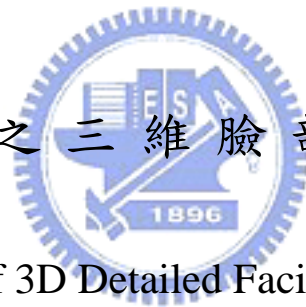


國立交通大學

資訊科學與工程研究所

碩士論文

合成具細紋之三維臉部表情之研究



Synthesis of 3D Detailed Facial Expression

研究生：莊榮元

指導教授：林奕成 博士

中華民國九十五年九月

合成具細紋之三維臉部表情之研究
Synthesis of 3D Detailed Facial Expression

研究生：莊榮元

Student : Chii-Yuan Chuang

指導教授：林奕成

Advisor : I-Chen Lin

國立交通大學
資訊科學與工程研究所
碩士論文



Submitted to Institute of Computer Science and Engineering

College of Computer Science

National Chiao Tung University

in partial Fulfillment of the Requirements

for the Degree of

Master

in

Computer Science

September 2006

Hsinchu, Taiwan, Republic of China

中華民國九十五年九月

合成具細紋之三維臉部表情之研究


研究生：莊榮元

指導教授：林奕成 博士

國立交通大學

資訊工程與科學研究所

摘 要



想要產生一張表情栩栩如生的三維人臉一向都要花費許多人力與時間，即使是經驗豐富的動畫師，仍然要花許多時間在人臉表情的細部調整與修改上。在今日的電影與遊戲製作中，真實表情的動作資料，可藉由三維動態捕捉技術擷取，並且被用來加速三維人臉表情的製作。然而，一般的三維動態捕捉技術僅能擷取較大維度的臉部幾何，對於脸部細微的紋理變化則很難有效捕捉。

在本篇論文中，我們針對脸部細紋的合成與表情的編輯，提出了一個以特徵點驅動的具細紋之三維脸部表情合成系統，此系統可分為兩個主要部分—三維脸部表情的編輯與脸部細微紋理的擷取。我們的系統提供了使用者一個較為直覺的編輯介面，並利用一群指定的脸部特徵點之間的幾何關係，使用一群範例影像估算並合成出新的表情的近似三維幾何結構與相對的法向量或二維外觀影像。除了用在單純的具細紋表情之編輯上，此一技術所產生的脸部細紋資料亦可進一步用在強化現有三維動態捕捉技術所產生的人臉表情上。

Synthesis of 3D Detailed Facial Expression

Student: Chii-Yuan Chuang

Advisor: Dr. I-Chen Lin

Department of Computer Science
National Chiao Tung University

ABSTRACT

Producing a life-like 3D facial expression is usually a labor-intensive process. Even for an experienced animator, it spends much time on trial-and-error procedure. In movie and game industries, motion capture techniques and 3D scanning, acquiring motion data from real persons, are used to speed up the production. However, acquiring dynamic and subtle details, such as wrinkles, on a face are still difficult or expensive.

In this thesis, we developed a feature-point-driven approach to synthesize novel expressions with details. Our work can be divided into two main parts: acquisition of 3D facial details and expression synthesis. By employing geometric relation between specific feature points, our system provides an intuitive editing tool to synthesize 3D geometry and corresponding 3D detailed normal or 2D textures of novel expressions. Besides expression editing, the proposed method can also be extended to enhance existing motion capture data with facial details.

Acknowledgements

First of all, I would like to thank my advisor, Dr. I-Chen Lin, for his guidance in the past two years. Also, I appreciate all members of Computer Animation & Interactive Graphics Lab for their help and comments. Finally, I am grateful to my family for their support and encouragement.



Contents

摘 要.....	I
ABSTRACT.....	II
ACKNOWLEDGEMENTS.....	III
CONTENTS	IV
LIST OF FIGURES.....	V
1. INTRODUCTION	1
1.1 Motivation.....	1
1.2 Introduction	2
1.3 System Overview.....	3
2. RELATED WORK.....	5
3. ACQUISITION OF EXPRESSION DETAILS	11
3.1 Reconstruction of 3D Geometry	11
3.2 Normal Recovery.....	11
3.3 Normal difference map	14
3.4 Post Processing of Normal Difference Maps.....	16
4. FEATURE-POINT-DRIVEN SYNTHESIS OF NOVEL EXPRESSIONS	17
4.1 Calculation of Blending Weights.....	17
4.2 Synthesis of Textures and Normals	20
4.3 Editing and Synthesis of 3D geometry.....	22
4.3.1 Editing Interface.....	23
5. EXPERIMENT AND RESULTS.....	25
5.1 Collection and Pre-Processing of Prototype Images	25
5.2 Results.....	26
6. CONCLUSION AND FUTURE WORK	34
6.1 Conclusion.....	34
6.2 Future Work	34
7. REFERENCE.....	35

List of Figures

Figure 1:(a) synthetic Neutral Face, (b) synthetic face of lifting eyebrow without details, (c) synthetic face of lifting eyebrow with estimated details, and (d) real face of lifting eyebrow.	1
Figure 2: The system overview	4
Figure 3 : Bregler’s video rewrite system. [5]	6
Figure 4: Extract an expression from the vector space of faces, and add it to the neutral face. [2]	7
Figure 5: Structure light image (a) and range image captured (b) by L. Zhang’s system. [25].....	8
Figure 6: The overview of Deng’s blendshape animating system [6].	9
Figure 7: Synthesis of facial detail by Golovinskiy et al [11]. (a) low-resolution mesh obtained from a commercial scanner. (b) synthesize detail on (a) using statistics extracted from high-resolution meshes in their database. (c) age the face by adjusting the statistics to match those of an elderly man.	10
Figure 8: Normal recovery	12
Figure 9: The left figure is the acquired image and the right one is the recovered normal array illustrated in R,G,B channels.	13
Figure 10: A face example.	14
Figure 11: The estimated surface normal (solid arrow) can be regard as actual surface normal (small dotted arrow) plus estimated error due to color variation (long dotted arrow).....	15
Figure 12: A normal difference map	15
Figure 13: The effect of normal difference map. The first row shows face normals of neutral expression. The second row is an expressed face. Without modifying surface normals, the surface normal is wrong. In third row, we use normal difference map to add the difference (vector c) between normal of neutral face (vector a) and normal of expressed map (vector b).....	16
Figure 14: The original normal difference map (a). And we use adaptive Gaussian filter to eliminate noise and try to preserve features like wrinkles (b).	16
Figure 15: w_i represents the blending weight. A novel is synthesized by interpolating prototypes.	18

Figure 17: We roughly divide a face into 8 sub-regions to increase the probable novel expression.....21

Figure 18: Left, without blending on sub-region boundary, the hard edge between two sub-regions is clear. Right, the hard edge is eliminated by blending on sub-region boundary.....22

Figure 19: The 3D head model with control points. The green triangles present the control points. User can modify its position by a keyboard and mouse.24

Figure 20: A set of prototype image consist of 3 different views. Left and right ones are used to recover the 3D geometry of feature points. Center view is use to acquire normal maps.25

Figure 21: The neutral face we choose. The blue marker is put on several specific position to observe the variation of expressions.....26

Figure 22: Our 3D head model with neutral face.27

Figure 23: The synthesized head model. The picture at right corner is corresponding texture.....27

Figure 24: Another synthesized head model. The picture at right corner is corresponding texture.....28

Figure 25: The synthesized head model. The picture at right corner is corresponding texture.....28

Figure 26: Synthesize a normal difference map of female to a male model.29

Figure 27: The synthesized head model (left) and a real picture.29

Figure 28: Figure (a),(b),(c),and (d) are four sets of synthesized expressions. The forehead wrinkles in (a), (c) and (d) are thinner because error during interpolation procedure. If we applied Gaussian filter (described in chapter 3.4) less times, there are more noises in the synthesis result (such as (b)).....30

Figure 29: Figure (a),(b),(c),and (d) are four sets of synthesized expressions.31

Figure 30: Figure (a),(b),and (c) are three sets of synthesized expressions.32

Figure 31: The prototypes which we used to synthesize fig28 and fig22~25.33

Figure 32: The prototypes which we used to synthesize fig29 and fig27....33

Figure 33: The prototypes which we used to synthesize fig30 and fig26.33

1. Introduction

1.1 Motivation

Although facial animation is popularly used in various kinds of media, generating realistic faces is still a labor-intensive work for animators. That is due to a face is the most expressive part of our appearance, and any subtle difference may have different meanings.

Recently, motion capture (mocap) techniques are popularly utilized to speed up the production of 3D facial animation. Dozens of markers are placed on feature points of a subject's face. Motion trajectories of these features can be recorded to drive meshes of a head model. However, there are still subtle portions, such as wrinkles or creases, whose variations are smaller than markers. These subtle portions are difficult to be acquired by mocap techniques.

In Figure 1, we can see the difference between a real face and synthetic faces with or without the facial details. With the same motion of lifting eyebrow, the detailed face is more expressive.

