Multi-layered Expression Synthesis

JIA-RU LIN AND I-CHEN LIN
Institute of Multimedia Engineering
College of Computer Science
National Chiao Tung University
Hsinchu, 300 Taiwan

In this paper, a feature-point-driven expression editing and synthesis framework is proposed. While the extensively used blend shape methods suffer detail losing during image blending, the proposed multi-layer method can retain the flexibility and variety of geometry editing and preserve detail features as well. For low-frequency sub-bands, optimization-based blend shape is presented for large-to-mid scale synthesis. In addition, statistics-based feature matching and enhancement are proposed for high-frequency details. Our results show that the proposed methods are adequate to high-resolution expression synthesis and detail-preserved image editing.

Keywords: facial expression, blend shape, image editing, steerable pyramid, detail preservation

1. INTRODUCTION

While facial editing and synthesis are popularly used in computer animation, advertisement and other interaction systems, producing realistic facial details is still a labor-intensive work. Due to our familiarity with facial appearance, animators have to carefully reproduce the delicate changes. By contrast, motion capture (Mocap) techniques are widely used for semi-automatic facial motion acquisition. In order to capture feature markers' motions, dozens of markers are placed on control points of a subject's face. But, these techniques still can't capture the subtle portions, such as wrinkles, creases, or pores.

To acquire facial expressions, there are also many techniques, such as high resolution 3D laser scanners and face-scanning dome [1]. These approaches provide convincing results, but it is inefficient to acquire all appearances that we need and these devices are highly expensive. On the other hand, many data-driven approaches are proposed to generate novel facial expressions from a set of example appearances, such as blend shape [2-7]. The concept of blend shape is to represent each example expression in convex vector space. Novel expression can be generated by using convex combination of those example expressions, also called prototypes. Using blend shape can synthesize various facial expressions. However, the high-resolution facial detail information, such as wrinkles and pores, may blur during image blending. To tackle the problem, we propose multi-layer facial detail analysis and synthesis approach for synthesizing expressions with large-scale geometry and preserving details as well. For each prototype image, we utilize steerable pyramid [8, 9] to extract various frequency bands and resolution-aware methods are applied to each sub-band respectively to preserve image details. The framework

Received November 17, 2008; revised March 17, 2009; accepted April 2, 2009.
Communicated by Tong-Yee Lee.
[*] This paper was partially supported by National Science Council, Taiwan with Grant No. NSC 95-2221-E-009-164-MY3.

can be divided into two parts: offline processing and online multi-layer expression synthesis. Fig. 1 demonstrates the framework of our system.

The offline procedure aligns those acquired expressions with the neutral face for producing prototype images. Furthermore, motion of each aligned pixel is estimated during image alignment. To keep the regional detail variations, we cluster a face into different regions by normalized graph cuts (NCuts) [10, 11]. The offline procedure only needs to be done once.

The online part is to synthesize novel expressions from prototype expressions. In addition to dealing with the lowest sub-band by blend shape, we propose using *statistics-based feature matching* and *high-band enhancement* approaches for high-band data synthesis. Instead of the weighted blending in spatial domain, statistics-based feature matching and high-band enhancement retrieves the high-band data according to estimate statistics parameters, and therefore, preserves the detail resolution. Our results show that the proposed multi-layered method synthesizes more detailed and less artificial expressions than the blend shape method.

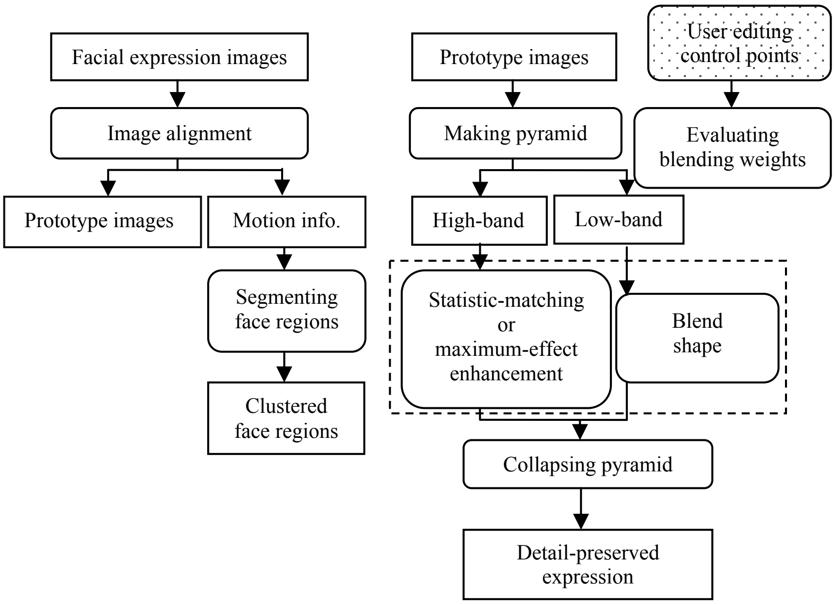


Fig. 1. The framework of multi-layer analysis and synthesis. Left: the offline process; right: the online process. Users only have to assign the control points and our system can automatically synthesize the corresponding expressions.

