國立交通大學

多媒體工程研究所

碩士論文

多層次臉部細紋之分析與合成

Analysis and Synthesis of Multi-layered Facial Details

研究生：林家如

指導教授：林奕成　助理教授

中華民國九十七年七月

多層次臉部細紋之分析與合成
Analysis and Synthesis of Multi-layered Facial Details

研究生：林家如　　Student：Jia-Ru Lin

指導教授：林奕成　　Advisor：I-Chen Lin

國立交通大學
多媒體工程研究所
碩士論文

A Thesis
Submitted to Institute of Multimedia Engineering
College of Computer Science
National Chiao Tung University
in partial Fulfillment of the Requirements
for the Degree of
Master
in

Computer Science

July 2008

Hsinchu, Taiwan, Republic of China

中華民國九十七年七月

多層次臉部細紋之分析與合成

研究生：林家如　　指導教授：　林奕成 博士

國立交通大學

多媒體工程研究所

摘　要

本論文提出一個可應用於產生人臉主要運動表情並保有細緻人臉細紋之影像多層次頻率分析與合成技術。我們使用對人臉的不同區域彼此間之相關係數作為人臉分群之準則，並利用分群出的人臉區域獨立進行表情的合成，藉以利用少量資料產生豐富變化之人臉表情。我們使用高頻資訊增強之控向金字塔分層法(steerable pyramid)對範例影像進行不同頻率成分之分解。在人臉高頻影像合成的過程，我們使用形態融合(blend shape)處理最低頻之資訊，並保留合適的高頻資訊，以作為後續標準差比對合成。藉由高低頻成分的獨立處理，我們的技術可以合成栩栩如生之人臉影像。將我們的系統與時間空間之相關性以及紋理對齊之技術結合，可呈現出具細緻紋理、豐富表情變化之人臉動畫。

關鍵字：形態融合，時間及空間連續性，normalized cuts，控向金字塔。

Analysis and Synthesis of Multi-layered Facial Details

Student: Jia-Ru Lin Advisor: Dr. I-Chen Lin

Institute of Multimedia Engineering

National Chiao Tung University

ABSTRACT

This thesis presents a multi-layer analysis and synthesis approach to synthesize novel expressions with fine facial details and global features. We analyze correlation between different face regions as the criterion for face segmentation, and process each sub-region independently for generating various appearances. In synthesis procedure, we use the high-band enhancement steerable pyramid for decomposing various frequency components. The lowest sub-band is used for primary deformation by blend shape. For high-band, the statistic-based feature matching and high-band enhancement synthesis is applied. Due to our high-band\low-band separate procedures, we can synthesize photorealistic facial expressions. Integrate our approach with spatial-temporal constraints and texture alignment, we can generate detail-preserved facial animation.

Keyword: blend shape, space-time coherence, normalized cuts, steerable pyramid.

Acknowledgements

I would like to greatly express my thanks to my advisor, Dr. I-Chen Lin. His patient and enthusiastic advice is the best support of the completion of this thesis. Most importantly, he devoted much time to guide me how to do academic research, analyze problems, look for possible solution, and accomplish research in the past two years. Also, I wish to appreciate all members of Computer Animation and Interactive Graphics Lab for their help and suggestion. With their help, the experiment of motion capture can be accomplished successfully. Eventually, I acknowledge my family for their support and encouragement.

Contents

List of Figures

Chapter 1. Introduction

1.1 Motivation

The technique of computer animation and virtual reality is widely used for 3D game and movie industry in decades. Also, 3D characters are popularly used in computer animation. However, producing realistic 3D character with facial details is still a labor-intensive work for animators. That is due to we are very familiar with observing facial appearance. The facial expression can communicate various kinds of feelings, and any slightly expression changes may express completely different meanings. Therefore, those details are difficult for manual editing.

The motion capture (Mocap) technique is common used method to acquiring motion. This technique can also use for generating facial animation. In order to capture facial animation, dozens of markers are placed on control points of subject's face. Tracking the variation of markers can acquire the facial motion, but they can't capture the subtle portions caused by skin deformation, such as wrinkles, creases, or pores.

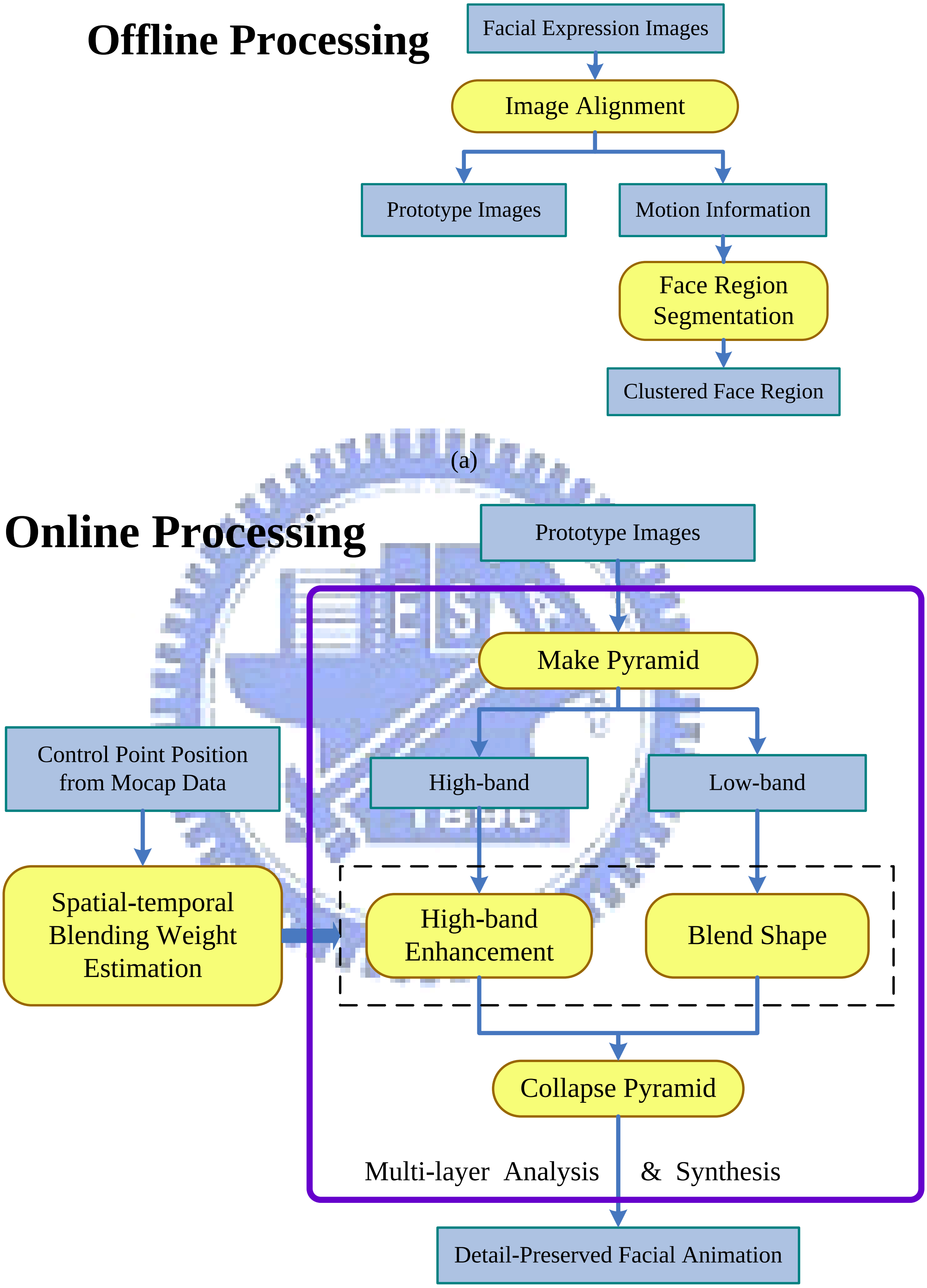
There are also many techniques to acquire face expressions, such as high resolution 3D laser scanners and face-scanning dome equipment. Those approaches provide convincing results, but it is inefficient to acquire all appearance that we need and their devices are highly expensive. On the other hand, many data-driven approaches are proposed to generate novel facial expressions from a set of example appearances, such as blend shape. The concept of blend shape is to represent each example expressions in convex vector space. The synthetic expression can be generated by using convex combination of those example expressions. Using blend shape can synthesize various facial expressions. However, the facial detail information, such as wrinkle and pores, may lose due to the process of image blending. For this reason, we propose multi-layer facial detail analysis and synthesis approach for synthesizing detail-preserved expressions.

1.2 Framework

The goal of our research is to synthesize facial expressions with fine facial details. The proposed framework can be divided into two parts: offline processing and online multi-layer expression synthesis. Figure 1 demonstrates the framework of our system.

The offline procedure aligns those acquired expressions with the neutral face for producing prototype images. Furthermore, we obtain per pixel motion information during image alignment. We partition face as 64×64 grids and use the received motion information for analyzing the correlation between each face grid pair. By using normalized cuts with the correlation as criterion, we cluster face as 12 sub-regions. This work only needs to be done once.

The online part is about combining prototype images with multi-layer approaches for synthesizing detail-preserved facial expressions. By integrating Mocap data with spatial-temporal coherence for evaluating blending weight, we can synthesize detail-preserved facial animation.



(a)

(b)

Figure 1: The framework of our system. (a) Offline processing. (b) Online processing.

1.3 Organization

This thesis is organized as follows. Chapter 2 introduces related work about the approach to acquire facial expressions and synthesize novel appearance. Chapter 3 explains the pre-processing procedure and the concept of normalized cuts for our face region segmentation. Chapter 4 proposed the framework of multi-layer facial detail analysis and synthesis. Chapter 5 illustrates the experiment result from our multi-layer framework. In the last chapter, we present the conclusion and the future work.

Chapter 2. Related work

In this chapter, we introduce existing approaches of synthesis novel facial expressions, which include the concept of blending shape [1-3][5-6], a multilinear model for face transfer [7], and a statistical model for analysis and synthesis of 3D facial details [8]. Then, we discuss the methods and equipment for acquiring facial expressions, such as face-scanning dome [11], a structured light system [12], and multi-scale facial expression capture equipment [13]. Finally, we introduce normalized graph cuts (NCuts) [16][17] which is employed in our face region segmentation in section 2.3.

2.1 Novel Facial Expression Synthesis

The concept of blending shape is extensively used for generating novel facial expressions from a set of example expressions. Pighin et al. (1998) [1] used convex combinations of the geometries and textures of example expression models to construct photorealistic facial expressions. Their system included expression mapping and expression editing. For expression mapping, their system first mapped one’s expression to another person by setting the same convex combination coefficients. Then, their system provided a user-friendly interface to edit new facial expressions interactively. However, their expression editing system required a user to specify the convex combination coefficients manually for generating new facial expressions.

The traditional expression mapping techniques have the shortcoming of lack of expression details caused by skin deformation. Zhang et al. (2006) [2] developed a geometry-driven facial expression synthesis system to improve the shortcoming of expression mapping. After assigning the feature point positions of the facial expression, their system can automatically synthesize the corresponding expression image with facial details. Besides, they modified Pighin's approach by subdividing a face into smaller sub-regions. Through blending the sub-regions of example expressions respectively and combining those sub-regions together seamlessly, their approach can receive various kind of synthesis expressions.

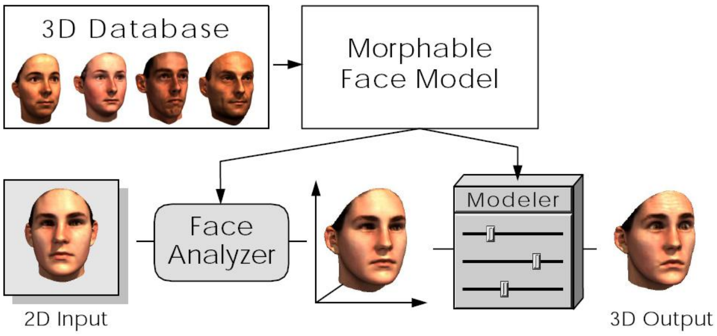


Figure 2: The system of modeling 3D texture face.

Blanz and Vetter (1999) [3] introduced a technique for modeling textured 3D faces, Figure 2 explains their system. They constructed a data set of 3D scanned face models, and transformed the geometry and texture of the prototypes to vector space. They used linear combination of the prototypes to synthesize new facial expression and used principal component analysis (PCA) [4] to analyze the face data set. In the process of synthesizing new facial expressions, they combined the average of face shape data set with weighted shape eigenvectors of covariance matrices to synthesize the new facial geometry. By using the same process, they can synthesize new face texture in the same way.

In 2003 [5], Blanz et al. further improved their method for photo-realism face animation of a signal image or video. They represented facial expressions in vector space and estimated by computing the difference between neutral face and expression face of the same person. After adding the facial expression vector to a neutral 3D face model, their system can transfer expressions across individuals. Figure 3 shows the concept of modified Leonrado's Mona Lisa with smile expression.