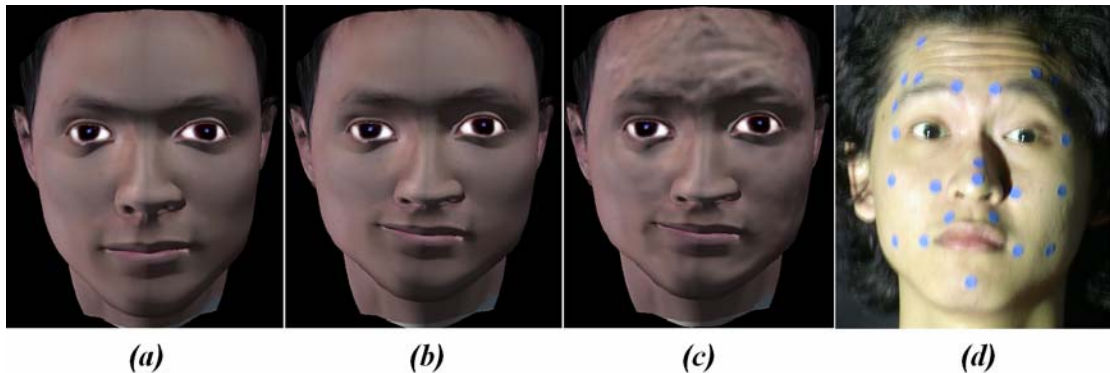


Figure 1:(a) synthetic Neutral Face, (b) synthetic face of lifting eyebrow without details, (c) synthetic face of lifting eyebrow with estimated details, and (d) real face of lifting eyebrow.

To overcome the lack of facial details, conventional producing procedures may use labor-intensive adjustment or 2D texture, which is difficult to relight and is used on specific applications such as close-up animation. On the other hand, 3D laser scanners with high resolution can also grasp facial details in the cutting-edge movie industry. Such an approach provides convincing results, but it is expensive and inefficient to scan all expressions which we need.

1.2 Introduction

As described in section 1.1, current approaches used to generate 3D facial expressions still have several problems. In this thesis, we would like to tackle the problems. Our first challenge is about the estimation of facial details.

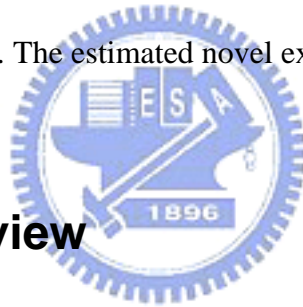
In our research, we try to synthesize novel expression from a set of prototype images. However, if we just applied the synthesized texture to a 3D head model, the model will be difficult to relight or alter view directions (such as image-based approaches). In order to overcome the problem of relighting and altering view directions, our system should reconstruct the detailed geometry. The reconstruction of geometry could be achieved by stereopsis. However, prominent features on our face are sparse, and the point correspondences of detailed wrinkles are difficult to evaluate. We make use of stereo triangulation for the approximate geometry of facial expression. To estimate the facial details, we propose using a shape-from-shading (SFS) method since SFS techniques can avoid unreliable point matching. In our work, we take several sets of photos as our sample expressions. Each set of photos has 3 different views, and all these photos are taken under the same single light source. The central view is used to estimate facial details. The other two views are used to reconstruct 3D positions of feature points.

The second challenge is expression synthesis. Since dragging feature points is one

of the most intuitive way for editing, our goal is to synthesize novel expressions and facial details according to configurations of user-defined feature points.

Given a set of feature point positions adjusted by users, our system utilizes an optimization approach to select appropriate prototype expressions from the data pool and calculates the best weights for blending. For interactive editing, the surface details are represented in terms of normal difference maps, which can be efficiently rendered by pixel shaders.

The major contribution of our system is that we propose an intuitive and inexpensive framework for acquisition and editing detailed expressions. Sample expressions are evaluated by using feature-point-driven face reconstruction and shape-from-shading techniques. The novel expressions can be synthesized by optimal projection in expression space. The estimated novel expressions can also be retargeted to models of other subjects.



1.3 System Overview

Our system can be divided into two parts: offline processing and online novel expression synthesis. Figure 2 shows an overview of our system.

Offline processing is about the preprocessing of our prototype images. After gathering our prototypes, we preprocess these images for conveniently synthesizing novel expression. This work only needs to be done once, so we separate it from online synthesizing procedures. This offline part will describe in chapter 3.1 and 3.2.

The online part will separately described in chapter 3 and 4.

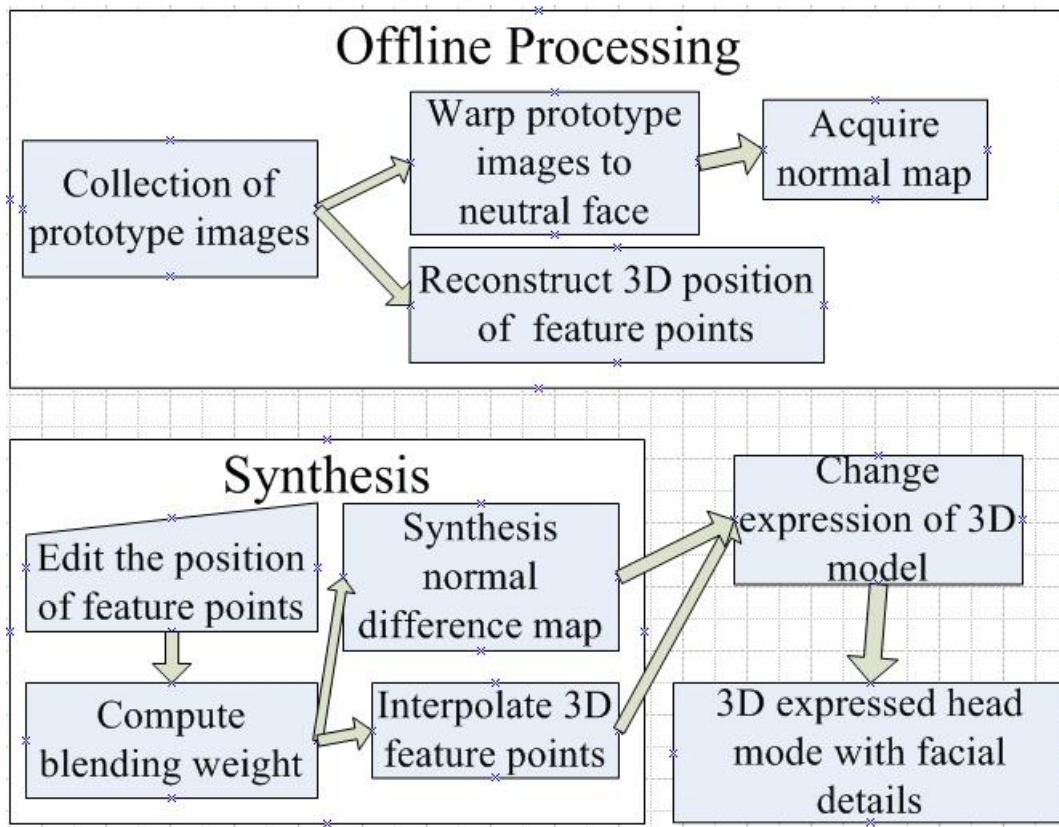


Figure 2: The system overview



2. Related Work

For facial animation, generating life-like facial expression is a challenging problem. Here, we will review some previous researches.

Facial animation can be roughly divided into two groups according to their basic structures. The first one is image-based facial animation, and the other is 3D model-based facial animation. Image-based approaches employ one or several real facial images to synthesize novel images. Because of the use of real images, image-based approaches can reach a photorealistic quality while there are sufficient samples in the data pool. The drawbacks of image-based approaches are difficult to relight, and to alter view directions. Image-based approaches are usually used for close-up animation.

In image-based approaches, Beier and Neely proposed a feature-based morphing approach [1]. This approach first used in Michael Jackson's MTV "black and white" and got a success. Now it is widely used in several image-based animations. In video rewrite [5], Bregler et al. used existing footage to synthesize a new video sequence. The system records mouth images in the training footage to match the phoneme sequence of the new audio track. Using this approach can automatically create new video of a person pronouncing words that he/she did not speak. Another approach to create talking heads was proposed by Ezzat in 2002, [7]. Since Bregler's system can not produce the mouth that isn't in their database, Ezzat et al. developed a variant of the multidimensional morphable model (MMM) to synthesize previously-unseen mouth configuration from the set of mouth prototypes.