2. RELATED WORK

To analyze and synthesize novel facial images or surfaces, example-based methods, such as blend shape, are the most practical solutions. In the following sections, we introduce several state-of-the-arts, and also mention the detail acquisition and other related articles.

2.1 Novel Facial Expression Synthesis

Blend shape is extensively used to generate novel facial expressions from a set of example expressions, called prototypes. Pighin *et al.* [2] used convex combinations of the geometries and textures of example models. However, their expression editing system required users manually specifying the convex combination coefficients for novel facial expressions. Zhang *et al.* [3] developed a geometry-driven expression synthesis system to tackle the troublesome problem of parameter tuning. After assigning feature point positions of facial expressions, their system can automatically estimate the corresponding coefficient for novel facial textures.

Blanz and Vetter [4] proposed an impressive system for 3D face modeling from a single image. They transformed geometry and texture of hundreds of scanned faces to principal component analysis (PCA) [12] vector space and used linear combination to synthesize new facial surfaces. In 2003, Blanz *et al.* [5] further extended their method for photo-realism facial animation. After adding the facial expression vector to a neutral 3D face model, their system can transfer expressions across individuals.

Ezzat *et al.* [6] applied the blending concept for speech-driven animation. A sequence of training video is first projected to vector subspace space, called multidimensional morphable model (MMM), and they mapped the phonemes to projected coefficients. Given a new speech sequence, novel facial animation can be estimated by optimizing the MMM coefficient trajectories. In 2005, Vlasic *et al.* [13] proposed a more general multilinear model to transfer facial motion to another character by matrix factorization. They parameterized the space with various attributes (*e.g.*, identity, expression and viseme), and thus, those parameters can be used to drive 3D textured face mesh for a re-target character. However, a large number of normalized face scan data are required for evaluation of multilinear model.

The abovementioned methods, based on subspace data blending, are proven to be reliably for synthesizing novel facial expressions. Nevertheless, high-frequency details are usually lost during data blending.

On the other hand, Golovinskiy *et al.* [14] presented a statistical model for retargetting 3D facial details such as pores and wrinkles. They used Weyrich's acquisition system [1] to acquire high resolution facial geometry across different genders, ages and races. By matching displacement image with desired statistic properties and combining it with the base mesh, new face geometry with details can be generated. The proposed method is inspired by this work. While this work focused on geometry transfer, our method focuses on expression editing. Moreover, since they extracted and synthesized statistical details from tiles, their approach cannot deal with coarse wrinkle cross over those tiles.

Recently, Ma *et al.* [15] used polynomial functions to approximate correlations between strains of large-scale facial surfaces and mid-scale wrinkle undulations. This approach represented and controlled surface details in an efficient and compact form. Suconphunt *et al.* presented a 3D expression editing system by 2D feature contours [16]. The goal of this work is similar to ours. Nevertheless, they proposed directly combining best-matched prototype segments from database, and our multi-layered approach required much fewer prototypes to keep detailed wrinkles.

2.2 Acquisition of Facial Expressions

In addition to wrinkle textures, Acquisition of 3D facial detailed geometry is also attractive in graphics research. In Weyrich's research [1], they used face-scanning dome to measure high-resolution face model and skin reflectance. The equipment consisted of 16 digital cameras, 150 LED light sources, and a commercial 3D face-scanning system. Their measurement system can acquire high quality facial details; however, this method required high cost.

Zhang *et al.* [17] presented a system that construct high resolution and dynamic face models from video sequences. However, to reconstruct dynamic fine facial geometry from space-time stereo encounters inherent low-capture-rate and self-occlusion problem of structure light system. To rectify this problem, Bickel *et al.* [18] proposed a multi-scale representation and acquisition technique for animating high resolution facial geometry and wrinkles. They classified facial expressions from fine scale (*e.g.* pores, moles, freckles, spots) to coarse scale (*e.g.*, nose, cheeks, lips, eyelids), and used corresponding equipment to capture different scale features. In their system, a valley-shape wrinkle model is applied to analyze position and shape of wrinkles from intensity variations in video. In 2008, Bickel *et al.* [19] further extended their method for animation. Their hybrid animation considered facial geometry as large-scale motion and fine-scale motion. They computed the large-scale motion by using the same linear shell deformation [18], and interpolated fine-scale facial details by radial-basis functions.

3. FACE SEGMENTATION

Our goal is to synthesize expressions with limited prototypes. For producing more various facial expressions, we propose partitioning a face into meaningful regions for the consequent synthetic procedure. Therefore, we first decompose a normalized face into 64 × 64 grids, and evaluate the correlation of motion between every two grids. With motion correlations, Normalized Cuts (NCuts) [10, 11] are then used to segment face into regions. The detailed steps are presented in the following sub-sections.

3.1 Pre-processing

After recording images with different facial expressions as our prototypes, without lose of generality, we select the first image or preparative expression as the neutral face. All other images are aligned with the neural face to avoid blurred or ghost effects during image blending. Users first assign the corresponding feature points and lines in each image, and the global head movement can be removed. Image warping [20] is then applied to align unassigned pixels with the neutral face. We assume the pixel displacement between aligned and unaligned images as the motion of that pixel.