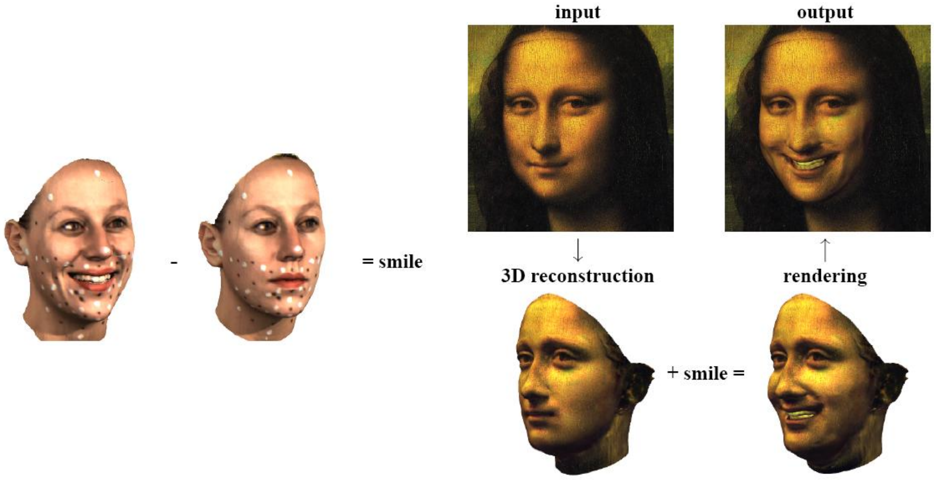


Figure 3: Extract the facial expression and add the expression to a neutral 3D face.

Ezzat et al. (2002) [6] created a speech animation module. They used a single video camera to record a human subject as he/she utters a predetermined speech corpus. After processing the corpus automatically and providing the novel speech as input, their module can synthesize a brand new video with the human subject's mouth uttering novel utterances. Their main concept was using the multidimensional morphable model (MMM) to synthesize previously unseen mouth configurations from a set of mouth image prototypes and using a trajectory synthesis technique for mapping an input phone stream to a trajectory of parameters in MMM space. Figure 4 shows the procedure of the speech animation system.

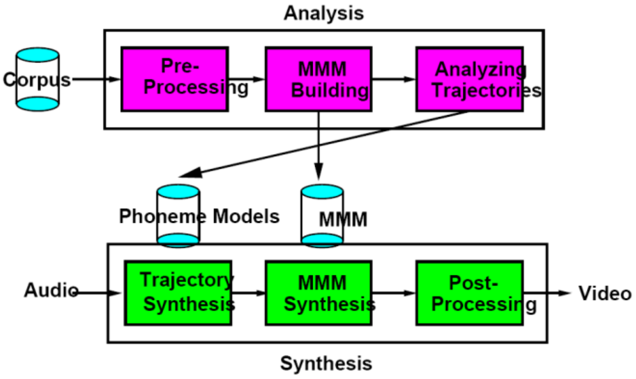


Figure 4: The system of video realistic speech animation module.

The blending shape technique is intuitive and easy to use for synthesis facial expression. For this reason, this technique is extensively employed in computer animation. However, this technique has the shortcoming of losing facial details in the image blending process.

Using motion capture system (Mocap) to capture facial expressions is also widely used in animation, but it requires covering densely markers over character's face. Vlasic et al. (2005) [7] proposed an approach to transfer facial motion to another character and only needed simply monocular video equipment. Their concept was based on multilinear model of 3D face meshes, which can parameterize the space of geometry variations as different attributes (e.g., identity, expression and viseme). Thus, they used those parameters to drive detailed 3D textured face mesh for a target character (Figure 5). However, a large number of normalized face scan data are required for evaluation of multilinear model.

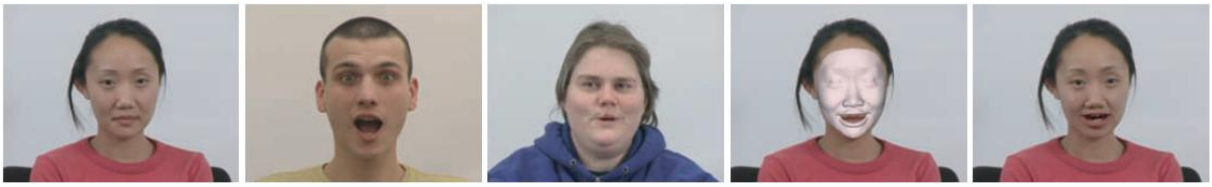


Figure 5: Face transfer with multilinear models can control animator's facial attributes, such as identity, expression, and viseme. This result shows they extracted the identity from first video, expression from the second, and viseme for the third one, then combined those

attributes back to the original video.

Golovinskiy et al. (2006) [8] introduced a statistical model for analytic and synthetic small 3D facial details such as pores and wrinkles. They used the acquisition system [Weyrich et al. 2005] to acquire the high resolution face geometry across different genders, ages and races. Then, they separated a high resolution face into smooth base mesh and a detailed displacement image, and they extracted the statistic of displacement image by steerable pyramid [9][10]. The steerable pyramid can decompose the input image into different orientation components. To extraction the statistic of the displacement image, they partitioned the image into tiles and analyzed the standard deviation of different orientation components from steerable pyramid. By matching the displacement image with the desired statistic properties and combining it with the base mesh, new facial expressions with details can be generated. Since the extraction of statistics process decomposed displacement image into titles, this approach can’t deal with the coarse wrinkle cross over those titles. An overview of the statistical model shows in Figure 6.

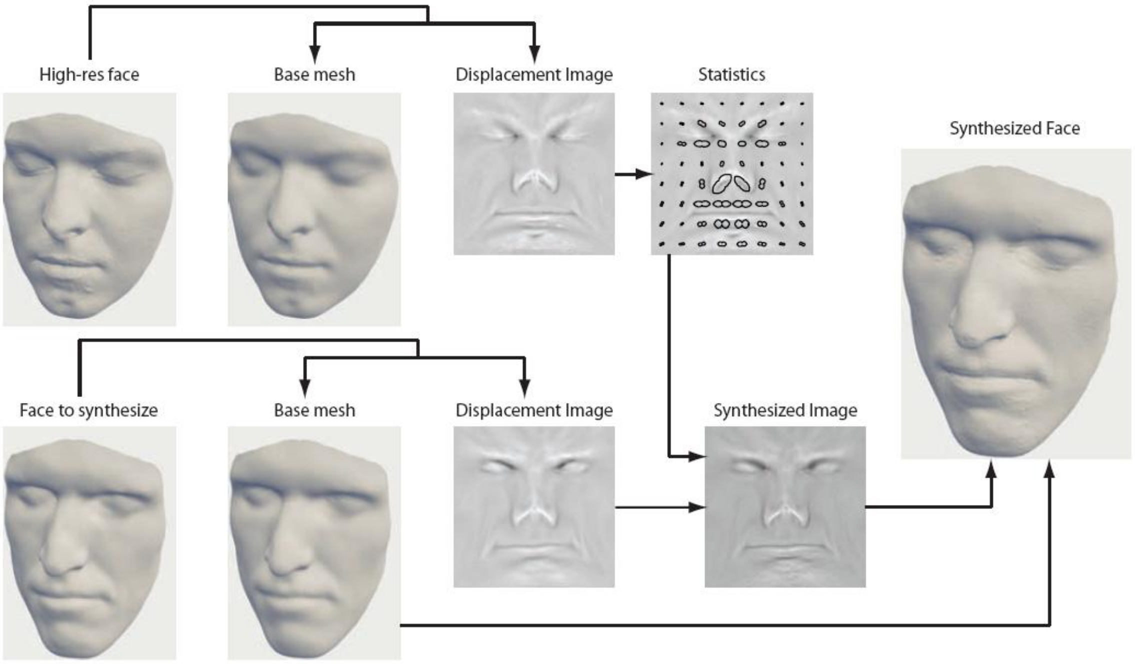
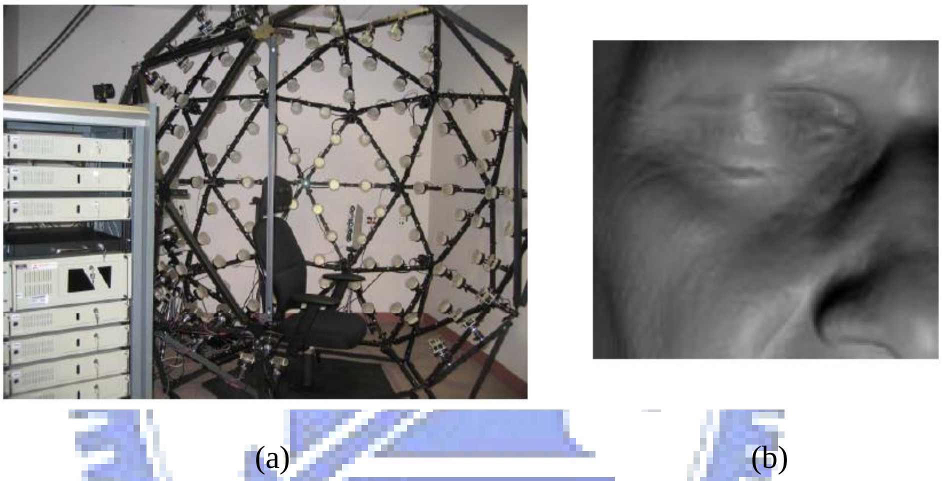


Figure 6: The statistical model of analysis and synthesis facial details.

2.2 Acquisition of Facial Expressions

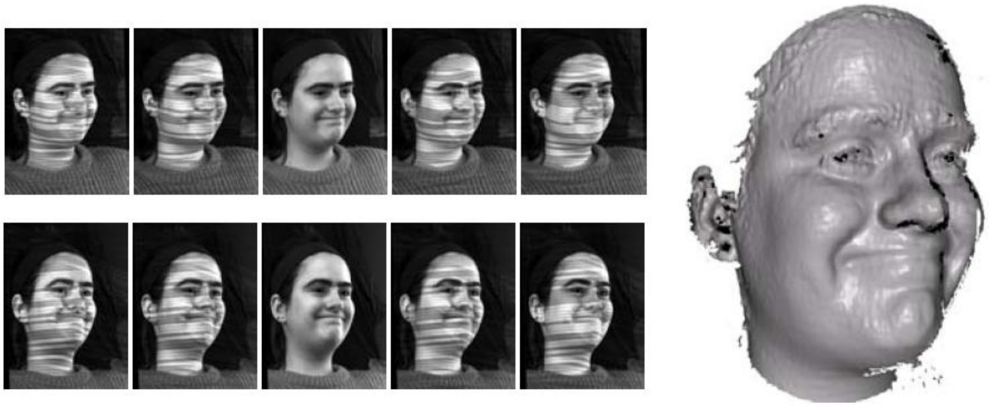
Facial detailed geometry influenced the realism of 3D face models greatly. In Weyrich's [11] research, they used face-scanning dome to measure the high-resolution face model and skin reflectance. The equipment consisted of 16 digital cameras, 150 LED light sources, and a commercial 3D face-scanning system, shows in Figure 7. Their measurement system can acquire high quality facial details, however, this method required high cost.



(a) (b)

Figure 7: (a) The equipment of face-scanning dome. (b) The high quality geometry.

Zhang et al. (2004) [12] presented a system that construct high resolution and dynamic face models from video sequences. They used the globally consistent spacetime stereo to avoid the problem of pixel discontinuity on the reconstructed face surfaces. Their system included the process of template fitting and template tracking. In template fitting process, they fitted the depth map estimated from stereo video to the template mesh. For template tracking, they used the vertex motion measured from optical flow to automatically fit the template mesh to the frame sequences. Zhang and colleagues used spacetime stereo to reconstruct objects that can move and deform over time. However, due to self-occlusion effects and lower capture rate, acquiring dynamic fine facial geometry was still difficult. Figure 8 shows the reconstructed face by spacetime stereo.



(a) (b)

Figure 8: (a) The structured light image from a pair of stereo videos. (b) The reconstruct face.

The technique to reconstruct face model from spacetime stereo encounters the inherent low-capture-rate and self-occlusion problem of structure light system. To rectify this problem, Bickel et al. (2007) [13] proposed a multi-scale representation and acquisition technique for animating high resolution facial geometry and wrinkles. They classified the 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 facial expressions. In their system, they used the face-scanning dome [Weyrich et al. 2006] to acquire high resolution facial details and MoCap to receive the coarse scale facial motion. Next, they transferred the Mocap motion to animate the high resolution 3D scan data for large-scale animation. They used additional two high resolution Basler cameras to capture the medium-scale expression wrinkles. Besides, they analyzed the position and shape of wrinkles from videos with the uniform B-Spline curve and a valley-shape wrinkle model [14]. After the shape estimation of wrinkles, they added this information to large-scale animation to produce medium-scale animation. Their approach can capture and animate different scale facial expressions. However, due to the predefinition of valley-shape model, the shape of synthesized wrinkles was limited. Figure 9 displays the synthetic results by using their multi-scale facial geometry model.