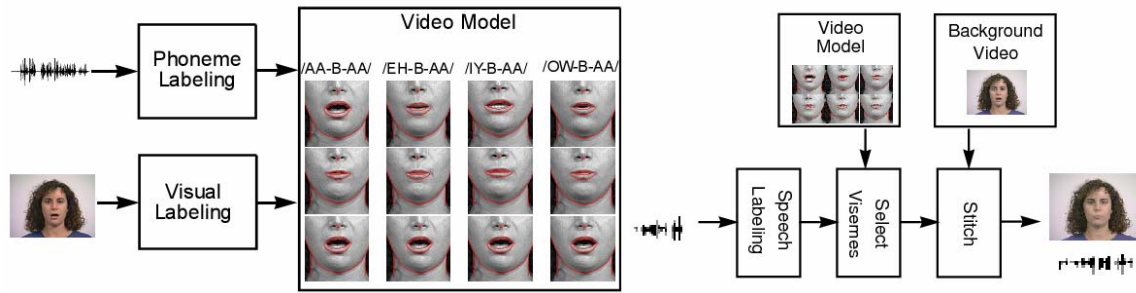


Figure 3 : Bregler's video rewrite system. [5]

Expression ratio image [16] is another method to generate expression with details. They captured the illumination change of one person's expression from images (called expression ratio image (ERI)). With geometric warping first, ERI can be mapped to other person's face image. Tu et al. [22] adapted the ERI approach for 3D model-based animation. They transformed ERI into normal maps. Then the normal map sequence can be applied to 3D models.

In contrast to image-based facial animation, 3D model-based animation is suitable to use in a 3D environment. However, it needs a lot of control parameters to deal with modeling, facial motion, and even surface reflection.

Waters [23] proposed pioneering face models with physical muscle (physical-based approach). They simulated the deformation of face muscle and tissue by utilizing an approximate model and physical model. The face model can be controlled by a set of parameters. To generate desired expressions, users need to adjust the value of parameters to control the face muscle. Their approach can simulate muscle and face tissue, but simulating facial details (such as subtle wrinkles and creases) are not efficient and are difficult to control. In Sifakis' research [21], they proposed an anatomically accurate face model controlled by muscle activations and kinematic bone degrees of freedom. Their novel algorithm can automatically compute control values for sparse motion capture marker input.

While physical based approaches simulate facial expression by physical model, the other researches extract the information from images or synthesize dynamic texture from images. Making face [12] takes image sequences of six camera views of an actress' face where 182 markers are pasted. Then the 3D deformation of the face can be accurately tracked by the large set of markers. At the same time, the image sequence (after removal and inpainting of marker regions) is used to create a texture map sequence for the 3D polygonal face model. The approach can result in a life-like performance, but it still has the weakness of texture map –difficult to relight and retarget.

Based on a large set of 3D scanned face examples, Blanz et al. [3] built a morphable head model. First, they transformed the shape and texture of the examples into a vector space representation. Then, by the linear combination of prototypes, new faces and expressions can be modeled. This approach is useful for model textured 3D faces. In 2003 [2], they further transferred facial expressions by computing the difference between two scans of the same person in a vector space of faces. This approach could extract the difference between a neutral face and an expressed face (ex. a smiled face). Then, the extracted expression could be added to images.

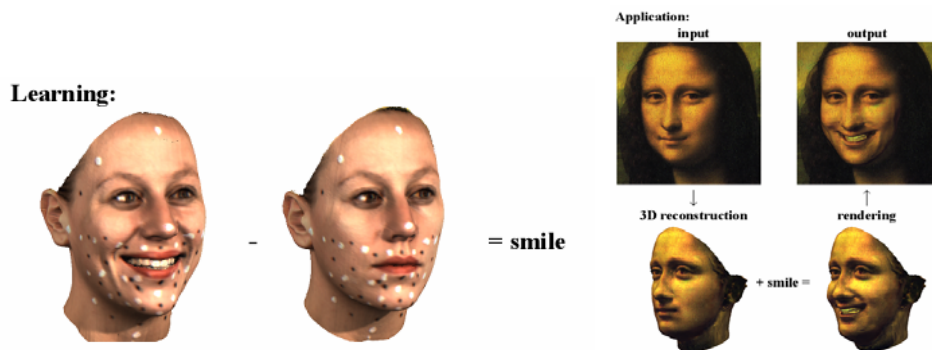


Figure 4: Extract an expression from the vector space of faces, and add it to the neutral face. [2]

Although the range data of a static face can be captured by laser scanners, the

facial expression of a dynamic face can not be captured by laser scanners. Zhang et al. [25] proposed a structure light approach to capture the dynamic variation of a face. Their system utilizes two projectors to project stripe patterns and six cameras. Through the deformation of stripe patterns, the depth map can be calculated. Besides, they presented a keyframe interpolation technique to synthesize in-between video frame and a controllable face model.

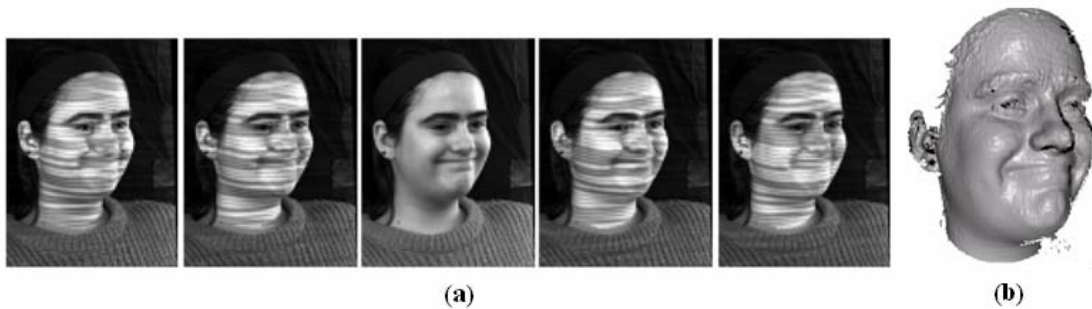


Figure 5: Structure light image (a) and range image captured (b) by L. Zhang's system. [25]



A geometry-driven approach proposed by Zhang et al. [27] synthesizes facial expression through the relation between positions of specific feature points and expressions. They utilized a vector space based on captured expression prototypes. Then with new positions of feature points input by users, a new 2D expression can be synthesized by solving an optimization problem in this vector space. We adapt their approach to extract expression details in normal map form (details will be introduced in other chapters later).

The blendshape (shape interpolation) method is a popular approach to computer animation (ex. the “Gollum” model in movie “Lord of the Rings”). However the cost of blendshape face animation is considerable. Therefore, Deng et al. [6] propose an approach to save production time. Their research is a semi-automatic technique to directly animate popularized 3D blendshape face models by mapping facial motion

capture data spaces to 3D blendshape faces paces. They defined reference mocap-video pairs with motions of human subjects and corresponding video frames. Then, they manually tune the blendshape weights to match each reference mocap-video pairs. Finally the blendshape weights to any new facial motion capture frame can be derived from manual-tuned weights and corresponding reference mocap-video pairs. The system can save the time of manually tuning blendshape weights.

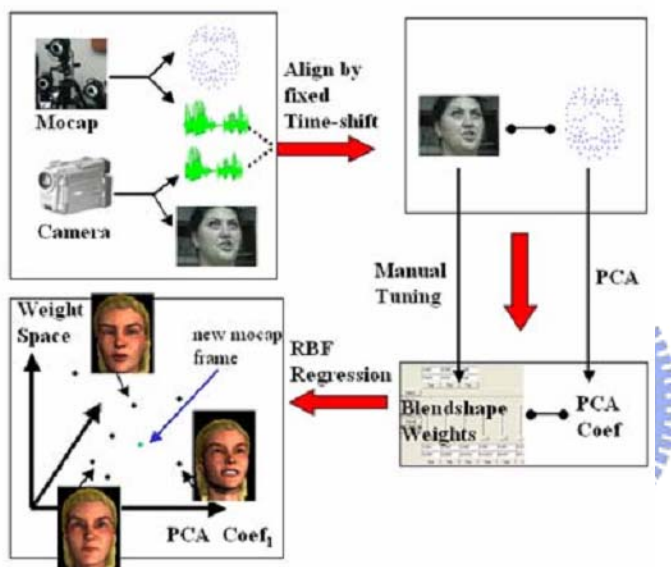


Figure 6: The overview of Deng's blendshape animating system [6].

In addition to facial details due to different expressions, Golovinskiy et al [11]. develop a statistical model for synthesizing detailed facial geometry due to aging. They acquire high-resolution face geometry of people across a wide range of ages, genders, and races. For each scan, they separate the skin surface details from a smooth base mesh using displaced subdivision surfaces. Then, they analyze the resulting displacement maps using the texture analysis/ synthesis framework of Heeger and Bergen, adapted to capture statistics that vary spatially across a face. Finally, they use the extracted statistics to synthesize plausible details on face meshes of arbitrary subjects.