After evaluating motion information of all prototype images, we use histogram equalization [21] to align the facial skin colors. At present, our prototype expressions are selected empirically. The selection can be replaced by automatically clustering expressions from video frames. Our 22 prototype images are shown in Fig. 2.

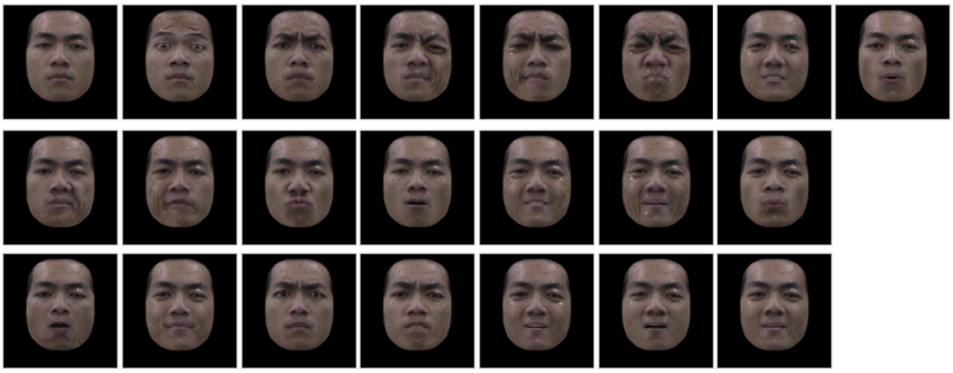
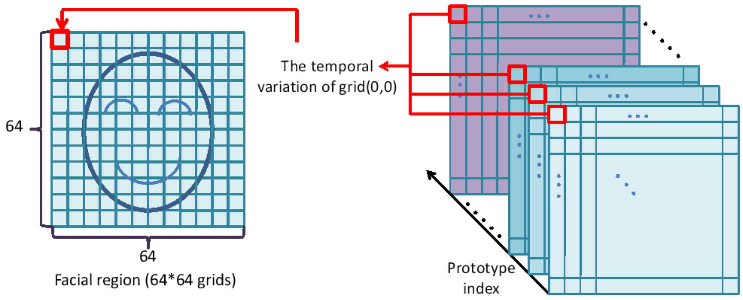


Fig. 2. Twenty two prototype images with various expressions. The resolution of each prototype image is 512 × 512 pixel^2.

3.2 Correlation Analysis

To reliably cluster a face into regions and for further statistic analysis, we first decompose face into 64 × 64 grids and use the statistic parameters in grids for data grouping. As shown in Fig. 3, we consider variation of those tiles at the same grid index but different prototype images as temporal variation The orders do not affect our analysis. We define the motion of the grid *a* at *t* as the average motion of pixels in the grid *a* at prototype *t*.



The temporal variation of grid(0,0)
64
64
Facial region (64*64 grids)
Prototype index

Fig. 3. The tiles in the same position among different prototype images can be regarded as the temporal variations of the corresponding grid.

The correlation between gird *a* and *b* can be defined as:

$$Correlation(a,b) = \frac{1}{n-1}\sum_{t=1}^{n}(\frac{a_t-\overline{a}}{S_a})(\frac{b_t-\overline{b}}{S_b}), \tag{1}$$

where a_t and b_t are the motion of grid *a* and grid *b* at time *t* (or prototype *t*). $\overline{a}$ and $\overline{b}$ are the average motions of grid *a* and grid *b* among all prototype images respectively. S_a and S_b are the standard deviation of grid *a* motion and grid *b* motion. Using the correlation formula (Eq. (1)), we can evaluate the correlation between each pair of grids. Since our human faces are nearly symmetric, in addition to the original prototypes, we also included mirrored prototypes for face segmentation. Therefore, the distributions of our segmented regions are symmetric, which is closer to users' cognition, but these regions are synthesized respectively.

3.3 Face Segmentation Based on Normalized Cuts

After evaluating correlation between each pair of grids, we use normalized cuts [10, 11] to partition face into regions of adequate sizes and shapes.

Normalized cut is aimed at minimizing the disassociation between two groups:

$$Ncut(A, B) = \frac{cut(A, B)}{asso(A, V)} + \frac{cut(A, B)}{asso(B, V)}, \tag{2}$$

where the A and B are two disjoint sets, $cut(A, B)$ is the total weight of edges that have been removed in partitioning procedure. $asso(A, V)$ is the total connection from nodes in A to all nodes in the graph, and $asso(B, V)$ is similar defined.