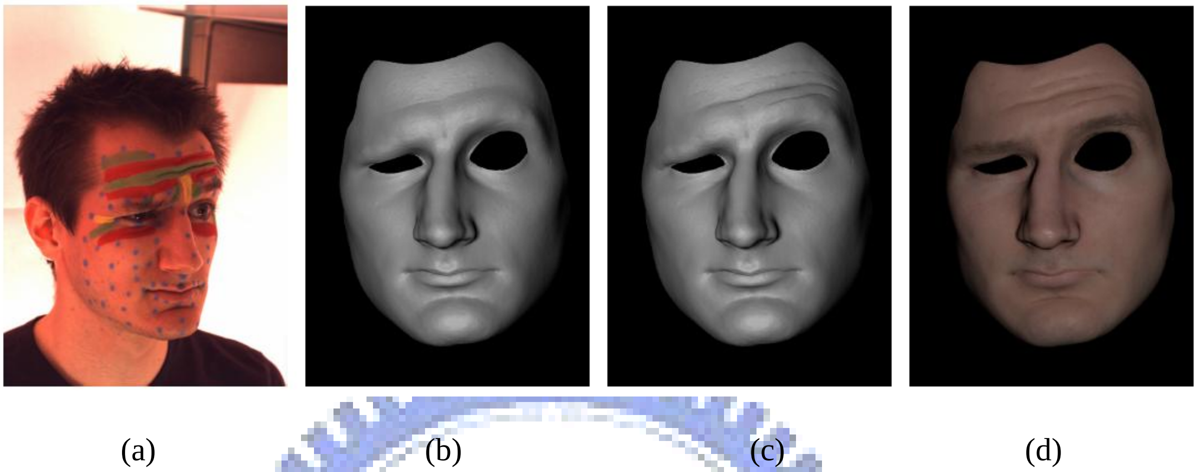


Figure 9: Animation high resolution face scanned by Mocap and video-driven wrinkle model. (a) Video frame. (b) Large-scale animation. (c) Medium-scale animation with wrinkles. (d) Skin-rendering.

In 2008 [15], Bickel et al. further modified their approach and presented a method for real-time animation of facial detail expressions. Their hybrid animation considered facial geometry as large-scale motion and fine-scale motion, Figure 10 shows their hybrid face animation pipeline. They computed the large-scale motion by using the same linear shell deformation [13], and incorporated a pose-space deformation technique for learning fine-scale facial details from a set of example poses. The pose-space deformation can learn the corresponding skin strain of wrinkle formation sparsely. After combining large-scale facial motion with fine-scale details, their approach can perform real-time animation of facial detail faces.

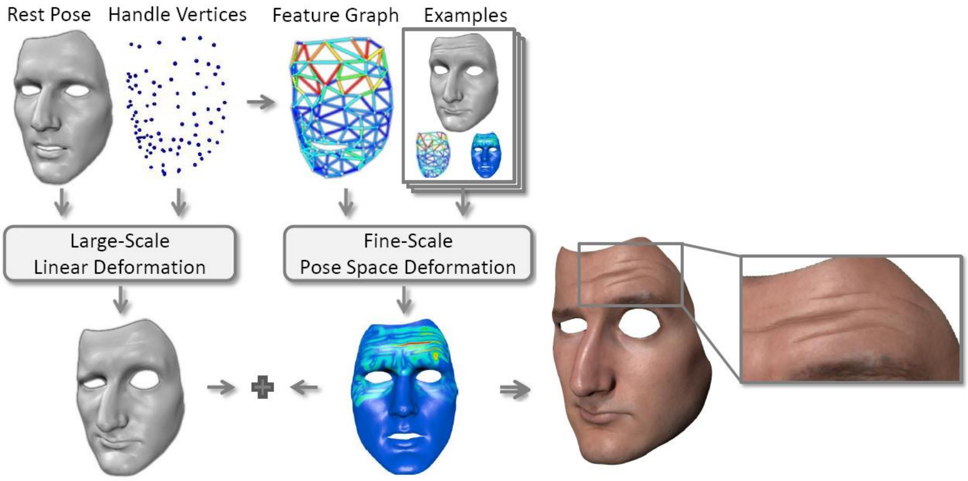
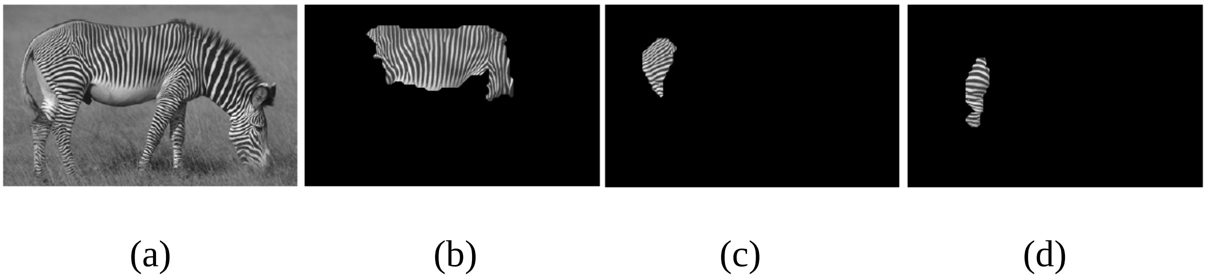
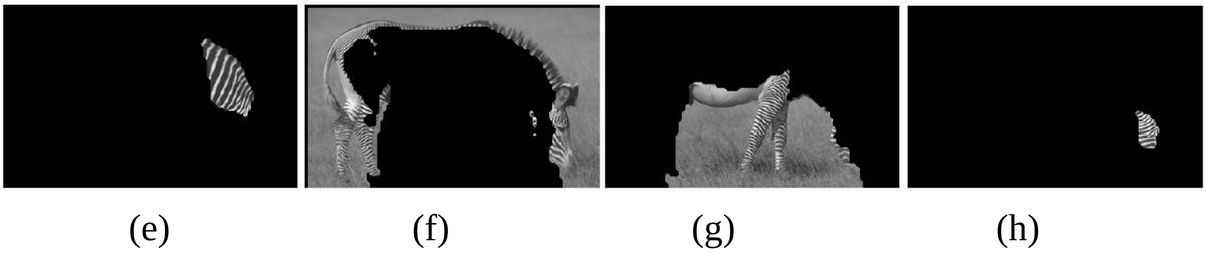


Figure 10: The pipeline of hybrid face animation.

2.3 Face Region Segmentation

Before synthesizing new facial expressions, we take the motion of each pixel into consideration for clustering face into different sub-region. We use normalized graph cuts (NCuts) as our clustering algorithm which was proposed by Shi and Malik [16][17]. Their clustering algorithm treated image segmentation as a graph partitioning problem and used the normalized cut value as a global criterion for segmenting the graph. The normalized cut criterion can measure both the dissimilarity between the different groups and the similarity within the groups. Besides, they used a generalized eigenvalue evaluation to optimal the criterion for producing encouraging images segmentation results. Figure 11 shows the results of normalized cuts approach on texture segmentation for a natural image of a zebra against a background.





(e) (f) (g) (h)

Figure 11: (a) An image of zebra. (b-h) the result of normalized cuts.

Chapter 3. Face Region Clustering

For producing more various facial expressions, we need to partition face into different regions for the consequential synthetic procedure. Therefore, we decompose face into 64×64 grids, evaluate the motion of each grid, and use those information for analyzing the correlation between each pair of grids. Finally, we employ the correlation between grids as the criterion in clustering process, and use normalized graph cuts (NCuts) algorithm to segment face into different sub-regions.

3.1 Pre-Processing

In order to synthesize various kinds of facial expressions, we record 22 images with different facial expressions. Without lose of generality, we select the first image or preparative expression as neutral face. To avoid blurred synthetic result during image blending process, we first translate head position to remove the movement of head pose and then use image warping [18] to align all images with the neutral face. In the image warping process, we assume the distance between each source-destination pixel pairs as the displacement of each pixel between destination expression face and neutral face.

After evaluating motion information of all prototype images, we process all prototype images by histogram equalization [19] which gather image’s histogram and adjust image’s contrast. Using histogram equalization to process all prototype images can eliminate the different color of face skin and still maintain facial texture. Therefore, we can construct seamless synthetic result from sub-region blending shape. After those pre-processing steps, we employ those 22 images as prototype images in the following analysis and synthesis process and partition each prototype image into 64×64 tiles for the consequential clustering process.

3.2 Correlation Analysis

For clustering a face into different sub-regions, we decompose the face region of image into 64×64 grids. Before the clustering process, we need to analyze the correlation between each pair of grids. Afterward, we employ the evaluated correlation as criterion to cluster those 4096 grids into different groups.

In statistics, correlation is an objectively standard to estimate the relationship between two data sets. Using correlation can evaluate the strength and direction of a linear relationship between two random variables. In order to analyze the correlation between each pair of grids, we also partition each prototype image into 64×64 tiles. Then we utilize per pixel displacement gather from image warping and average the displacement of those pixels within tile as the tile displacement. We employ the average displacement to represent the motion of each prototype tile.

Having the motion data of each prototype tiles, we consider those tiles at the same position among all prototype images as the temporal variation (Figure 12). Therefore, we can regard the motion of tiles at the same position among different prototype image as the movement of the corresponding grid at a different time step.

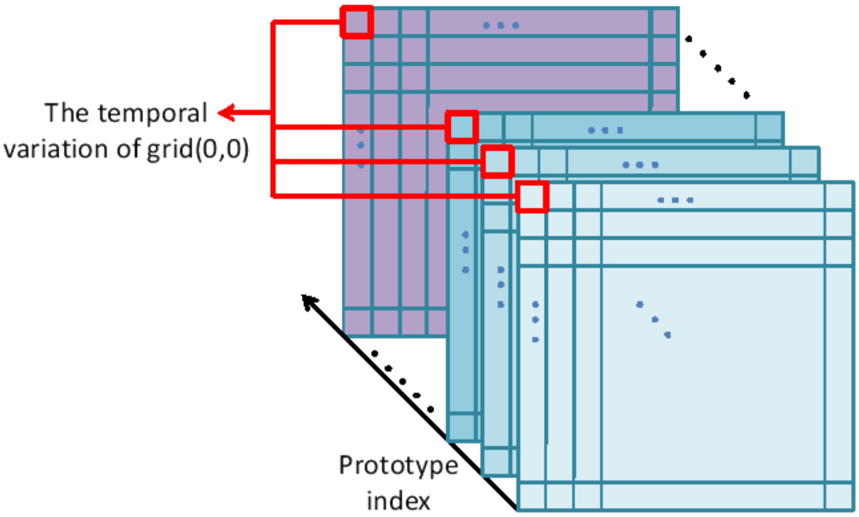


Figure 12: The tiles lie in the same position among different prototype images can represent the temporal variation of the corresponding grid.

We employ the motion of grid at different time as information of each gird for analyzing the correlation between each pair of grids. The correlation between gird *a* and grid *b* can define as:

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

a_t and b_t are the motion of grid *a* and grid *b* at time *t*, that is, those grids belong to prototype *t*. $\bar{a}$ and $\bar{b}$ are respectively the average motion of grid *a* and grid *b* among all prototype images. S_a and S_b are the standard deviation of grid *a* motion and grid *b* motion. The normalized procedure for computing correlation can estimate the correlation of two variable sets with different measurement units. The two data sets have positive linear relationship if the correlation coefficient approaches to 1, and have negative linear relationship if the correlation coefficient approaches to -1. The absolute value of correlation is closer to either 1 or -1, the stronger relationship between two variables.

Using the correlation formula (Equation 1), we evaluate the correlation between each pair of grids, and use the correlation as criterion to cluster face segment.

3.3 Face Region Segmentation by Normalized Cuts

After employing each grid motion to analyze the correlation between each pair of grids, we use normalized cuts [16] [17] for partitioning face into different sub-regions.

For graph partitioning, Shi and Malik used *normalized cut* (Equation 2) as the partition criterion to measure the disassociation between two groups:

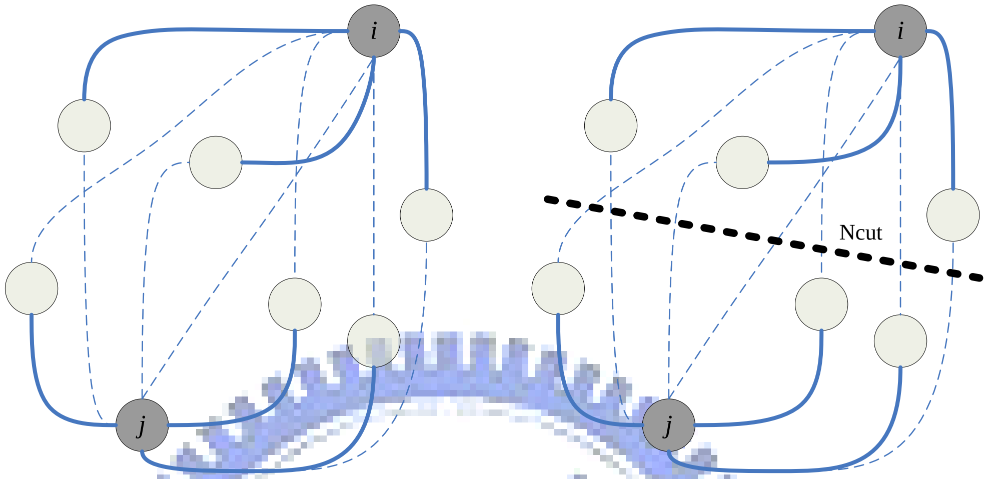
$$Ncut(A,B)=\frac{cut(A,B)}{asso(A,V)}+\frac{cut(A,B)}{asso(B,V)} \tag{2}$$

, where the *A* and *B* are two disjoint sets, *cut*(*A*,*B*) is the total weight of edges that have been removed in the partition 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. In partition process, they optimized the *normalized cut* to minimize the disassociation between the groups.

Our purpose is to segment the face grids into different groups. Therefore, we use the face grids to set up a weighted graph *G* = (*V*, *E*), and define the weight on each edge as the correlation between grid pairs (Equation 3). Consequently, the edge weight w_{ij} between grid *i* and grid *j* can define as:

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

correlation (i, j) is the correlation between grid i and grid j, the correlation coefficient can acquire by equation 1. Figure 13 shows an example of normalized cuts.