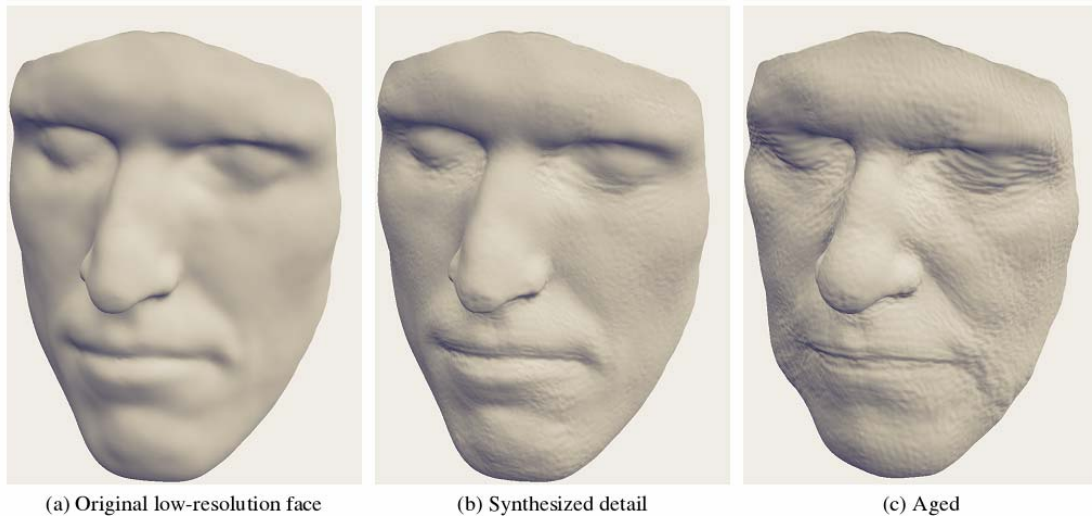


Figure 7: Synthesis of facial detail by Golovinskiy et al [11]. (a) low-resolution mesh obtained from a commercial scanner. (b) synthesize detail on (a) using statistics extracted from high-resolution meshes in their database. (c) age the face by adjusting the statistics to match those of an elderly man.

In addition to facial animation, our research is related to extract the surface variation from images. Here, we introduce some research related to extract surface variation. Horn et al. [13] proposed that the surface normals can be recovered from the intensity variations of an image. He took an optimization method that iteratively minimized errors. Fang et al. [8] adapted Horn's approach. They simply utilize Lambertian reflection model to extract the normal map from single image. Their approach spends less time and doesn't need expensive equipment.

3. Acquisition of expression details

In our research, we synthesize novel expression from several sets of prototype samples. In order to acquire 3D detailed face structure, we use stereo triangulation for conspicuous markers in two views. While morphing a generic model according to these 3D positions, we can acquire an approximate geometry. However, stereo triangulation is unreliable for detailed variation, since the point correspondences are difficult to find. Therefore, we utilize a modified shape-from-shading (SFS) technique to deal with the facial details.

The reconstruction of 3D geometry will be introduced in 3.1 and our SFS will be described in 3.2. Section 3.3 and 3.4 describe the process of our estimated normals.

3.1 Reconstruction of 3D Geometry

Through camera calibration and stereo reconstruction, we can reconstruct 3D position of each marker on the subject by simply using a purely algebraic approach.

Given the projection matrices \mathbf{M} and \mathbf{M}' and corresponding points p and p' , we can rewrite the equations $p = \mathbf{M}\mathbf{P}$ and $p' = \mathbf{M}'\mathbf{P}$ as:

$$\begin{cases} p \times \mathbf{M}\mathbf{P} \equiv 0 \\ p' \times \mathbf{M}'\mathbf{P} \equiv 0 \end{cases} \Leftrightarrow \begin{pmatrix} [p_x]M \\ [p'_x]M \end{pmatrix} \equiv 0$$

This is an over-constraint system and we can solve \mathbf{P} easily by a linear least-square method.

3.2 Normal Recovery

Horn [13] and H. Fang et al. [8] proposed recovering the surface from an image under the Lambertian reflectance model. Human skin, in contrast, has view-invariant diffuse reflection and also specular and subsurface scattering reflection. But, the

diffuse term still dominates the image intensity. To make our recovery process more efficient and feasible, we use a modified SFS approach similar to the one proposed by H. Fang [8]. The modified SFS approach we used is work with a single image under a single light source. There are other approaches that could estimate normal from multiple light sources. However, these approaches are only good for still object. When the subject is not still, the corresponding points between images with different light source is difficult to track.

Because of the diffuse term dominating the image intensity, our approach utilizes Lambertian reflection model taking into account of diffuse term only, we also assume that the color of the skin is uniform. Based on these assumptions, the variations of skin color in image is only due to the variations of the angle between surface normals and incident light directions. With the assumptions, the normal can be efficiently extracted from single images as follow.

As shown in Figure 8, let L be the unit vector of the light source direction. To evaluate the surface normal N_{xy} of a pixel in the image, first, we have to estimate a projection vector P_{xy} as shown in Eq. (1).

$$P_{xy} = \nabla I_{xy} - (\nabla I_{xy} \cdot L)L \quad (1)$$

where $\nabla I_{xy} = (\frac{\partial I_{xy}}{\partial x}, \frac{\partial I_{xy}}{\partial y}, 0)$ is the image gradient and P_{xy} means the projection of vector ∇I_{xy} to the plane perpendicular to L .

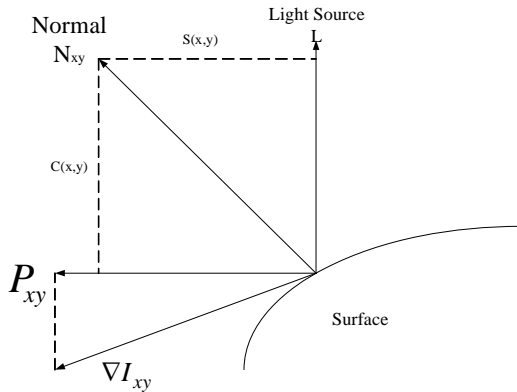


Figure 8: Normal recovery

First, we project the gradient of the image to the plane perpendicular to the incident light vector. The normal of the pixel can be calculated through the cosine and the sine value estimated from the image.

We assume that darkest intensity value, I_{\min} , implies the intensity of ambient light in the scene and the brightest value, I_{\max} , indicates the intensity when a pixel faces the light source. The I_{\min} and I_{\max} can be simply calculated by the sum of R, G and B channel. That is

$$I_{\max} = \max((R_{xy} + G_{xy} + B_{xy})/3)$$

$$I_{\min} = \min((R_{xy} + G_{xy} + B_{xy})/3)$$

where R_{xy} , G_{xy} , and B_{xy} mean the values of Red, Green, and Blue channels, for pixels (x, y) . Then the cosine of angle between the surface normal and the incident light direction can be evaluated as follows:

$$C(x, y) = (I_{xy} - I_{\min}) / (I_{\max} - I_{\min})$$

Therefore, the sine value $S(x, y)$ between the surface normal and the incident light direction can be calculated:

$$S(x, y) = \sqrt{1 - C(x, y)^2}$$

After the calculation of $C(x, y)$, $S(x, y)$ and P_{xy} , the normal can be estimated through eq.(2)

$$N_{xy} = C(x, y)L + S(x, y)P_{xy} / \|P_{xy}\| \quad (2)$$

Finally, the normal vector will be normalized. We demonstrate the estimated normals in Figure 9, where red, green, and blue components of each pixel represent the x, y and z components of normal vector.

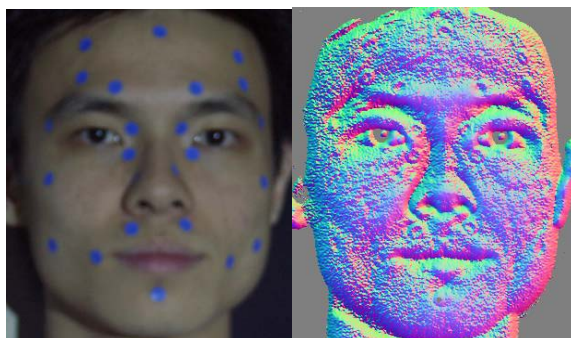
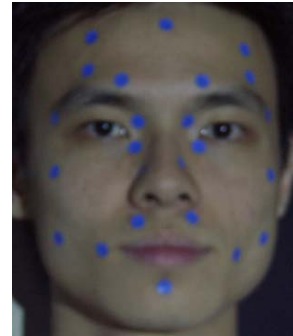


Figure 9: The left figure is the acquired image and the right one is the recovered normal array illustrated in R,G,B channels.

3.3 Normal difference map

In the above subsection, we describe how to extract surface normals from a single image. Nevertheless, when we applied the shape-from-shading (SFS) technique based on the uniform-skin-color assumption, some defective normals occurs. As we know, the human skin color is not completely uniform. Furthermore, color variations on acnes, scars etc. will make the image gradients changes dramatically. Hence, we propose using normal difference map to alleviate defects.



In Figure 10, there are several situations may lead to computing errors. The first one is discontinuous regions such as eyebrows, lip and eyeballs. This type of color discontinuity is large and occurs on expectable regions. Thus, we can simply skip normal estimation in these regions.