Our purpose is to segment facial grids into different groups. Therefore, we regard facial grids as nodes and set up a fully-connected weighted graph $G = (V, E)$. The edge weight w_{ij} between grid i and grid j can be defined as:

$$w_{ij} = e^{-\|1-Correlation(i,j)\|}. \tag{3}$$

After evaluating the weight value, we can set $\boldsymbol{W}$ as an $N \times N$ symmetrical matrix with $W(i, j) = w_{ij}$, N is the amount of all grids. And let $\boldsymbol{D}$ be an $N \times N$ diagonal matrix with d on its diagonal, where

$$d(i) = \sum_j w(i, j).$$

Then we minimize the normalized cut by solving the generalized eigenvalue system:

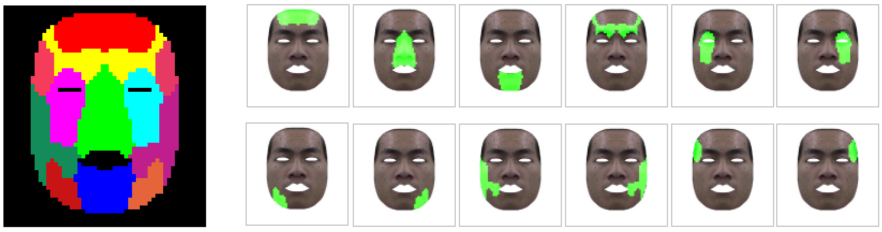
$$(D - W)y = \lambda Dy. \tag{4}$$

Instead of recursive binary cuts, we use the subspace of n top eigenvectors and K-means clustering for more efficient K-way partition.

In order to determine a reasonable number of clusters, we define a measurement function to keep the balance between cluster number and the correlation within each group:

$$\left\langle n^* \right\rangle = \arg\min\{k_1 \times n + k_2 \times \ (1 - cor)\}, \tag{5}$$

where n is the number of clusters, cor is the average correlation among all groups, k_1 and k_2 are the parameters for adjusting the influence of each term. By minimizing the measurement function, we can determine the number of face regions. Fig. 4 shows the clustering result. In the synthesizing procedure, we deal with each cluster respectively.



(a) (b)

Fig. 4. (a) The result of face clustering. Grids marked by the same color belong to the same cluster; (b) Regions of a face segmented by normalized cuts.

4. MULTI-LAYER FACIAL DETAIL SYNTHESIS

After segmenting a face into motion-consistency regions, existing data-driven methods, *e.g.* feature-point-driven blend shape [3], can automatically synthesize the corresponding textures according to feature point's configuration. Nevertheless, high-frequency details, such as pores and wrinkles, usually blur during blending process. Therefore, we propose improving the blend shape process by a *multi-layer analysis and synthesis* approach to preserve the facial details.

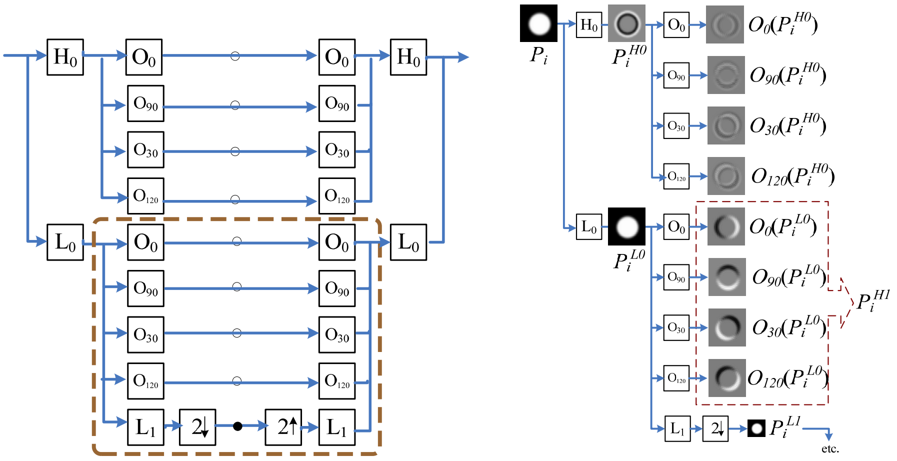
Our main concept is first separating grids of all prototypes into various sub-band images, including high-pass images, low-pass images, and various orientation sub-bands. Next, for a novel expression, we evaluate high sub-band images by our detail-preserved methods and synthesize low sub-band images by optimization-based blend shape. Finally, we combine those processed components in each sub-band and can obtain novel facial expressions.

4.1 Sub-band Extraction by Steerable Filter

We modify the framework of steerable pyramid [8, 9] for sub-band decomposition and integration. Steerable pyramid is similar to the well-known Laplacian pyramid, and can decompose an image into several frequency bands. Moreover, it further divides each frequency band into a set of orientation bands. Since the steerable pyramid is self-inverting, we can apply the same filter for image reconstruction.

Fig. 5 shows our modified framework of steerable pyramid. For preserving more high frequency information, we further decompose the high-band of each prototype image into four orientation sub-bands. Besides, we adopt the first derivative of the 2-dimensional, circularly symmetric Gaussian functions, rotating 0°, 90°, 30°, and 120° around horizontal, as steerable filters for sub-band extraction.