(a) A graph G

(b) A NCuts on G

Figure 13: An example of graph G, we compute the correlation between each grid pairs. (a) In consideration of the correlation between grid i/grid j and all other grids. If the correlation between grid pair is more than an absolute threshold, the edge is assigned to solid line, otherwise the edge is assigned to dashed line. (b) The normalized cuts can partition high correlation grids into a group and separate the low correlation grids.

After evaluating the weight value, we can set $\mathbf{W}$ as a $N \times N$ symmetrical matrix with W(i,j) = w_{ij} , N is the number of total grids. And let $\mathbf{D}$ be a $N \times N$ diagonal matrix with $\boldsymbol{d}$ on its diagonal, where $d(i) = \sum_j w(i, j)$. Then we minimize the *normalized cut* by solving the generalized eigenvalue system (Equation 4),

$$(\mathbf{D} - \mathbf{W})y = \lambda \mathbf{D}y \qquad \textbf{(4)}$$

Solve equation 4 for eigenvectors with the smallest eigenvalues, and use all of the top eigenvectors can obtain a K-way partition. That is, the n top eigenvectors can be used as n

dimensional indicator vectors for each grid. The number of groups segmented is controlled directly by the number of *n*. Selecting *n* top eigenvectors can identify up to 2^n groups.

In order to determine the number of clusters in our face region clustering procedure, we define a measurement function (Equation 5) to keep the balance between cluster number and the correlation within each group:

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

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 sub-regions. Figure 14 shows the face region clustering result, and we synthesize novel facial expression within each cluster respectively.

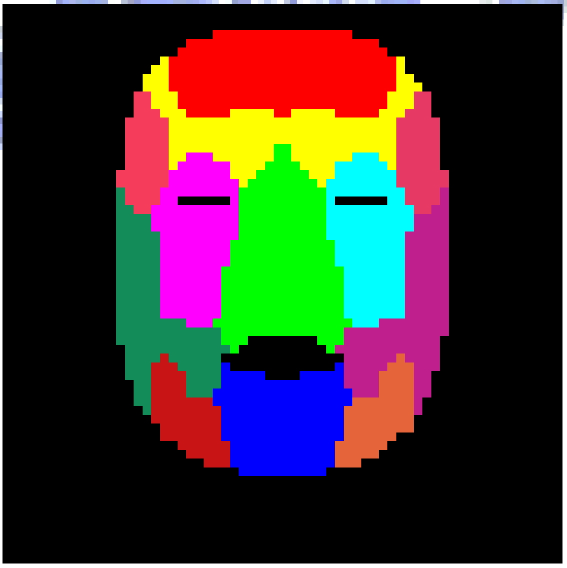


Figure 14: The face clustering result. The grid marked with the same color indicates those grid belong to the same cluster.

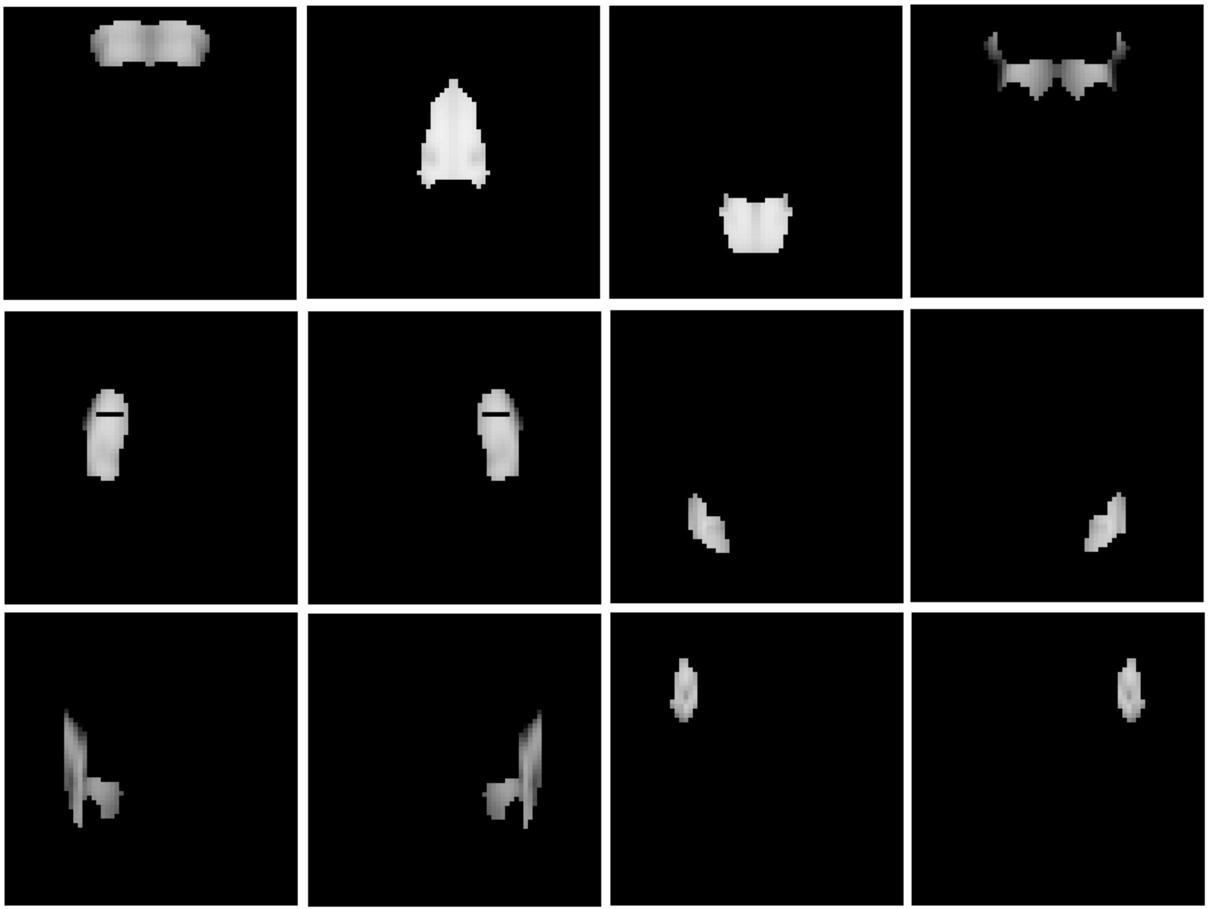


Figure 15: Those images demonstrate the average correlation of each grid within group. The intensity of each grid represents the value of average correlation. Those images correspond the correlation 0.8~1.0 with intensity 0~255.

Chapter 4. Multi-layer Facial Detail Synthesis

Even though we can obtain various facial expressions, it seems impossible to acquire all feasible expressions. Therefore, we propose a facial detail-preserved technique to synthesize novel facial expressions from prototype images. We integrate the modified steerable pyramid with statistics-based matching and high-band enhancement for synthesizing facial details preserving expressions. With our multi-layer analysis and synthesis procedure, we can synthesize detail-preserved appearance and avoid the blurred result by traditional blending shape.

4.1 Produce Detail-Preserved Expressions

Existing data-driven approaches can generate new facial expressions from blending a dataset of example expressions. However, facial details of the newly generated facial expressions, such as pores and wrinkles, will blur during image blending process. Therefore, we propose improving the conventional blend shape process by using a *multi-layer analysis and synthesis* approach to preserve the facial details in the image blending procedure.

Our main concept is first separating all example expressions into various sub-band images, such as high-pass images, low-pass images, and various orientation sub-bands. Next, we compose high sub-band images and low sub-band images respectively for synthesizing novel expressions. Finally, we reconstruct novel facial expressions by combining those components in each sub-band for detail-preservation. For decomposing all example images into different sub-bands and recombining new facial expressions, we employ the framework of steerable pyramid [9][10] .

Steerable pyramid is one kind of image pyramids. An image pyramid can transform an image into different sub-bands by convolution and sub-sampling. For the successive level of each sub-band, the sub-sampling factor is increased by a factor of two in each dimension. Decomposing an image by using the hierarchy of image pyramid yields a set of sub-band images of different sizes that correspond to different frequency bands. The original image can simply be recovered by inverting the sequence of operations which be used in the image decomposition process.

The steerable pyramid is similar to Gaussian or Laplacian pyramids, it can decompose the image into several spatial frequency bands. Moreover, it can further divide each frequency band into a set of orientation bands. The system diagram of steerable pyramid shows in Figure 16.

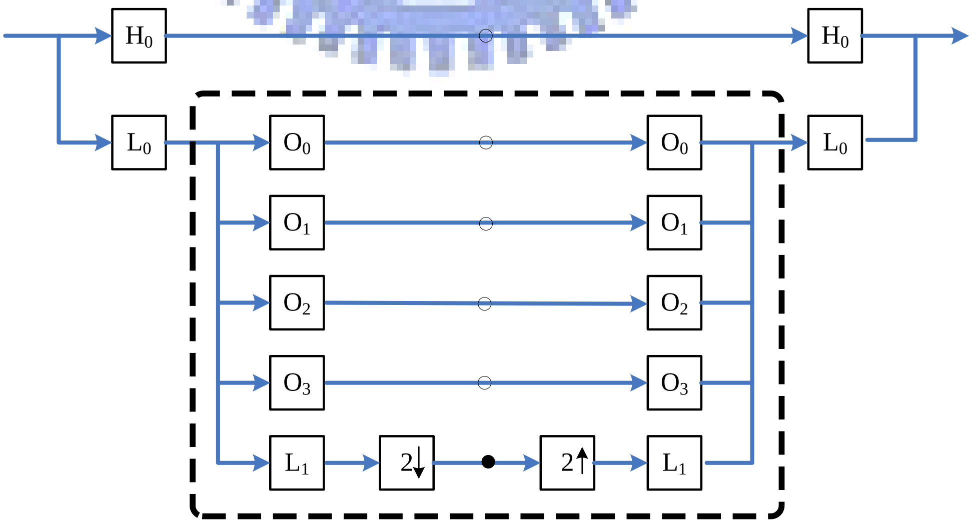


Figure 16: System diagram of the steerable pyramid. Each square box represents the

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_i are band-pass filters in different orientation. The hollow circles represent the decomposed sub-band images. The pyramid can construct recursively by repeat the process enclosed by the dashed rectangle at the location of solid circle.

Figure 16 demonstrates the structure of decomposition and reconstruction steerable pyramid. The left-hand side of the diagram is the analysis part, decomposing an image into a series of different scale and orientation sub-bands. The procedure is also called *make-pyramid*. On the contrary, the right-hand side of the diagram is the synthesis part, applying the invert operations of analysis part to the decomposed sub-bands and combines those sub-bands to reconstruct the original image. The procedure is called *collapse-pyramid*.

The steerable pyramid transform begins with an input image filtered by a high-pass/low-pass splitting filter. The split low-band can further decompose into different oriented band-pass sub-bands by steerable filters, low-pass sub-band by low-pass filter and down-sampling. The steerable filter is a filter of arbitrary orientation which can be synthesized by linear combination of a set of "basis filters". The next level of the pyramid is constructed from the low-pass sub-band filtered by a set of steerable filters and low-pass filters, and this process can repeat recursively in the make-pyramid procedure. The characteristic of steerable pyramid is self-inverting, that is, filters of synthesis part are the same as filters of analysis part. Therefore, we can reconstruct the original image by inverting the operations of analysis part.

4.2 Multi-layer Expression Synthesis

4.2.1 High-band Enhancement Steerable Filter

Since the analysis part of steerable pyramid can decompose images into various scale and orientation sub-bands and the original image can be recovered back through the synthesis part, we further modify the framework of steerable pyramid and integrate the modified framework with the procedure of our multi-layer facial detail synthesis. Figure 17 shows the modified framework of steerable pyramid. For preserving more high frequency information, we further decompose the high-band of each prototype images into one level various orientation sub-bands. Besides, we adopt the first derivative of the 2-dimensional, circularly symmetric Gaussian function rotate 0°, 90°, 30°, and 120° about horizontal as steerable filters for extracting the orientation sub-bands.