Figure 10: A face example.

The second condition is unusual color discontinuity which occurs randomly according to the conditions of different persons, such as acnes or scars. This kind of discontinuity is often small. Thus, we can alleviate the error by normal difference map described bellow. If the region of this discontinuity is abnormally large, manual assistance is required.

The third condition leads to error is the gradual variation of skin color. Although the color of facial skin is mostly continuous, it has gradual variation between some regions. This variation will cause error of the estimated normal values. Fortunately, the skin color just varies gradually, thus we can also eliminate the effect which due to color variation by recorded normal difference instead of estimated normal.

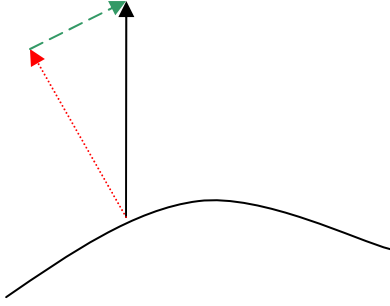


Figure 11: The estimated surface normal (solid arrow) can be regarded as actual surface normal (small dotted arrow) plus estimated error due to color variation (long dotted arrow)

As shown in Figure 11, the estimated surface normal can be regarded as actual surface normal plus estimated error due to gradual color variation. Therefore, we propose an approach called normal difference map to eliminate the estimated error. This approach is shown in Figure 12. Since the estimated error due to gradual variation of skin color is almost the same in difference expression, we can eliminate

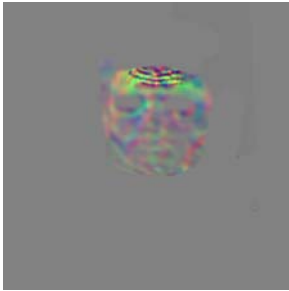


Figure 12: A normal difference map

the estimated error by the method bellow. First, we select a neutral face from our collection of expression prototypes. Then the normal map can be acquired from image by approach described in section 4.1. Finally the normal difference map can be calculated by subtraction of the normal map of expressed face to the normal map of neutral

face. The procedure can be formulized as follow:

$$NDM = NM_{exp} - NM_{neu}$$

where NDM is the normal difference map, NM_{exp} is the normal map of novel expression and NM_{neu} is the normal map of the neutral face.

The information, which contains in the normal difference map, is the difference of face normal between the neutral face and the face of novel expression. This means

that we can add the facial details to 3D model simply by modifying the original surface normal according to the normal difference map. The third row of Figure 13 shows the calculation of normal difference map.

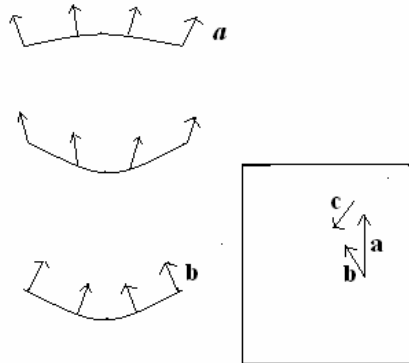


Figure 13: The effect of normal difference map. The first row shows face normals of neutral expression. The second row is an expressed face. Without modifying surface normals, the surface normal is wrong. In third row, we use normal difference map to add the difference (vector c) between normal of neutral face (vector a) and normal of expressed map (vector b).

3.4 Post Processing of Normal Difference Maps

Because of the pixel alignment problem, input noise, and digitization, etc, our normal difference maps have some unavoidable estimation errors. We utilize an adaptive Gaussian filter to reduce noise problem. Around the wrinkle region, we select the 3-by-3 filter mask. In other regions, a larger mask filter will be used. This procedure can remove the noise effectively and make the normal difference map much smoother. Figure 14 shows the result where the adaptive Gaussian filter is applied.

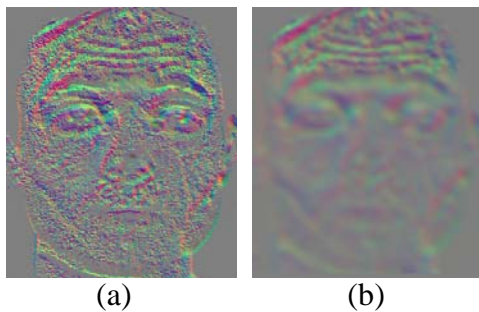


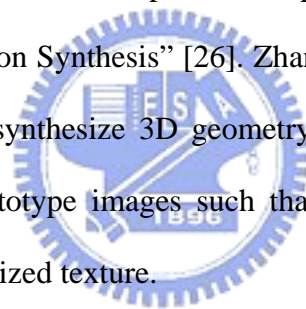
Figure 14: The original normal difference map (a). And we use adaptive Gaussian filter to eliminate noise and try to preserve features like wrinkles (b).

4. Feature-Point-Driven Synthesis of Novel Expressions

In the previous chapters, we describe how to acquire facial details from an image. However, it is infeasible to acquire all possible facial expressions. Thus, we develop an approach to synthesize novel expressions from prototypes.

To synthesize a novel expression, we assume that expressions are highly related to the movement of specific feature points. Here, we assume that similar expressions will have similar movements of feature points.

Q. Zhang et al proposed an optimized approach in “Geometry-Driven Photorealistic Facial Expression Synthesis” [26]. Zhang et al focused on synthesizing 2D textures, but we further synthesize 3D geometry and normals and also applied additional constraints on prototype images such that we can acquire facial details (surface normal) from synthesized texture.



4.1 Calculation of Blending Weights

As described before, our system synthesizes not only 2D textures but also 3D geometries. We take the concept of vector space interpolation to deal with both of these two targets. That is, we regard the expression as the combination of 3D geometry and appearance (2D textures or normal difference maps). Because the deformation of a face muscle is continuous, the geometry of a specific wrinkle can be interpolated by several geometries of related wrinkles with proper interpolating weight. Furthermore, a 2D image presents the difference of shading. Because the different shading of expressions under the same lighting condition presents the difference of face normal, the interpolation of two images (2d expression) represents

the corresponding variations of interpolated surface normal. If we establish a proper vector space to represent the 3D geometry and 2D appearance of expressions, we could approximate novel expression (include 3D geometry and surface normal) by interpolating several prototypes with proper blending weights. Therefore, to synthesize a novel expression we have to calculate the blending weights first. The interpolation can be represented as Figure 15 and the calculation of blending weights is described below.

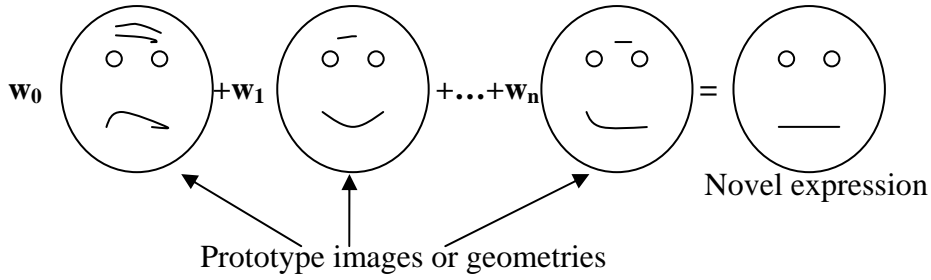


Figure 15: w_i represents the blending weight. A novel is synthesized by interpolating prototypes.

To calculate the blending weights, we utilized an expression vector space. Each expression is represented as $E_i = (G_i, I_i)$ where E_i , G_i represents an expression, and geometry respectively. $I_i = (N_i, T_i)$ where N_i is the surface normal and T_i is the face texture. Let $H(E_0, E_1, \dots, E_m)$ be the space of all possible convex combinations of these examples.

$$H(E_0, E_1, \dots, E_m) = \left\{ \left(\sum_{i=0}^m w_i G_i, \sum_{i=0}^m w_i I_i \right) \mid \sum_{i=0}^m w_i = 1, \text{ and } w_0, w_1, \dots, w_m \geq 0 \right\}$$

When we synthesize a new texture, the implicit normal map is synthesized at the same time. Thus, the facial details can be acquired from the synthesized texture in normal map form. We can represent novel expressions as follow:

$$E_{(new)} = G_{(new)} N_{(new)} T_{(new)}$$

$$\text{where } G_{(new)} = \sum_{i=0}^m w_i G_i, T_{(new)} = \sum_{i=0}^m w_i T_i, \text{ and implicit } N_{(new)} = \sum_{i=0}^m w_i N_i$$

As shown above, the blending weight w_i is related to normal, texture, and geometry.

Therefore, we can calculate the weights from one of the three components. In our approach, geometric relation between prototypes is employed to get the blending

weights. Let G_i^S denotes the feature point set of prototype expressions in our data pool (how to capture the 3D positions of feature points will be described in chapter 5)

and G^N denotes the set of new positions of feature points which is given by users.

By projecting G^N into the convex hull of $G_0^S G_1^S \dots G_m^S$, the weights can be found.