Fig. 5 (b) shows an example of decomposing an image P_i and it also demonstrates our symbol definition for each sub-band. In our multi-layer analysis and synthesis procedure, we use a set of prototype images $\{P_i\}_{i=1 \text{ to } m}$, m is the number of all prototypes. Each prototype image P_i is decomposed by high-pass/low-pass split filter and then generate high sub-band P_i^{H0} and low sub-band P_i^{L0}. Let $O_0(.)$ represents the operator of horizontal steerable filter. $O_{90}(.)$, $O_{30}(.)$, and $O_{120}(.)$ represent the rotated horizontal operator by 90°, 30°, and 120° respectively. By decomposing level j low-band P_i^{Lj}, we can produce high-band P_i^{Hj+1} and low-band P_i^{Lj+1} of level $j+1$. High-band P_i^{Hj+1} can be re-



(a) (b)

Fig. 5. (a) The modified framework of steerable pyramid in our multi-layer facial detail synthesis procedure. The left-hand side of the diagram is analysis part; the right-hand side is synthesis part. Each square box represents convolution or down/up sampling operations: H_0 is a high-pass filter, L_i are low-pass filters of level i sub-band and O_j are band-pass filters in different orientation. The hollow circles represent the decomposed sub-band images. The pyramid can construct recursively by repeating the process enclosed by the dashed rectangle at the location of solid circle; (b) The illustration of a steerable pyramid and symbol definition.

garded as composing of different orientation components: $O_0(P_i^{Lj})$, $O_{90}(P_i^{Lj})$, $O_{30}(P_i^{Lj})$, and $O_{120}(P_i^{Lj})$.

By comparing reconstructed image with the ground truth, our experiments showed that a 3-level pyramid provided satisfactory intensity ranges and details in analysis and synthesis of prototypes with 512 × 512 pixels. More frequency levels can barely improve the quality. Therefore, our modified framework of steerable pyramid decomposes all prototypes into 4 orientations at 3 levels of frequency bands, and employs those sub-bands for synthesizing novel facial expressions.

In order to synthesize detail-preserved expressions and reduce blurred results, optimization-based blend shape is only applied to the lowest sub-band P_i^{L3} and other high-band images are processed by statistic-matching-based methods.

4.2 Feature-Point-Driven Blend Shape

We assume that facial geometry has high relation with facial appearance. That is, the positions of control points between similar facial expressions are with high similarity. By presenting the geometry and appearance of prototype images in vector space, novel facial expression can be generated from convex combination of prototype images with proper blending weights. Therefore, we represent the ith expression E_i as $E_i = (G_i, P_i)$, where G_i is geometry (position configuration of feature points) and P_i is the prototype image i. Let $H(E_0, E_1, \ldots, E_m)$ be the space of all possible convex combination among all example expressions, *i.e.*,

$$H(E_0, E_1, \ldots, E_m) = \left\{ \left(\sum_{i=0}^{m} w_i G_i, \sum_{i=0}^{m} w_i P_i \right) | \sum_{i=0}^{m} w_i = 1, \text{and } w_0, w_1, \ldots, w_m \geq 0 \right\}. \quad (6)$$

Therefore, novel expression E^{new} can be represented by a particular set of w as:

$$E^{new} = (G^{new}, P^{new}), \text{ where } G^{new} = \sum_{i=0}^{m} w_i G_i, P^{new} = \sum_{i=0}^{m} w_i P_i. \quad (7)$$

Since we separate a face into regions for more variety during synthesis, we include notation of region R in Eq. (6) as:

$$H^R(E_0^R, E_1^R, \ldots, E_m^R) = \left\{ \left(\sum_{i=0}^{m} w_i^R G_i^R, \sum_{i=0}^{m} w_i^R P_i^R \right) | \sum_{i=0}^{m} w_i^R = 1, \text{and } w_0^R, w_1^R, \ldots, w_m^R \geq 0 \right\}, \quad (8)$$

where the G_i^R denote the vector of E_i's control point positions within or on the boundary of R. P_i^R denotes the region R of prototype image P_i. w_i^R is the blending weight for region R of prototype i. Accordingly, each region R of the synthetic image can be generated by:

$$E^R = (G^R, P^R), \text{ where } G^R = \sum_{i=0}^{m} w_i^R G_i^R, P^R = \sum_{i=0}^{m} w_i^R P_i^R. \quad (9)$$

4.3 Weight Evaluation by Optimization

In conventional blend shape, weights w_i or w_i^R are assigned by users, but it is not intuitive. Since we consider that facial appearance is highly related to facial geometry, we only ask users to adjust positions of feature points, *e.g.* eye or mouth corners. Our system can estimate the weight according to geometry features, and use the weight to synthesize novel facial appearance.

Let $G^R{}_{new}$ denotes region R's control points position of novel expression. Given $G^R{}_{new}$, we want to find the blending weight for interpolating $G_0{}^R, \ldots, G_m{}^R$. This problem can be formulated as an optimization problem as the weight evaluation in [3]:

$$\begin{aligned} &\text{Minimize: } \left(G_{new}^R - \sum_{i=0}^{m} w_i^R G_i^R \right)^T \left(G_{new}^R - \sum_{i=0}^{m} w_i^R G_i^R \right), \\ &\text{Subject to: } \sum_{i=0}^{m} w_i^R = 1, w_i^R \geq 0 \ \text{ for } i = 0, 1, \ldots, m. \end{aligned} \quad (10)$$

After optimizing the blending weights of each region by a sequential quadratic programming (SQP) method, we can apply those coefficients for synthesizing novel facial appearances.

4.4 Statistics-based Feature Matching

Applying blend shape to novel expression synthesis is effective, especially with only few prototype images. However, when blending a large number of prototype images, high frequency details at pores or wrinkles, usually lost during blending process.

