaries $\Delta_{\mathbf{I}}$ with the sketch dictionaries $\Delta_{\mathbf{S}}$, we can generate a sketch over scales using the same generating steps. Some examples of the sketches are shown in Fig.1 and Fig.10.

Our sketch has more details than the state-of-the-art face sketch work[4], though there is still more work to do before rendering stylistic cartoons. We argue that it is much more convenient to define and change the style in this generative representation.

3. Experiments

To verify the framework we proposed, experiments were conducted based on 350 frontal face images chosen from different genders, ages and races — 200 for training and 150 for testing. All the images are resized to four different resolutions: 32×32, 64×64, 128×128 and 256×256 pixels respectively. The landmarks and sketches on the training are manually labelled.

In the first experiment, we report on the face reconstruction, learning of dictionaries, and sketching. Results are shown in Fig.1 and Fig.10.

In the second experiment, we compare the efficiency of the three models: (i) the 1-layer global AAM model with more landmarks and PCA components, (ii) the 2-layer component models, and (iii) the 3-layer model. To be fair, we measure the total description length (coding length) of the 200 images plus the size of the codebook.

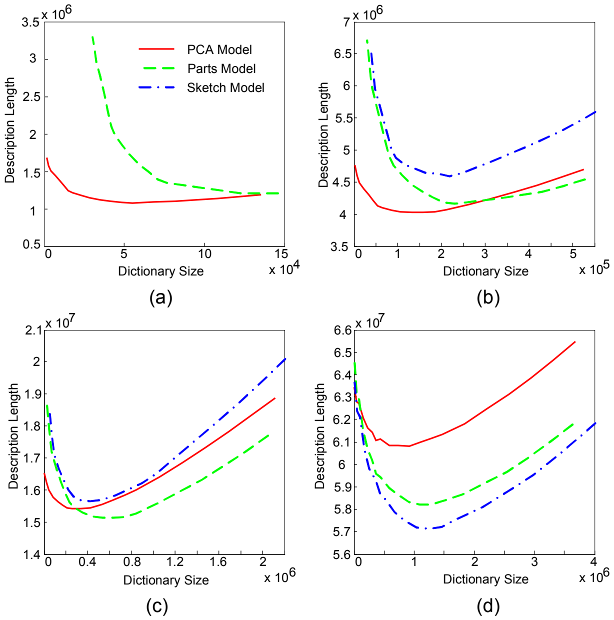
$$DL = L(\Omega_I; \Delta) + L(\Delta)$$

,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:

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

,where M denotes the number of data and $|\Delta|$ the dictionary size. For example, in the 1-layer global AAM model, it is the pixel number of mean-texture and eigen-texture used plus twice the point number of mean-shape and eigen-shape used. In Fig. 9, we plot how the coding length of the models changes with different dictionary sizes. At low resolution like 32×32 and 64×64, the DL of 1-layer global AAM model is shorter than 2-layer component model or 3-layer sketch model. At high resolution like 128×128 and 256×256, the component model and sketch model outperform respectively in the sense of coding efficiency. By applying the criterion of MDL(*minimum description length*),



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

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

we are able to select the most ”sufficient and compact” generative model for coding a given set of face data at certain resolution. It’s worth mentioning that there may be more appropriate criterion than MDL for certain objects like human faces. Other than minimizing the overall residue of the reconstructed face image, people may be more interested in keeping certain features on the face. For example, the wrinkles, see Fig.7 is very important for human to tell the age, gender or expression of a certain individual, while modeling them increases the coding length as much as the other strong facial features but reduce less residue. We may think these kind of features are of high ”sensitivity”. In the future, psychological experiments can be conducted to systemically study this phenomenon.

4. Summary and Future Work

In this paper we present a three layer generative model for high resolution face representation. The model incorporates diverse representations and allows varying dimensions for details and variabilities. In ongoing research, we are adding richer features including mustache, lighting variabilities. We’d also like to extend the model for stylish cartoon sketch in non-photorealistic rendering[4], super-resolution in image processing, and low bit face communication in wireless platforms through tracking the sketches over time. We argue that this high resolution representation should also improve other applications, such as to acquire precision 3D face model by stereo, expression analysis[13].

Acknowledgement

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

References

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

[2] S. Brenan, ”Caricature Generator”, *MIT Master Thesis*, 1982.

[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”, *Proc. of ICCV*, 2001.

[4] H. Chen and Z.Q. Liu, C. Rose, Y.Q. Xu, H.Y. Shum and D. Salesin, ”Example-based composite sketching of human portraits”, *Proc. 3rd Int’l Symp. on Non-photorealistic animation and rendering*, pp.95-153, 2004.