Thus, the estimation of blending weights can be written as an optimization problem:

$$\text{minimize} \left\{ \left(G^N - \sum_{i=0}^m w_i G_i^S \right)^T \left(G^N - \sum_{i=0}^m w_i G_i^S \right) \right\},$$

$$\text{subjects to: } \sum_{i=0}^m w_i = 1, w_i \geq 0, \text{ for } i = 0, 1, \dots, m.$$

The objective function of the optimization problem above can be rewritten as:

$$W^T g^T g W - 2G^{N^T} g W + G^{N^T} G^N$$

where $g = (G_0^S, G_1^S, \dots, G_m^S), W = (w_0, w_1, \dots, w_m)$. This optimization is a positive

semi-definite quadratic programming problem with linear constraints. This type of

problem can be solved by various methods such as interior-point method or active set

method. In our approach, we used the active set strategy to solve this optimization

problem. The flow chart is described in Figure 16.

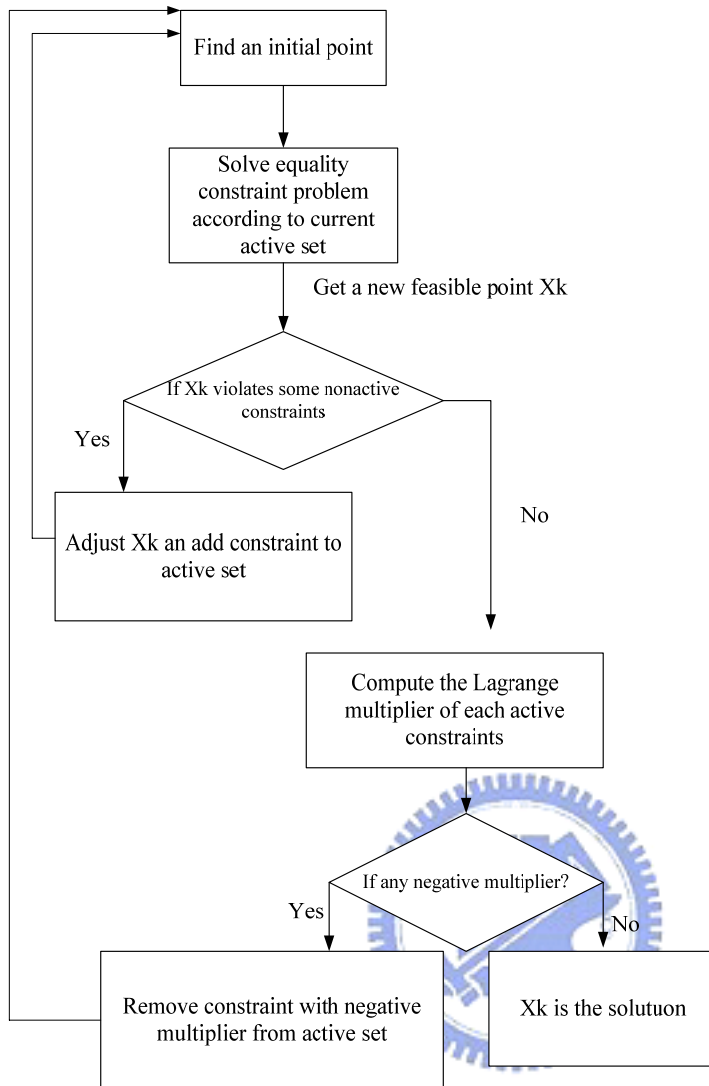


Figure 16: The flow chart of active set strategy.

4.2 Synthesis of Textures and Normals

After calculating the blending weights, we can synthesize the textures or normals by interpolating the prototypes. For convenience of interpolation, we align the pixels of prototypes first. The alignment is performed by an image warping technique which is proposed by Beier et al.[1]. After alignment of pixels, the synthesis can be performed as weighted sum of corresponding pixels in all images.

Since there are only 15~25 prototypes in our database, if simply interpolating the prototypes, the variations of novel expression will be few. Thus, we divide a prototype

image into 8 sub-regions. With the 8 sub-regions, we could synthesize in each sub-region individually. That is to say each sub-region can have different sub-expression. By dividing sub-regions, we can compose more expressions under a smaller data pool.

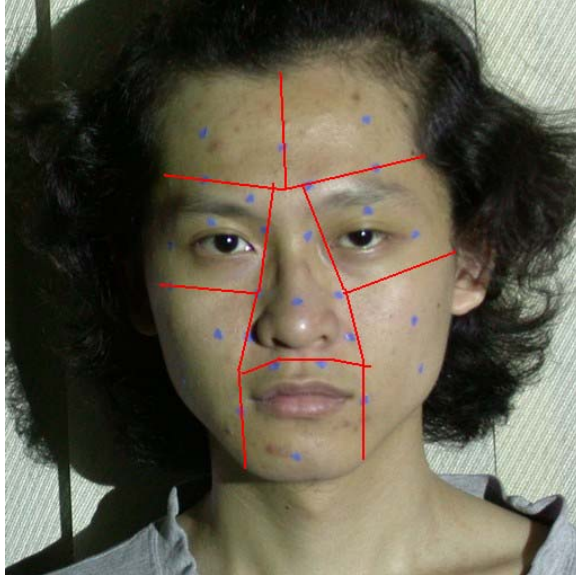


Figure 17: We roughly divide a face into 8 sub-regions to increase the probable novel expression.

To avoid image discontinuities along the sub-region boundaries, we do gradual blending between the sub-region boundaries. At the sub-region boundaries, we do the blending as below:

$$P = w * P^i + (1 - w) * P^j$$

where P represents pixel in boundary, P^i represents the pixel in region i , and P^j represents the pixel in region j . The blending weight w will gradually change from one to zero when moving from region i to region j .

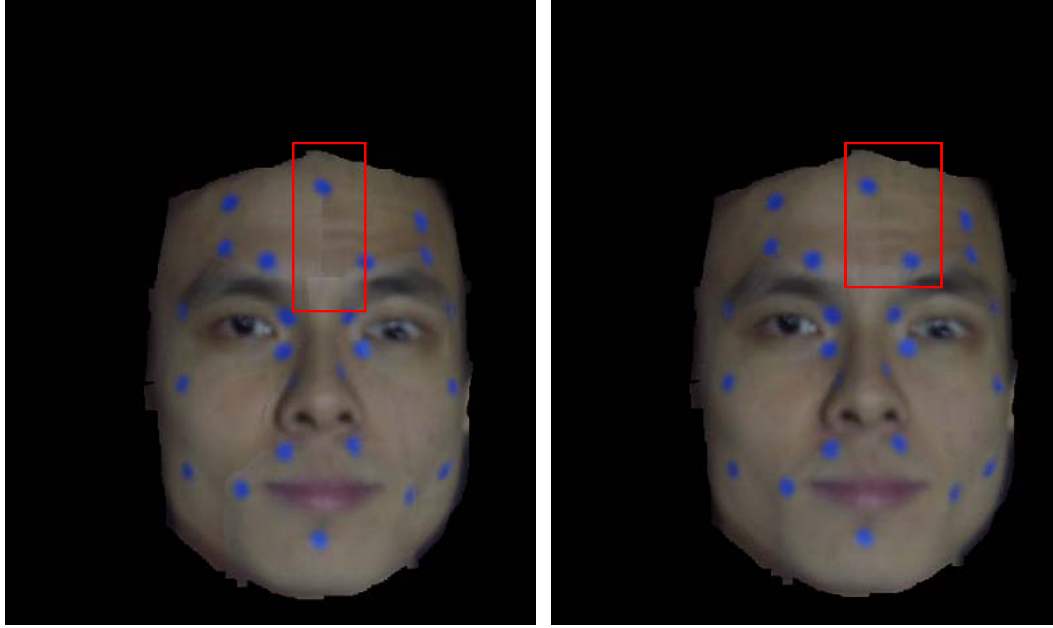


Figure 18: Left, without blending on sub-region boundary, the hard edge between two sub-regions is clear. Right, the hard edge is eliminated by blending on sub-region boundary.

4.3 Editing and Synthesis of 3D geometry

In previous two sections, we describe how to calculate blending weights and synthesize 2D textures. In this sub-section, we describe the synthesis of 3D geometry and how to use the novel geometry to deform the 3D head model.

Since the blending weights is derived from the vector space which consists of the geometries of prototype images, the 3D geometry of novel expression can also be calculated by simply interpolating the geometries of prototype images. That is:

$$G_{(new)} = \sum_{i=0}^m w_i G_i^S \quad \text{where } w_i \text{ and } G \text{ is defined same as section 4.1.}$$

Once users have input a new configuration of feature points, our system will project the configuration to expression space formed by prototypes. After we get the 3D geometry of novel expression, we can use the geometry to deform the 3D generic head model to produce the novel 3D expression.