On the other hand, if we apply the blending as parameter evaluation and synthesize high-band images by appropriate data indexing, the high frequency features will be preserved. Therefore, we propose using statistics-based matching for high-bands of prototypes.

The simplest method is to extract histogram of each high-band grid and select the gird with the closest histogram for synthesis in high-band. However, storing and analyzing those histograms is burdensome. Golovinskiy *et al.* [14] found the width of histogram function is a dominant factor. Therefore, we adopt the concept and use the standard deviation of each grid to substitute histogram of each grid.

In the offline stage, we evaluate the standard deviations of prototype high band data. To synthesize high-band of novel expressions, the feature-point driven weights are used to blend standard deviation values of a grid among prototypes, and we select the prototype grid with the closest standard deviation value for synthetic expression. $\sigma(P_i^f, k)$ represents the standard deviation of grid k belonging sub-band f of prototype i. The standard deviation of corresponding synthesized grid *Sigma* can become:

$$Sigma = \sum_{i=0}^{m} w_i^R \sigma(P_i^f, k), \text{ where gird } k \in R. \tag{11}$$

The grid k of the synthesized image in sub-band f can be determined by finding the best matched prototype i:

$$\langle i^* \rangle = \arg\min_i (Sigma - \sigma(P_i^f, k)), \text{ for } i = 0, 1, \ldots, m. \tag{12}$$

4.5 Maximum-effect Enhancement

By selecting a proper grid for each synthetic image high-band, we can maintain the high frequency information in synthetic results. However, when dealing with prototype images with highly uneven standard deviation values, using the closest standard deviation as criterion may cause discordant synthetic results. Therefore, we further propose using maximum-effect enhancement for such situations.

Instead of applying the prototype with closest standard deviation, maximum-effect enhancement selects the prototype high-band with maximum blending weight as the corresponding synthetic high-band of the region. After blending the lowest sub-band and combining high-band with maximum blending weight, we can synthesize novel facial expressions by collapsing pyramid.

Due to our particular synthesis methods for high-band and low-band respectively, the proposed system can maintain more high frequency information for novel facial expressions with photorealistic facial details, such as wrinkles and pores. It also retains global geometry deformation, and mid-scale features, such as the wrinkle cross over forehead. In our experience, when the largest estimated weight is more than 0.65, maximum-effect enhancement is empirically appropriate for high-band synthesis. Otherwise, statistics-based feature matching provides better results.

5. EXPERIMENT AND RESULTS

In our first experiment, we captured various facial expressions of a performer and selected those frames with representative expressions as our prototype images. We put 41 markers on the performer's face and use motion capture device to trace the motion of markers for more precise alignment. The positions of feature points were selected according to motion variations on the face. After determining prototype images, we used image warping to align all example expressions with the neutral face.

As mentioned in section 3, to synthesize more various expressions from a few example expressions, we used normalized cuts to segment a face into regions. Consequently, a face was divided into 12 different regions, as show in Fig. 4. A region can only be affected by feature points within its territory.

In our research, we employed the modified steerable pyramid to analyze image information for synthesizing high-band and low-band of novel expression respectively. We integrated the statistics-based feature matching and maximum-effect enhancement for maintaining high-band information in synthetic procedure. We evaluated our system by cross-validation, where the ground-truth images are not included in the prototype image set. Fig. 6 demonstrates a comparison between our results with those of feature-point-driven blend shape [3]. Our results maintain more details around the cheeks and corners of mouth. Since our goal is mainly at wrinkles or creases, we omit the eyes and lips, which require more feature points for editing. Fig. 7 demonstrates synthetic texture of another performer without post-warping.

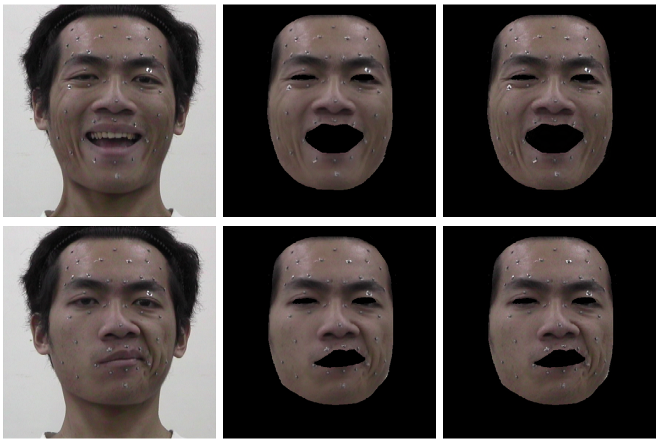
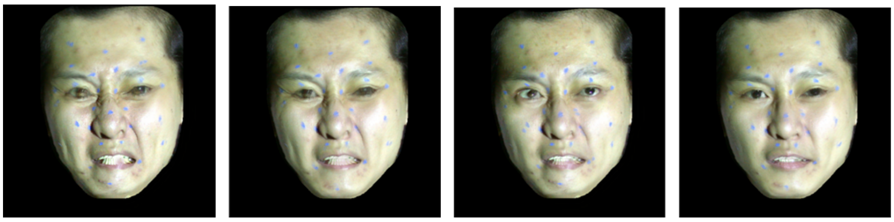