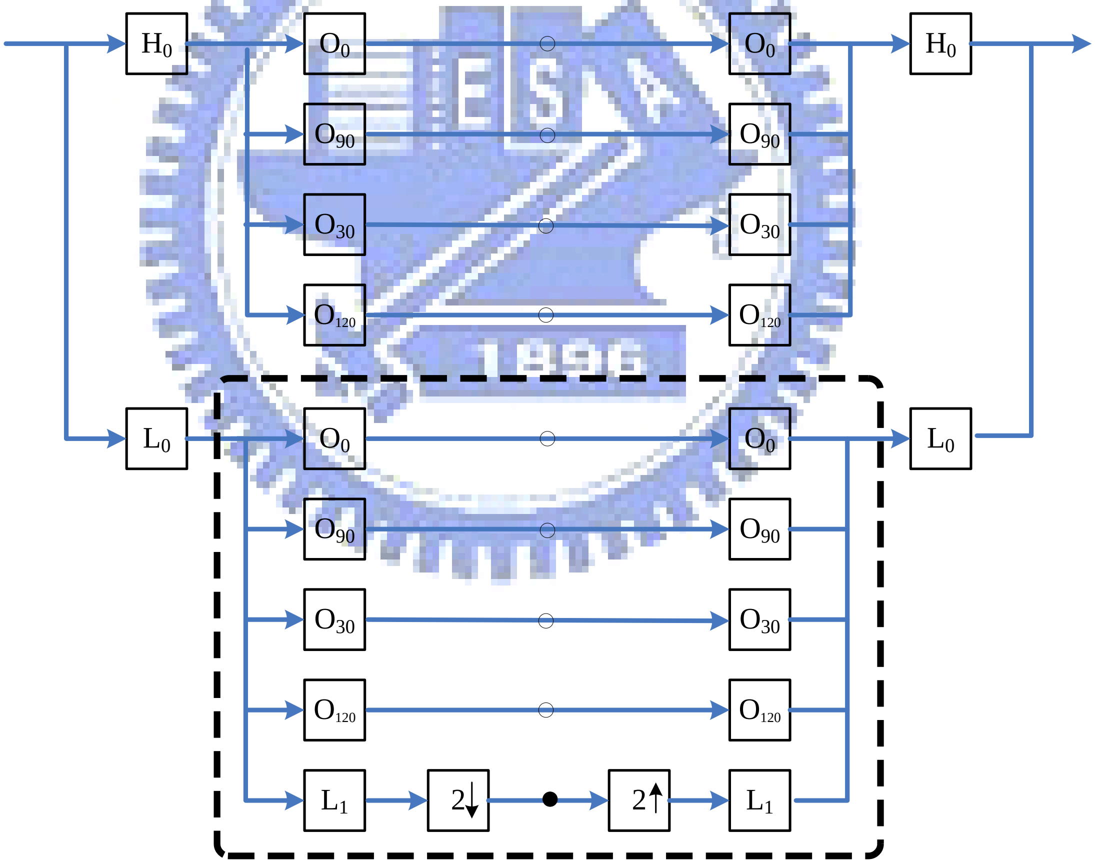


Figure 17: The modified framework of steerable pyramid for our multi-layer facial detail synthesis procedure. Those symbols in the modified system diagram have the same meaning as Figure 16. The structure is similar to the original steerable pyramid, and we further decompose the high-band into various orientation sub-bands.

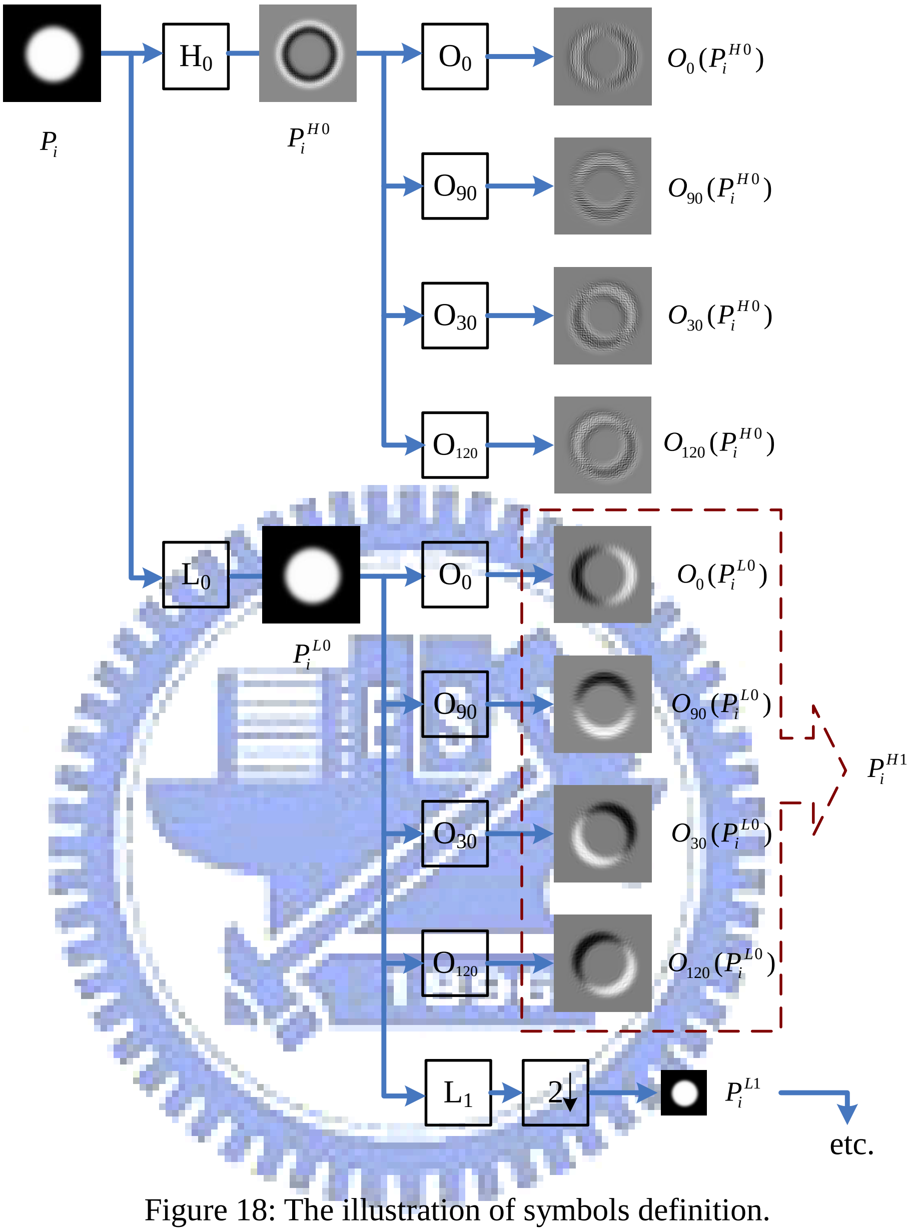


Figure 18: The illustration of symbols definition.

We use Figure 18 that shows the procedure of decomposing an image P_i and the correspondent symbol of each sub-band to demonstrate our symbol definition. In our multi-layer analysis and synthesis procedure, we use a set of prototype images $\{P_i\}_{i=1}^{m}$, 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 $O_0(\ldots)$ represent the operator of horizontal steerable filter. $O_{90}(\ldots)$, $O_{30}(\ldots)$, and $O_{120}(\ldots)$ 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 regard 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 the reconstructed image with the ground truth, we found that making more than three-level pyramids can let the reconstructed image with satisfactory intensity ranges and details as well. Therefore, we apply the modified framework of steerable pyramid to decompose all prototype images into 4 orientations and 3 level frequency bands, and employ those sub-bands for synthesizing novel facial expressions. In order to synthesize detail-preserved expressions and reduce blurred result, we process the lowest sub-band P_i^{L3} and other high-band of each prototype images independently.

4.2.2 Blend Shape

We assume that facial geometry has high relation with facial appearance. That is, the position of control points between similar facial expressions is 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 weight. Therefore, we consider each expression E_i can represent as $E_i = (G_i, P_i)$, where G_i is geometry, and P_i is prototype image. Let $H (E_0 ,E_1 ,..., E_m)$ be the space of all possible convex combination among all example expressions, i.e.,

$$H(E_0,E_1,\ldots,E_m)=\left\{\left(\sum_{i=0}^{m} w_iG_i,\sum_{i=0}^{m} w_iP_i\right)\mid \sum_{i=0}^{m} w_i=1,\text{ and } w_0,w_1,\ldots,w_m\geq 0\right\} \qquad (6)$$

Therefore, novel expression can represent as follows:

$$E^{new}=\left(G^{new},P^{new}\right)\text{ , where } G^{new}=\sum_{i=0}^{m} w_iG_i,\ P^{new}=\sum_{i=0}^{m} w_iP_i \qquad (7)$$

Since each face sub-region have different sub-expressions, we use normalized cuts to segment face into different sub-regions for generating more various facial expressions from small data pool. For this reason, we modified equation 6 to synthesize each sub-region *R*:

$$H^R(E_0^R,E_1^R,\ldots,E_m^R)=\left\{\left(\sum_{i=0}^{m} w_i^RG_i^R,\sum_{i=0}^{m} w_i^RP_i^R\right)\mid \sum_{i=0}^{m} w_i^R=1,\text{ and } w_0^R,w_1^R,\ldots,w_m^R\geq 0\right\} \qquad (8)$$

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

$$E^R=\left(G^R,P^R\right)\text{ , where } G^R=\sum_{i=0}^{m} w_i^RG_i^R,\ P^R=\sum_{i=0}^{m} w_i^RP_i^R \qquad (9)$$

Applying blend shape concept to synthesize novel expression is effective, especially only having few prototype images. On the contrary, using a large number of prototype images for blend shape, some high frequency details, such as pores or wrinkle, may be lost during blending process. Therefore, we propose statistics-based example matching instead of blending high-band information.

4.2.3 Extraction of Statistics

Since high frequency detail information will lose during blending shape, we synthesize each high-band of novel expression by evaluating the fine detail information of each prototype. Our goal is to produce high-band of synthetic expression with satisfactory detail information. Therefore, we partition each prototype high-band image as 64 × 64 grids and evaluate detail information of each grid. Extracting histogram of each high-band grids and selecting the gird with proper histogram for synthesizing corresponding grid in high-band can ensure that the detail information don't lost in blending procedure. However, storing and analyzing those histograms is burdensome. Golovinskiy et al. find the main different between histogram within the same grid of different prototypes is their width [8]. Therefore, we adopt the concept and use standard deviation of each grid to substitute for analyzing histogram of each grid. Furthermore, this approximation implies significant compression of data.

For synthesizing high-band of novel expressions, we evaluate the statistic value from blending standard deviation of the same grid among prototype images, and select the grid with the closest standard deviation value as the corresponding gird for synthetic expression. $\sigma(P_i^f, k)$ represents the standard deviation of grid *k* belong sub-band *f* of prototype *i*. The standard deviation of corresponding synthetic grid *Sigma* can derive from:

$$Sigma = \sum_{i=0}^{m} w_i^R \sigma(P_i^f, k), \text{ where gird } k \in R \tag{10}$$

The grid *k* of synthetic image in sub-band *f* can be determined by:

$$\langle i^* \rangle = \arg\min_{i} \left(Sigma - \sigma(P_{i}^{f}, k) \right), \text{for } i = 0,1,\ldots,m. \tag{11}$$

By selecting a proper grid for each synthetic image high-band, we can maintain the high frequency information of synthetic result. However, when deal with less prototype images, using the closest standard deviation as criterion to select proper high-band grid may cause incorrect synthetic result. Therefore, we integrate statistics-matching based frame work with high-band enhancement synthesis.

4.2.4 High-band Enhancement Synthesis

In order to integrate the benefits of blend shape and statistic model, we propose high-band enhancement synthesis for maintaining details and reducing blur effects as well. We separate the lowest sub-band P_i^{L3} and other high-band of each prototype images independently. For synthesizing the lowest sub-band of novel expression, we blend the lowest sub-band of each prototype in the pyramid of each face region respectively. For reducing the blur effect and avoiding estimated error of statistic among small data set, we select the sub-region of prototype high-band with maximum blending weight as the corresponding sub-region of synthetic high-band. After blending the lowest sub-band and combining high-band with maximum blending weight, we can synthesize novel facial expression by collapse pyramid.

Owing to processing high-band and low-band independently, we can maintain more high frequency information, synthesize novel facial expressions with photorealistic facial details such as wrinkles and pores, and retain global features, such as the wrinkle cross over forehead. In our experience, if the blending weight of all prototype images is average, using statistics-matching based can produce better result. On the contrary, when certain prototype image has larger blending weight, high-band enhancement is suitable for synthesis.

4.3 Expression Analysis

The goal of analysis is to estimate the blending weight w_i^R for synthesizing the sub-region R of novel facial expressions. Since we consider that the facial appearance is highly related to facial geometry, we apply control points position to determine facial expression. Thus, we analyze the blending weight from control points of all prototype images, and use the blending weight to synthesize novel facial appearance.

Let G_{new}^R denote the sub-region R's control points position of novel expression. Given G_{new}^R , 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:

$$\text{Minimize } : target = \left(G_{new}^R - \sum_{i=0}^{m} w_i^R G_i^R \right)^T \left(G_{new}^R - \sum_{i=0}^{m} w_i^R G_i^R \right), \tag{12}$$

$$\text{Subject to}: \sum_{i=0}^{m} w_i^R = 1, w_i^R \geq 0 \;\text{ for } i = 0,1,\ldots,m.$$

After optimizing the blending weight of each sub-region, we can apply those coefficients for synthesizing novel facial appearance.

4.4 Spatial and Temporal Constrain

By analyzing the blending weight in the previous section, we can synthesize novel facial expression with predefined control point position. Our approach can also apply to synthesize animation with facial detail expressions. For acquiring control point position with temporal coherence, we record one subject’s facial motion by Mocap for determining the blending weight at each frame. Since the facial expression involve very fine motion, even slightly breathing can disturb the skin and cause noises. In order to reducing those noises, we use the temporal coherence for optimal blending weight in each face sub-region *R* by minimizing the *target* term and additional smooth term:

$$C^R(y) = target + \lambda y^T W W y^T \tag{13}$$

y_t is a m-dimensional parameter vector that control blending weight at frame *t*. *y* is a vertical concatenation of all individual y_t :

$$y = \begin{bmatrix} y_t \\ \vdots \\ y_T \end{bmatrix} \tag{14}$$

We employ λ to determine the trade-off between both terms, and *W* is a first order difference operator for smoothness term:

$$W = \begin{bmatrix} -I & I & & & & \\ & -I & I & & & \\ & & & \ddots & & \\ & & & & -I & I \end{bmatrix} \tag{15}$$

Chapter 5. Experiment and Results

In this chapter, we introduce our experiment and demonstrate our result. At the beginning, we show the step for acquiring and pre-processing prototype images and Mocap data. Then, we illustrate the synthesized result from our multi-layer analysis and synthesis framework. We compare our hybrid approach with blend shape for synthesizing ground truth.