First, users have to select a set of corresponding pairs $\{\mathbf{p}_i, \mathbf{q}_i\}$, where \mathbf{p}_i is the feature point position of our synthesizing expression and \mathbf{q}_i is the corresponding point position on the generic model. Once the displacement of each feature point position $\mathbf{u}_i = \mathbf{q}_i - \mathbf{p}_i$ was calculated, we use scattered data interpolation $S(\mathbf{p})$ to estimate the displacement of other vertices on the original mesh. There are plenty of methods proposed to construct the interpolation function. We adopted the radial basis function which form is:

$$S(\mathbf{p}) = \sum_i c_i \phi(\|\mathbf{p} - \mathbf{p}_i\|) + M(\|\mathbf{p} - \mathbf{p}_i\|) + t$$

where ϕ is radial symmetric basis function, and c_i are displacement coefficients, and M, t are affine terms. To determine c_i, M and t , we solve a set of linear equations that includes interpolation constraints $\mathbf{u}_i = S(\mathbf{p}_i)$. Many different basis functions ϕ have been proposed. We use: $\phi(r) = e^{-r/64}$.

Once we construct the scattered interpolation function, the displacement of all vertices can be evaluated. In our system, we use 20~30 corresponding points between synthesizing expressions and the generic model.

4.3.1 Editing Interface

Figure 19 shows our editing interface. Our editing interface has 20~30 control points according to our prototypes. These control points can be selected with mouse and modified by keyboard. We use the 3D geometry reconstructed from our prototype image as constraints. Each control point can only be moved in the range of constraint.

Once users have input a new configuration of feature points, our system will project the configuration to expression space formed by prototypes. Then, the technique will be used to deform the 3D head model.

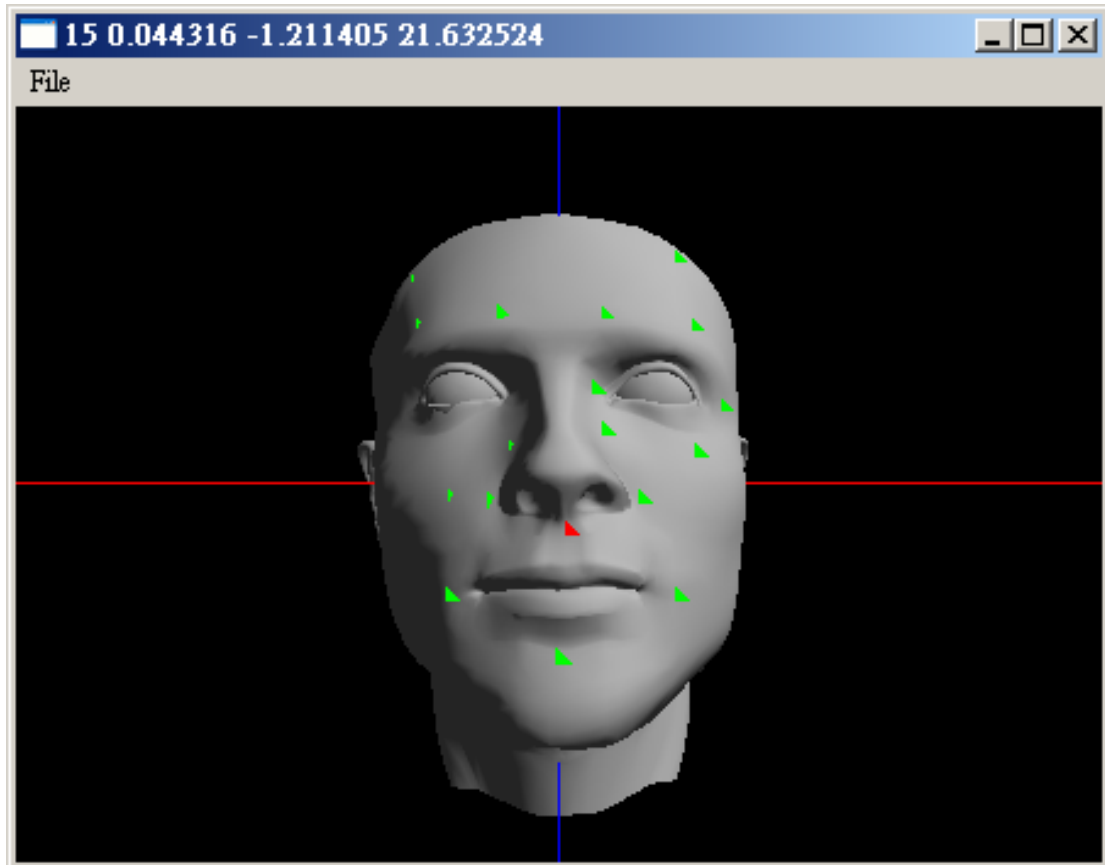


Figure 19: The 3D head model with control points. The green triangles present the control points. User can modify its position by a keyboard and mouse.

5. Experiment and Results

In this chapter, we will describe our experiment and show our results. At the beginning, we introduce the collection and pre-processing of our prototype images. Then, we will show our editing system and the final results.

5.1 Collection and Pre-Processing of Prototype

Images

In our system, we use 15~25 prototype images for synthesizing 2D novel expressions. In order to acquire feature details from our synthesized images, our prototype images were taken under an illumination-controlled environment.

In the environment, a projector is used as the only one light source. Three cameras are used to take pictures from three different views of our model. Figure 20 shows a set of our prototype images.

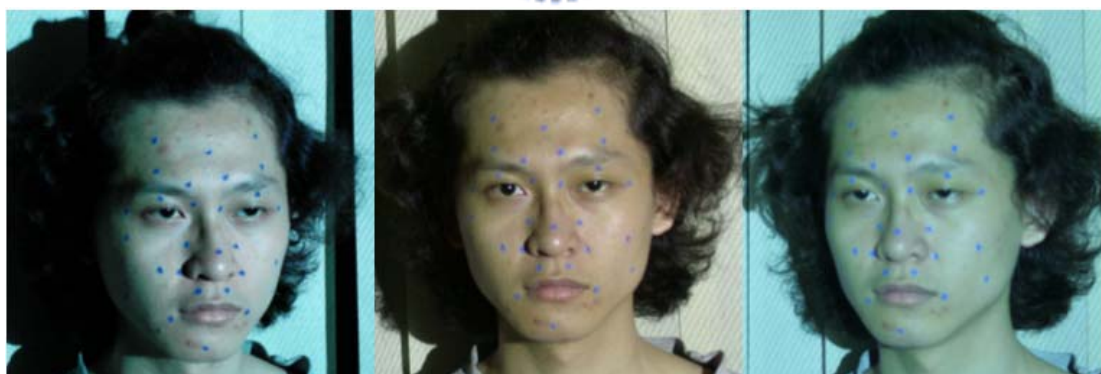


Figure 20: A set of prototype image consist of 3 different views. Left and right ones are used to recover the 3D geometry of feature points. Center view is use to acquire normal maps.

As shown in Figure 21, we put a set of (usually 20~30 markers, in the figure 23) markers on the actor's face. These markers are pasted on the position whose variations are representative under different expressions. We will use the 3D positions

of these markers to synthesize novel 3D expression.

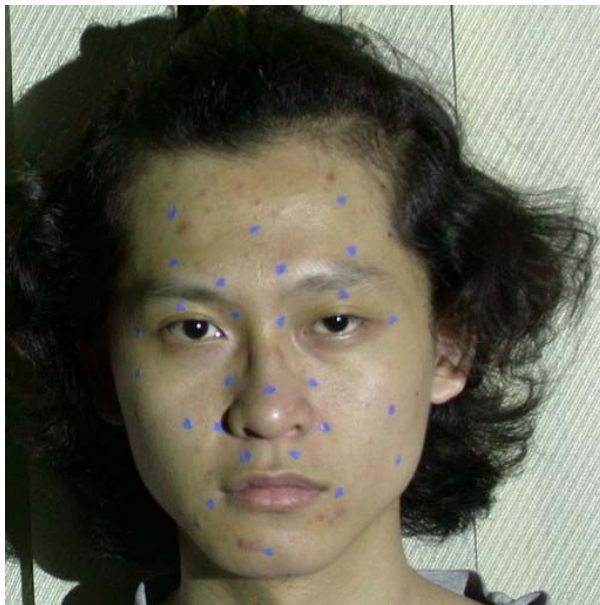


Figure 21: The neutral face we choose. The blue marker is put on several specific position to observe the variation of expressions.

5.2 Results

The 3D head model used in the thesis has 6078 vertices and 6315 polygons. Every vertex has a predefined normal vector.

Figure 23~30 are sets of our product. These expressions can be edited simply by dragging feature point positions. Then utilizing the technique described in chapter 4, approximate 3D geometries and normal difference maps were synthesized. Figure 31~33 are three sets of prototypes. Because there are no conspicuous wrinkles on the forehead in the prototype sets shown in Figure 32 and 33, the corresponding results (shown in Figure 29 and Figure 30) also have no forehead wrinkles.