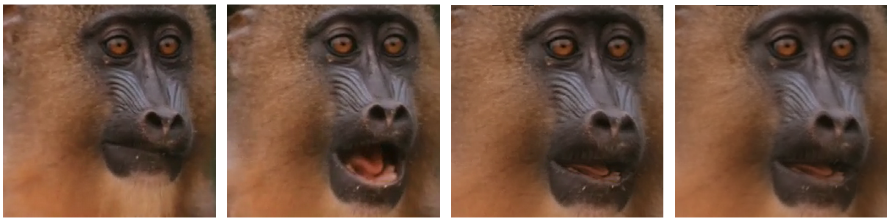
Fig. 6. The left images are the ground truth expressions. The center images are the result by blend shape. The right images are results by our approach. The synthesized images are post-warped to fit feature points.

Fig. 8 demonstrates an example of Mandrill. Figs. 8 (a) and (b) are Mandrill's expression with mouth closing and mouth opening respectively. We use those images as prototypes for synthesizing Mandrill's expression with half opening mouth. Figs. 8 (c) and (d) are synthesized results by our approach and blend shape respectively. The result



(a) Our approach. (b) Blend shape. (c) Our approach. (d) Blend shape.

Fig. 7. Comparison of expression textures, where feature points are aligned to the neutral face without post-warping; (a) and (c) are synthetic results by our approach; (b) and (d) are the corresponding synthetic results by blend shape. Under the same blending weight, using our approach can generate more apparent facial wrinkles than those by blend shape.



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

Fig. 8. Detail-preserved image editing[1]; (a) and (b) are prototype images; (c) and (d) are synthetic results by our approach and blend shape respectively.

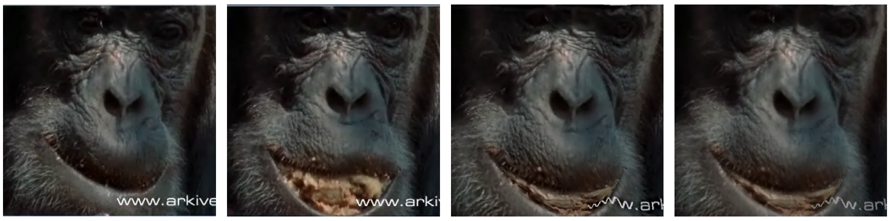


Fig. 9. Editing expression of chimpanzee[1]; (a) and (b) are prototype images of chimpanzee; (c) and (d) are synthetic results by our approach and blend shape respectively.

by our approach is with more details than that of blend shape, especially around the region of Mandrill's cheeks and mouth. Fig. 9 shows another example, using our approach can maintain more obvious wrinkle and contour of chimpanzee's face. Please refer to authors' web pages for demo images at higher resolutions.

Furthermore, to evaluate the effectiveness of the proposed method, we performed user evaluation for three issues: *detail quality*, *faithfulness*, and *artificiality*. Eleven subjects with experiences in image editing participated in these tests.

The first issue "detail quality" is to evaluate how much detail is preserved. We provided prototype images blurred by various magnitudes of Gaussian filtering as reference images. The scores of original images were 10; scores of images blurred by 0.3-pixel-radius Gaussian filtering were 9; scores of images blurred by 0.6-pixel-radius Gaussian filtering were 7, and so forth. Subjects were asked to evaluate 12 pairs of synthesized

[1] Image source: http://www.arkive.org/.

images by the proposed method and blend shape with a random order. In the second issue “faithfulness”, subjects were asked to compare 12 pairs of the synthesized images with the ground-truth. Score 10 was perfect; 8 was satisfactory; 6 was acceptable; 4 was artificial, and so forth.

In the third test “artificiality”, 10 real prototype images, 10 images by our methods and 10 images by blend shape were mixed and shown one-by-one in a random order. Subjects were asked to distinguish whether an image was synthesized. The score was the number of images that were regarded as artificial ones.

The results of three tests are shown in Table 1. For detail quality and faithfulness, the proposed method got better grades than the blend shape. For artificiality evaluation, even 3 of 10 ground-truth images were regarded as artifact. Our score 3.27 was close to the real data set.

Table 1. User evaluation of images by our method, blend shape and ground-truth.

	Detail quality	Faithfulness	Artificiality
ground-truth	–	–	3
The proposed method	7.87	7.62	3.27
Blend shape	7.10	7.07	3.73

6. CONCLUSION AND FUTURE WORK

The goal of the proposed work is to synthesize feature-point-driven expression with only a few examples. A novel detail-preserved multi-layer framework is herein presented. Given a set of prototype images of a subject, normalized-cut-based clustering is first applied to segment the face according to motion correlations. The segmented face is further decomposed into multiple frequency and orientation sub-bands. For the lowest-level-band, optimization-based blend shape is utilized to synthesize the large to mid scale geometry features, *e.g.* bulges or coarse creases. Statistic-based feature matching and enhancement are proposed to synthesize detail wrinkles and other highly-textural features, where high-frequency properties are retained.