5.1 Acquiring and Pre-Processing with Prototype images

In our experiment, we use high-definition video (HDV) to capture various facial expressions of actor and select those frames with representative expressions as our prototype images. As show in Figure 19, we put 41 markers on the actor's face and use Mocap to trace the motion of markers when actor makes a series of expressions. The position of markers is selected by marker movement can represent different facial expressions. We project the Mocap data to image coordinate for expression analysis procedure.

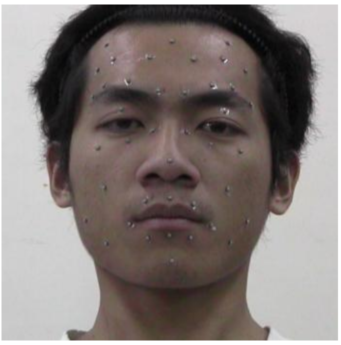


Figure 19: The neutral face with 41 markers for acquiring facial motion.

After determining prototype images, we use image warping to align all example expressions with natural face. Figure 20 shows the 22 prototype images after pre-processing.

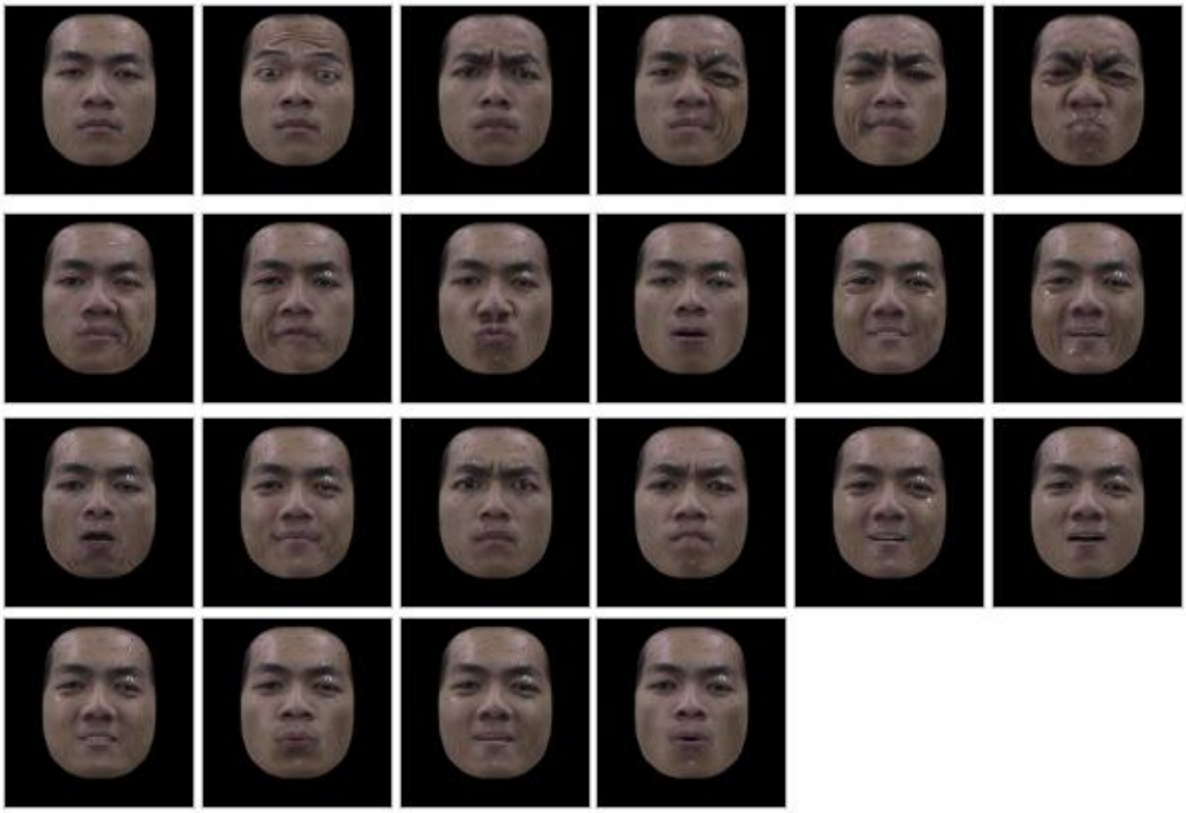


Figure 20: The 22 prototype images after normalization.

5.2 Face Region Clustering by Normalized Cuts

For synthesizing more various expressions from few example expressions, we use normalized cuts to cluster face into different sub-regions. In our normalized cuts process, we set k_1 = 0.0005, k_2 = 0.5 respectively. And we only use normalized cuts to cluster the gird which is part of face region. Clustering face region by normalized cuts may generate some isolated components, and we merge those isolated components into adjacent region. Consequently, we can cluster face into 12 different sub-regions, as show in Figure 21. And we assign the Mocap markers position as the control point of each face sub-region.

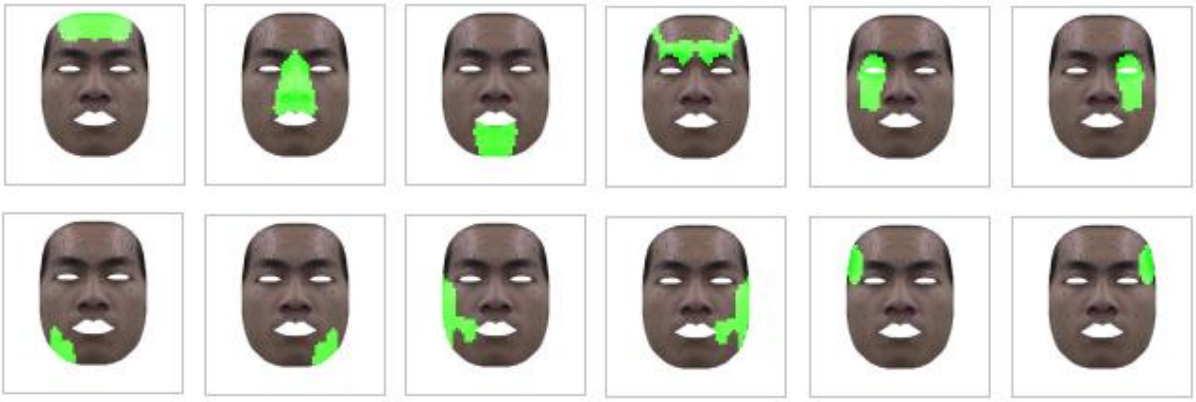


Figure 21: Sub-region of face from normalized cuts.

5.3 The Result of Multi-Layer Analysis and Synthesis

In our research, we employ the modified steerable pyramid to analyze image information for synthesizing the high-band and low-band of novel expression respectively. We integrate the statistics-matching based and high-band enhancement for maintaining the high-band information in synthesis procedure. The following results demonstrate our approach compare with blend shape for synthesizing ground truth.

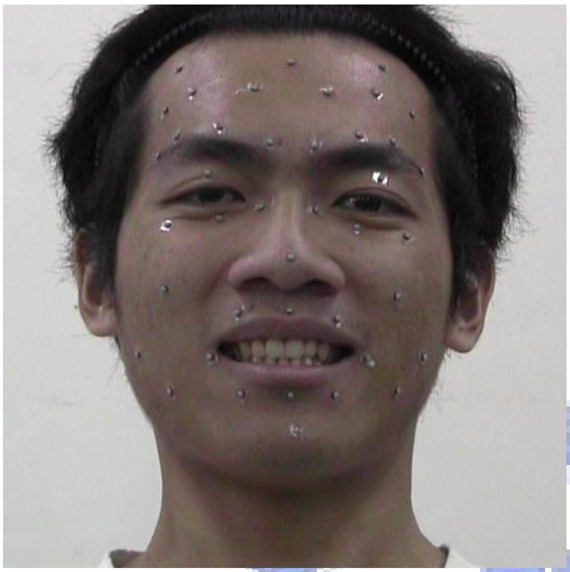
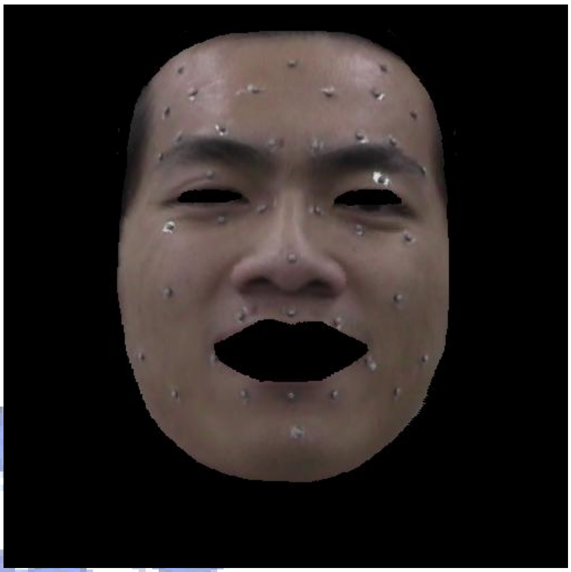
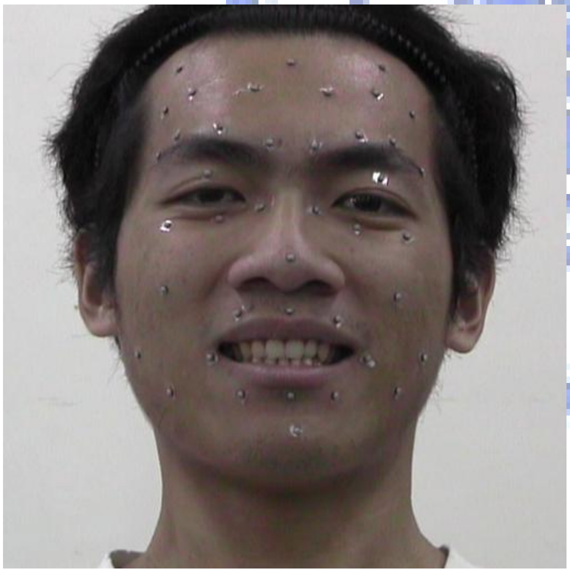
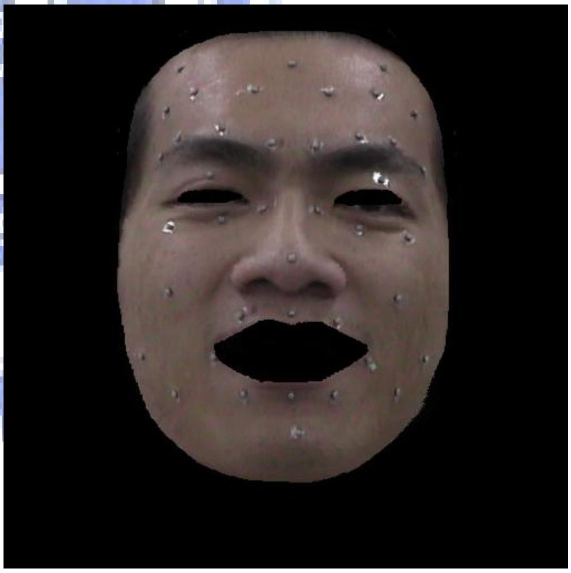
Maintain satisfactory intensity	
	