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

[6] T.F.Cootes, G.J. Edwards and C.J.Taylor. ”Active appearance models”, *Proceedings of ECCV*, 1998

[7] R.H. Davies, T.F. Cootes, C.Twining and C.J. Taylor, ”An Information theoretic approach to statistical shape modelling”, *Proc. British Machine Vision Conference*, pp.3-11, 2001

[8] K.S. Fu, *Syntactic Pattern Recognition*, Prentice-Hall, 1982.

[9] P.L. Hallinan, G.G. Gordon, A.L. Yuille, and D.B. Mumford, ”Two and three simensional patterns of the face”, *A.K. Peters, Natick*, MA, 1999.

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

[11] 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.

[12] 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.

[13] 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.

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

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

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

[17] S. Ullman and E. Sali, ”Object classification using a fragment-based representation”, *First IEEE Int’l Workshop, BMVC*, 2000.

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

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

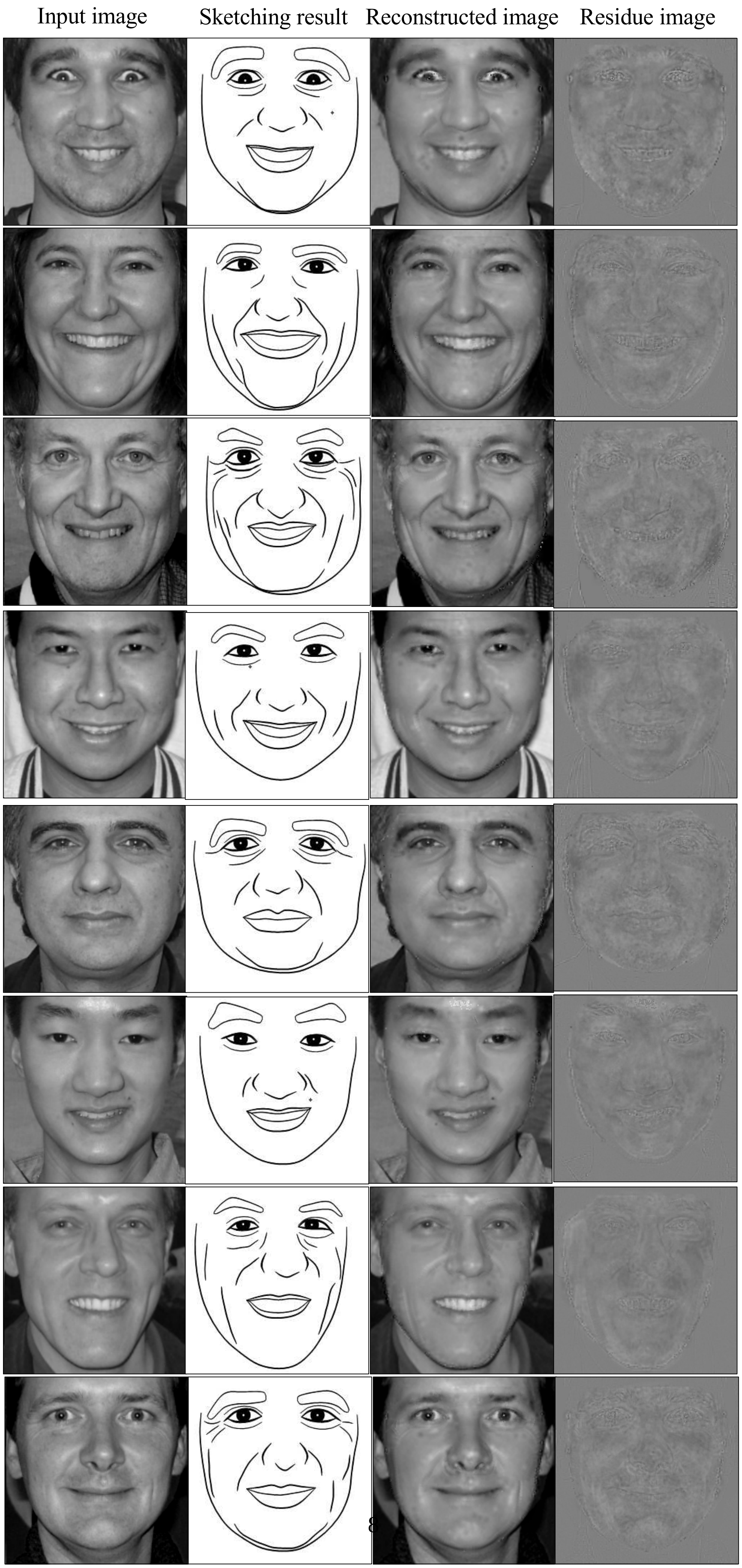


Figure 10. More results of reconstructed image, generated sketch and residue image of our model.