Our experiments and user evaluation show that more detail features are preserved by the proposed methods than by blend shape. The proposed method can be extended to 3D surface editing. It can also be extended to synthesize more detailed lips and eyes by more precise contour mapping. In our future work, we plan to combine our approach with spatial-temporal constrains and alignment for detail-preserved facial animation.

REFERENCES

1. T. Weyrich, W. Matusik, H. Pfister, B. Bickel, C. Donner, C. Tu, J. McAndless, J. Lee, A. Ngan, H. W. Jensen, and M. Gross, “Analysis of human faces using a measurement-based skin reflectance mode,” *ACM Transactions on Graphics*, Vol. 25, 2006, pp. 1013-1024.
2. F. Pighin, J. Hecker, D. Lischinski, R. Szeliski, and D. H. Salesin, “Synthesizing realistic facial expressions from photograph,” in *Proceedings of ACM Siggraph*, 1998,

pp. 75-84.

3. Q. Zhang, Z. Liu, B. Guo, D. Terzopoulos, and H. Y. Shum, "Geometry-driven photorealistic facial expression synthesis," *IEEE Transactions on Visualization and Computer Graphics*, Vol. 12, 2006, pp. 48-60.
4. V. Blanz and T. Vetter, "A morphable model for the synthesis of 3D faces," in *Proceedings of Siggraph*, 1999, pp. 187-194.
5. V. Blanz, C. Basso, T. Poggio, and T. Vetter, "Reanimating faces in images and video," in *Proceedings of Eurographics*, Vol. 22, 2003, pp. 641-650.
6. T. Ezzat, G. Geiger, and T. Poggio, "Trainable videorealistic speech animation," *ACM Transactions on Graphics*, Vol. 21, 2002, pp. 388-398.
7. Z. Deng, P.Y. Chiang, P. Fox, and U. Neumann, "Animating blendshape faces by cross-mapping motion capture data," in *Proceedings of ACM Siggraph Symposium on Interactive 3D Graphics and Games*, 2006, pp. 43-48.
8. D. J. Heeger and J. R. Bergeny, "Pyramid-based texture analysis/synthesis," in *Proceedings of ACM Siggraph*, 1995, pp. 229-238.
9. W. T. Freeman and E. H. Adelson, "The design and use of steerable filters," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 13, 1991, pp. 891-906.
10. J. Shi and J. Malik, "Normalized cuts and image segmentation," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 1997, pp. 731-737.
11. J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22, 2000, pp. 888-905.
12. I. T. Jollife, *Principal Component Analysis*, Springer-Veriag, New York, 2002.
13. D. Vlasic, M. Brand, H. Pfister, and J. Popovi, "Face transfer with multilinear models," *ACM Transactions on Graphics*, Vol. 24, 2005, pp. 426-433.
14. A. Golovinskiy, W. Matusik, H. Pfister, S. Rusinkiewicz, and T. Funkhouser, "A statistical model for synthesis of detailed facial geometry," *ACM Transactions on Graphics*, Vol. 25, 2006, pp. 1025-1034.
15. W. C. Ma, A. Jones, J. Y. Chiang, T. Hawkins, S. Frederiksen, P. Peers, M. Vukovicz, M. Ouhyoung, and P. Debevec, "Facial performance synthesis using deformation-driven polynomial displacement maps," *ACM Transactions on Graphics*, Vol. 27, 2008, pp. 121:1-10.
16. T. Suconphunt, Z. Mo, U. Neumann, and Z. Deng, "Interactive 3D facial expression posing through 2D portrait manipulation," in *Proceedings of International Conference on Graphics Interface*, 2008, pp. 177-184.
17. L. Zhang, N. Snavely, B. Curless, and S. M. Seitz, "Spacetime faces: high resolution capture for modeling and animation," *ACM Transactions on Graphics*, Vol. 23, 2004, pp. 548-558.
18. B. Bickel, M. Botsch, R. Angst, W. Matusik, M. Otaduy, H. Pfister, and M. Gross, "Multi-scale capture of facial geometry and motion," *ACM Transactions on Graphics*, Vol. 26, 2007, Article 33.
19. B. Bickel, M. Lang, M. Botsch, M. A. Otaduy, and M. Gross, "Pose-space animation and transfer of facial details," in *Proceedings of ACM Siggraph/Eurographics Symposium on Computer Animation*, 2008, pp. 57-66.
20. T. Beier and S. Neely, "Feature based image metamorphosis," in *Proceedings of Siggraph*, 1992, pp. 35-42.

21. T. Acharya and A. K. Ray, *Image Processing: Principles and Applications*, Wiley-Interscience, NJ, 2005.

Jia-Ru Lin (林家如) received the B.S. degree from the Department of Computer Science and Information Engineering, National Central University, Taiwan, in 2006. In 2008, she received the M.S. degree from Institute of Multimedia Engineering, National Chiao Tung University, Taiwan. Her research interests include computer graphics, computer vision, and image processing.

I-Chen Lin (林奕成) is an Assistant Professor in the Department of Computer Science, National Chiao Tung University, Taiwan. His research interests include computer graphics and animation, especially in facial and character animation, motion capture, and image-based modeling. He received B.S. and Ph.D. degrees in Computer Science from National Taiwan University in 1998 and 2003, respectively. He is a member of ACM SIGGRAPH and IEEE.