Ground truth	Blend shape
	
Ground truth	Our approach

Figure 22

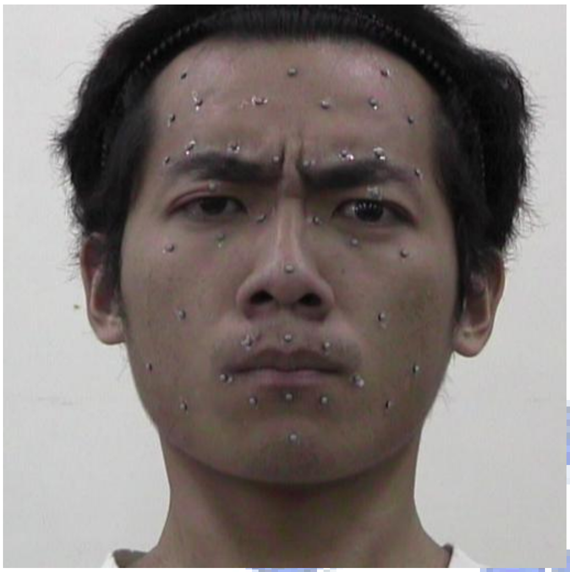
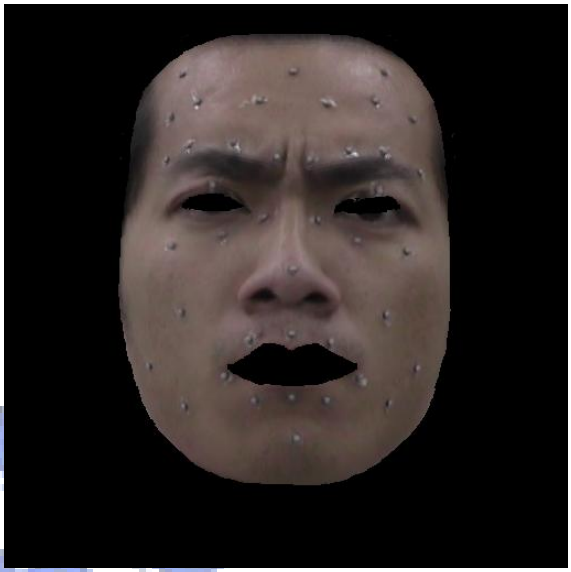
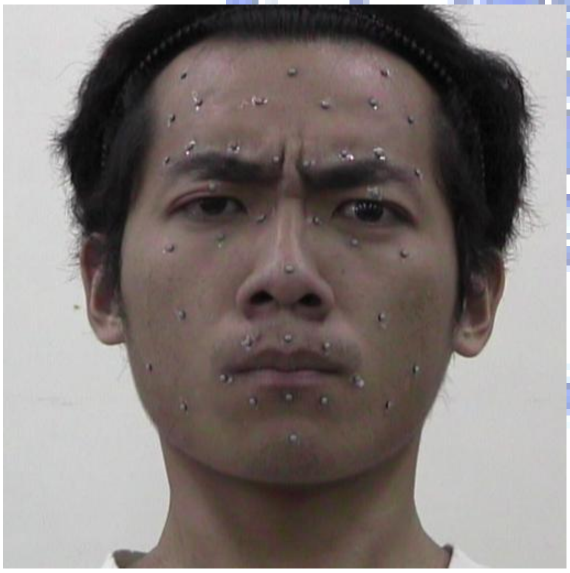
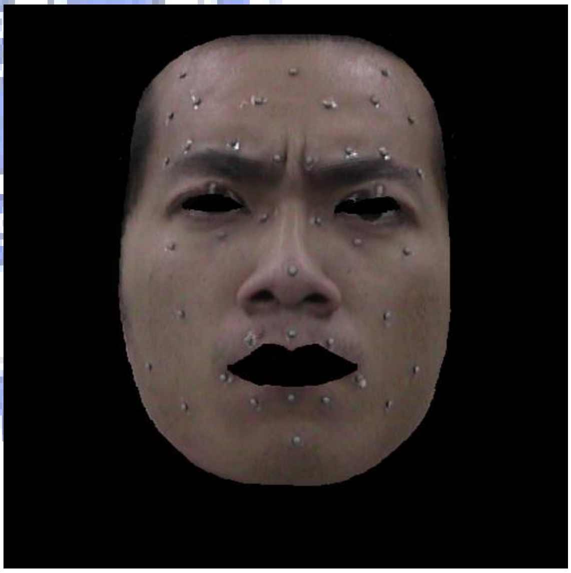
Maintain satisfactory intensity	
	
Ground truth	Blend shape
	
Ground truth	Our approach

Figure 23

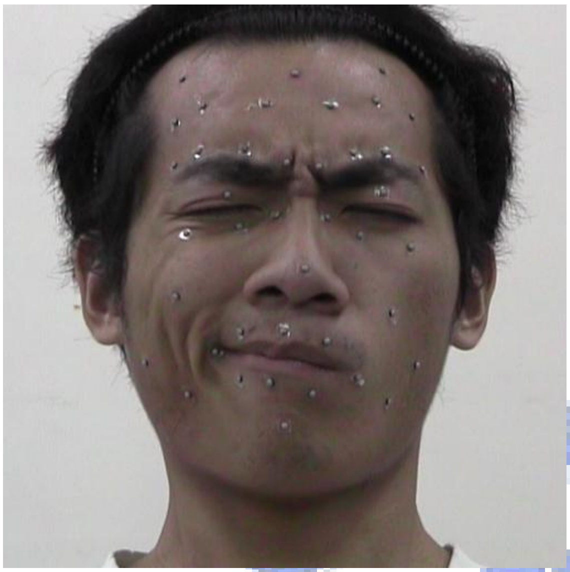
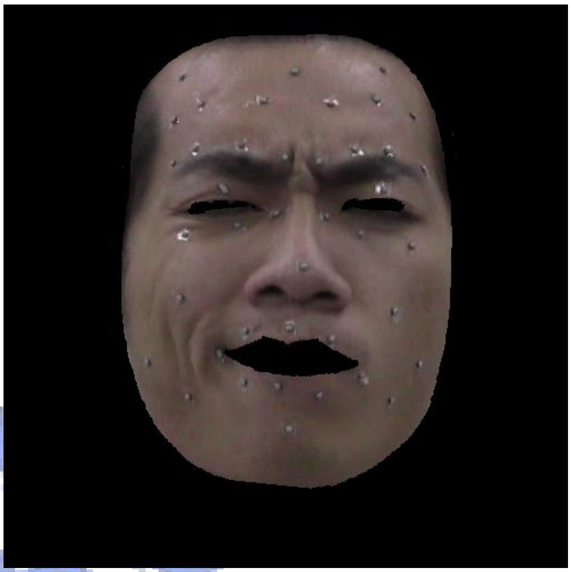
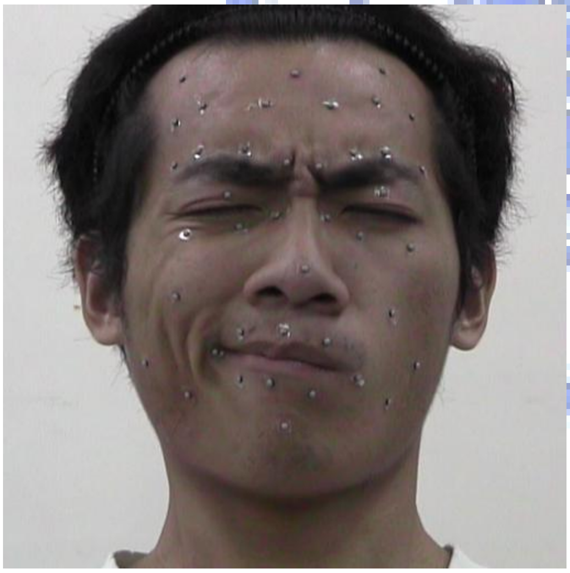
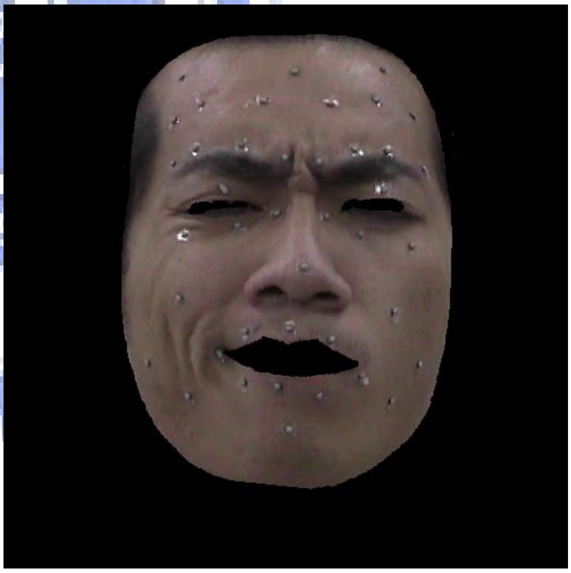
Maintain facial details	
	
Ground truth	Blend shape
	
Ground truth	Our approach

Figure 24

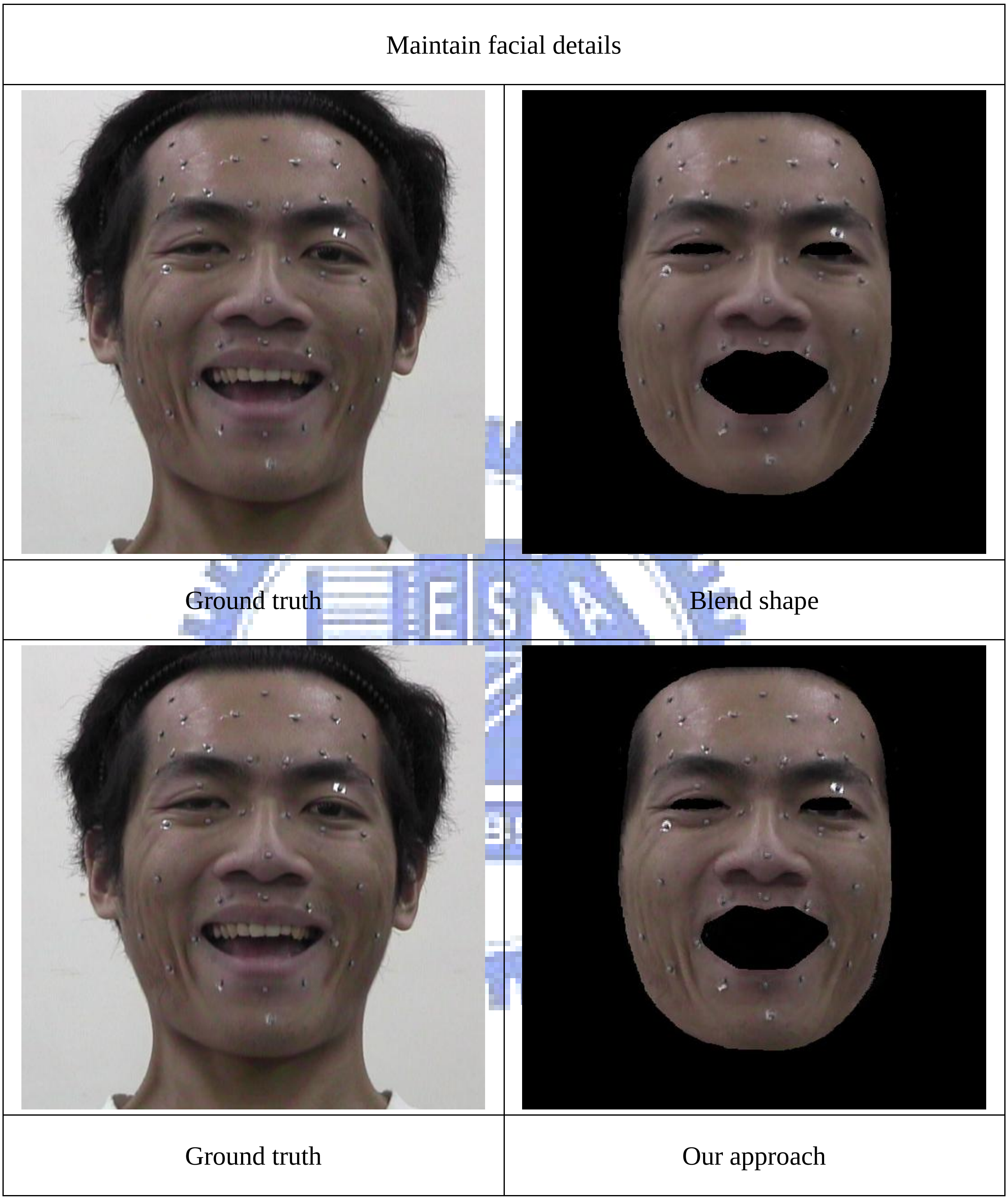

Maintain facial details
Ground truth
Blend shape
Ground truth
Our approach

Figure 25

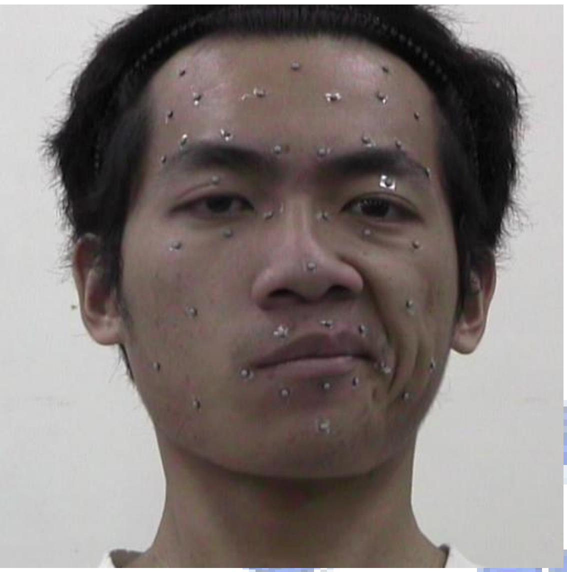
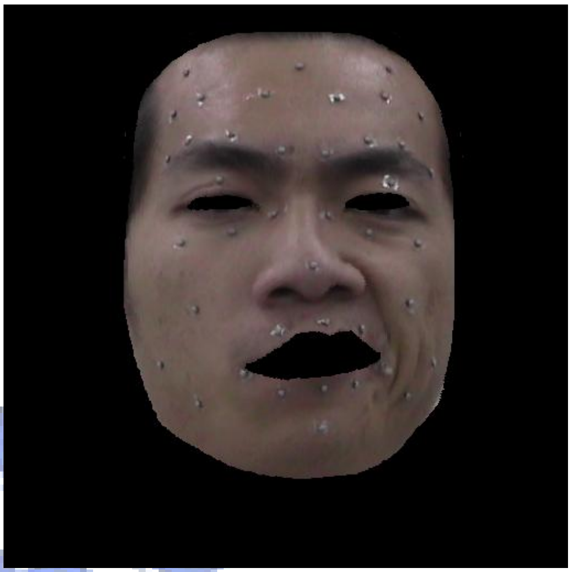
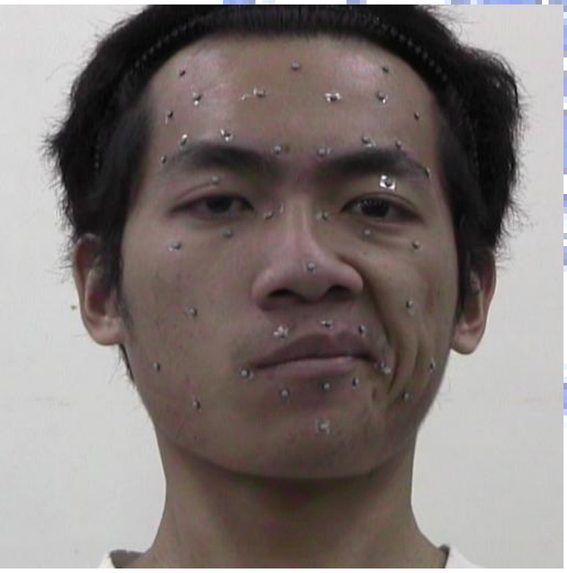
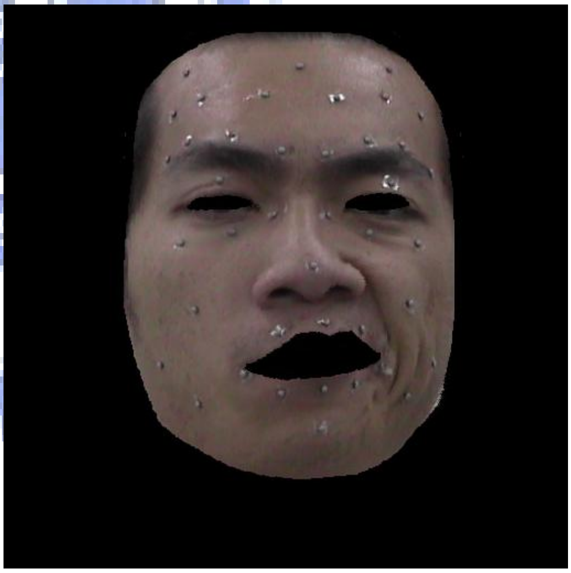
Maintain facial details	
	
Ground truth	Blend shape
	
Ground truth	Our approach

Figure 26

5.4 Application: Facial Animation

We combine the multi-layer analysis and high-band enhancement synthesis with spatial-temporal constraints for producing facial animation. Furthermore, we align the selected prototype in each synthetic region within four frames to avoid texture flicker effect.

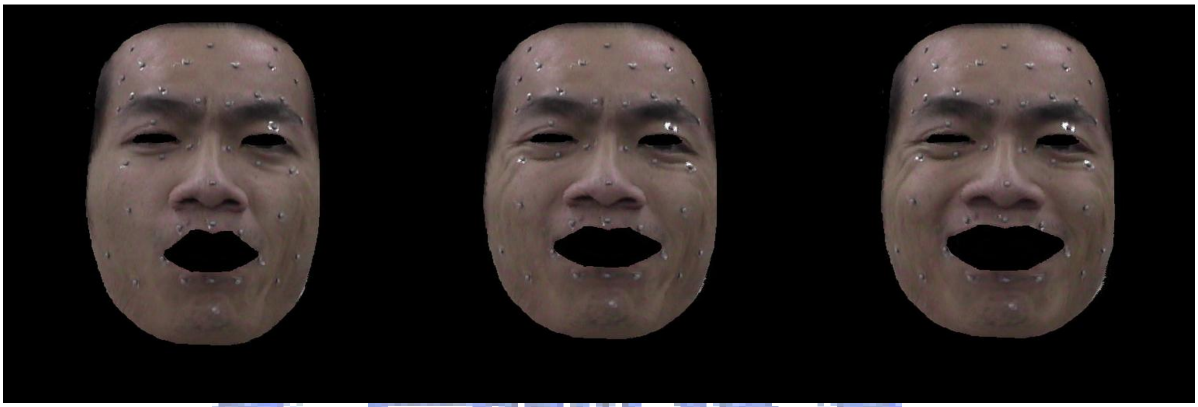


Figure 27: Happy expression.

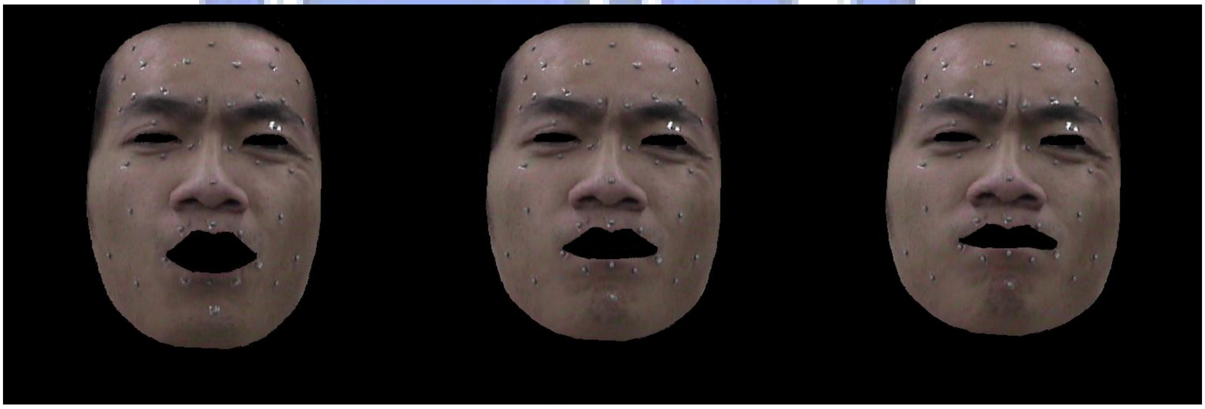


Figure 28: Sad expression.

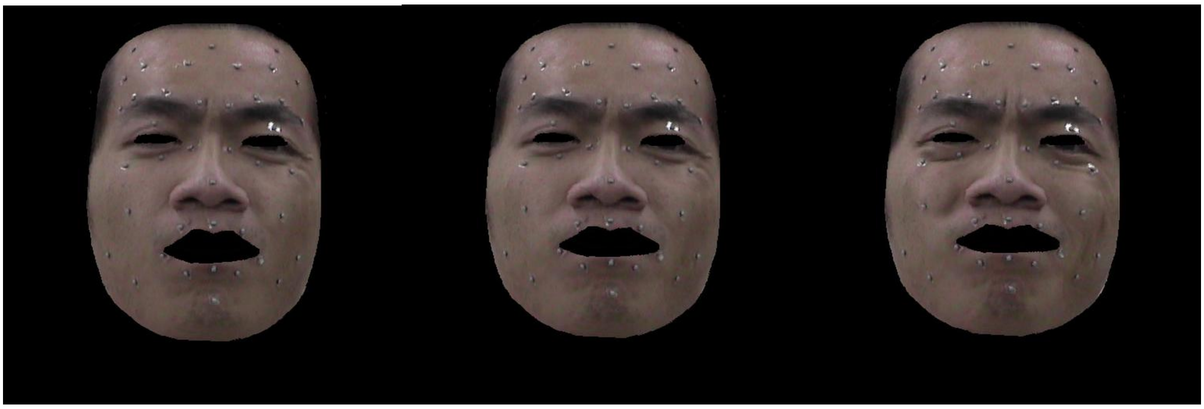


Figure 29: Cry expression.

Chapter 6. Conclusion and Future work

6.1 Conclusion

In this thesis, we propose a multi-layer facial detail analysis and synthesis model to synthesize detail-preserved expressions. We employ correlation between different face region as criterion for partitioning a face into different sub-regions and using those sub-regions to synthesize various appearances. For reducing blurred result and maintaining more detail information, we process the lowest sub-band and other high-band of synthetic image independently. We directly blend the lowest sub-band to synthesize lowest sub-band and consider the blending weight as the criterion to select the proper high-band region for synthesizing high-band. Our approach can synthesize novel photorealistic expressions and apply to producing detail-preserved facial animation.

Our contribution include: (1) a novel multi-layer analysis and synthesis framework for synthesize expressions with fine details and global features. (2) An optimal approach with spatial-temporal constraints and texture alignment for generating detail-preserved facial animation.

6.2 Future work

We use correlation as criterion for normalized cuts to partition face into different sub-regions. Therefore, the grids within each cluster have high correlations. In our future work, we plan to use this concept to retrieve more reliance control points. Besides, since our synthesis policy is selecting the satisfactory high-band information, the high-band detail may flicker when frame change. Therefore, we can apply texture alignment for future work.

Reference

[1] Frédéric Pighin, Jamie Hecker, Dani Lischinski, Richard Szeliski, and David H. Salesin, "Synthesizing Realistic Facial Expressions from Photograph", Proc. of ACM SIGGRAPH '98, ACM Press, pp.75-84, 1998.

[2] Qingshan Zhang, Zicheng Liu, Baining Guo, Demetri Terzopoulos, and Heung-Yeung Shum, "Geometry-Driven Photorealistic Facial Expression Synthesis", IEEE Trans. on Visualization and Computer Graphics Vol.12, No.1, pp.48-60, 2006.

[3] Volker Blanz and Thomas Vetter, "A Morphable Model for the Synthesis of 3D Faces", Proc. of SIGGRAPH '99, ACM Press, pp.187-194. Los Angeles, 1999.

[4] Ian T. Jollife. "Principal Component Analysis", Springer-Veriag, New York, 2002.

[5] Volker Blanz, Curzio Basso, Tomaso, Poggio, and Thomas Vetter, "Reanimating Faces in Images and Video", Proc. of EUROGRAPHICS Vol.22, No.3, pp.641-650, 2003.

[6] Tony Ezzat, Gadi Geiger, and Tomaso Poggio, "Trainable Videorealistic Speech Animation", Proc. of ACM SIGGRAPH 2002, ACM Transactions on Graphics, Vol. 21, Issue 3, pp.388-398, 2002.

[7] Daniel Vlasic, Matthew Brand, Hanspeter Pfister, and Jovan Popovi, "Face Transfer with Multilinear Models", Proc. of ACM SIGGRAPH 2005, ACM Transactions on Graphics, Vol.24, Issue 3, pp.426-433, Aug. 2005.

[8] Aleksey Golovinskiy, Wojciech Matusik, Hanspeter Pfister, Szymon Rusinkiewicz, and Thomas Funkhouser, "A Statistical Model for Synthesis of Detailed Facial Geometry", Proc. of ACM SIGGRAPH 2006. ACM Transactions on Graphics Vol.25, Issue 3, pp.1025-1034, July 2006.

[9] David J. Heeger and James R. Bergeny, "Pyramid-based Texture Analysis/Synthesis",

Proc. of ACM SIGGRAPH’95, New York, USA, pp. 229–238.

[10] William T. Freeman and Edward H. Adelson, “The Design and Use of Steerable Filters”, IEEE Transactions on Pattern Analysis and Machine Intelligence Vol.13, No.9, pp.891–906, 1991.

[11] Tim Weyrich, Wojciech Matusik, Hanspeter Pfister, Bernd Bickel, Craig Donner, Chien Tu, Janet McAndless, Jinho Lee, Addy Ngan, Henrik Wann Jensen, and Markus Gross, “Analysis of Human Faces using a Measurement-Based Skin Reflectance Mode”, Proc. of ACM SIGGRAPH 2006, ACM Transactions on Graphics, Vol.25, Issue 3. pp.1013-1024, 2006.

[12] Li Zhang, Noah Snavely, Brian Curless, Steven M. Seitz, “Spacetime Faces: High Resolution Capture for Modeling and Animation”, Proc. of ACMSIGGRAPH 2004, ACM Transactions on Graphics, Vol.23, Issue 3. pp.548-558, 2004.

[13] Bernd Bickel, Mario Botsch, Roland Angst, Wojciech Matusik, Miguel Otaduy, Hanspeter Pfister, and Markus.. Gross, “Multi-Scale Capture of Facial Geometry and Motion”, Proc. of ACMSIGGRAPH 2007, ACM Transactions on Graphics, Vol.26, Issue 3, Article 33, 2007.

[14] Yosuke Bando, Takaaki Kuratate, and Tomoyuki Nishita, “A Simple Method for Modeling Wrinkles on Human Skin”, Proc. of the 10th Pacific Conference on Computer Graphics and Applications, pp.166-175, October 09-11, 2002.

[15] Bernd Bickel, Manuel Lang, Mario Botsch, Miguel A. Otaduy, and Markus Gross, “Pose-Space Animation and Transfer of Facial Details”, Proc. of the 2008 ACM SIGGRAPH/Eurographics Symposium on Computer Animation, pp.57-66.

[16] Jianbo Shi and Jitendra Malik, “Normalized Cuts and Image Segmentation”, Proc. of IEEE Conf. Computer Vision and Pattern Recognition, pp.731-737, 1997.

[17] Jianbo Shi and Jitendra. Malik, “Normalized cuts and image segmentation”, IEEE Trans. Pattern Anal. Mach. Intell., Vol.22, No.8, pp.888–905, Aug. 2000..

[18] Thaddeus Beier and Shawn Neely, “Feature Based Image Metamorphosis”, Proc. of SIGGRAPH '92, Computer Graphics, pp.35-42, 1992.

[19] Tinku Acharya and Ajoy K. Ray, “Image Processing: Principles and Applications”, Wiley-Interscience 2005.