Our current results still have some defects. The first one is that synthesized facial details sometimes become thinner than we expected. That is due to errors such as pixel alignment during interpolation procedure. The second one is our system might produce some expressions which a real person may not express. That is due to our constraints between two sub-region are weak. Therefore, some strange combination may be synthesized if a user is not careful.

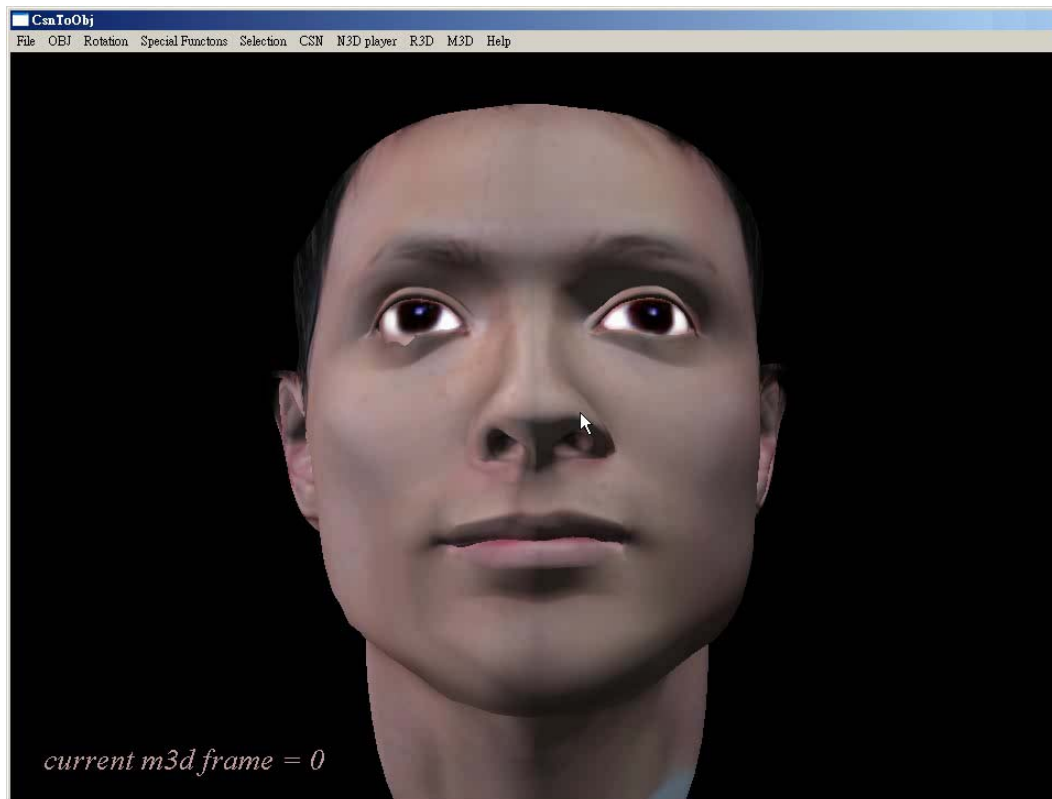


Figure 22: Our 3D head model with neutral face.

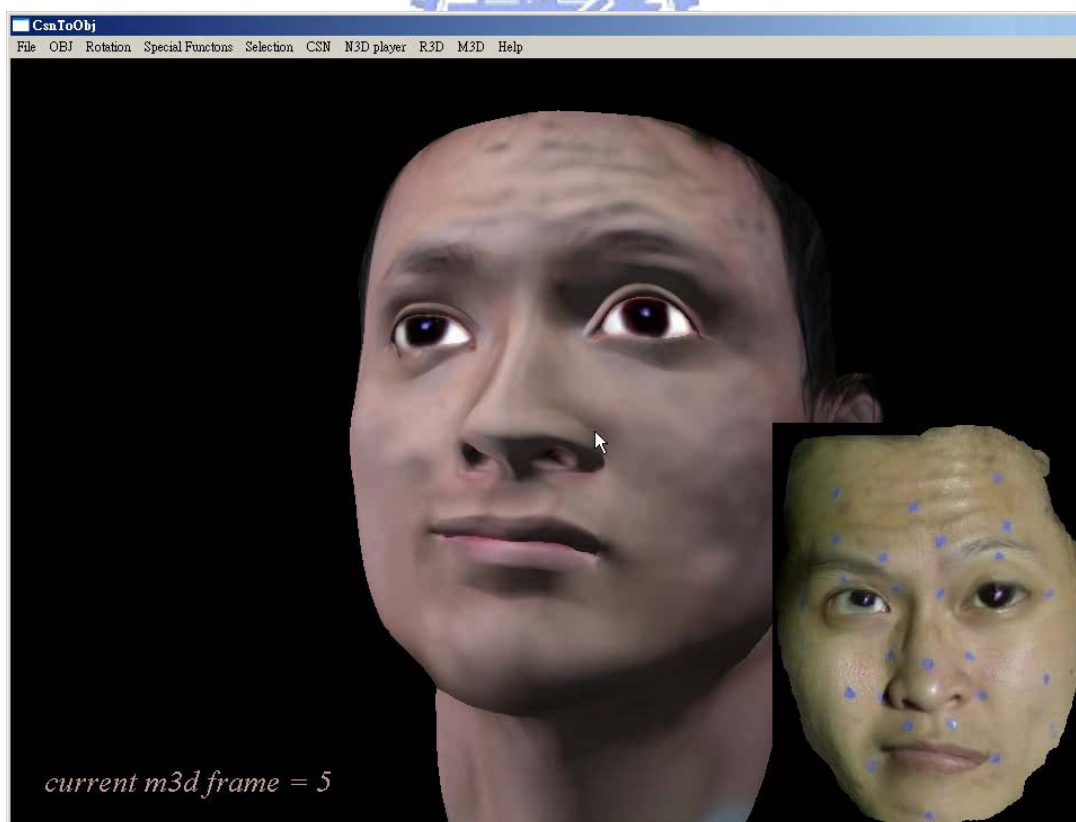


Figure 23: The synthesized head model. The picture at right corner is corresponding texture.

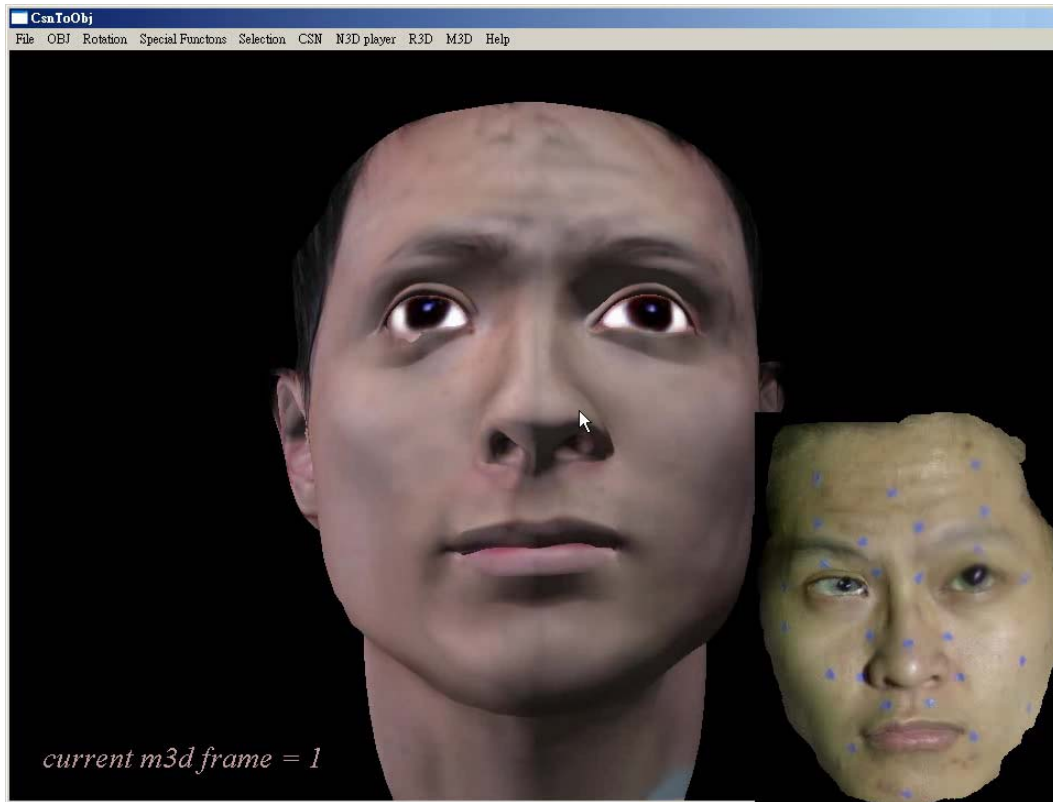


Figure 24: Another synthesized head model. The picture at right corner is corresponding texture.

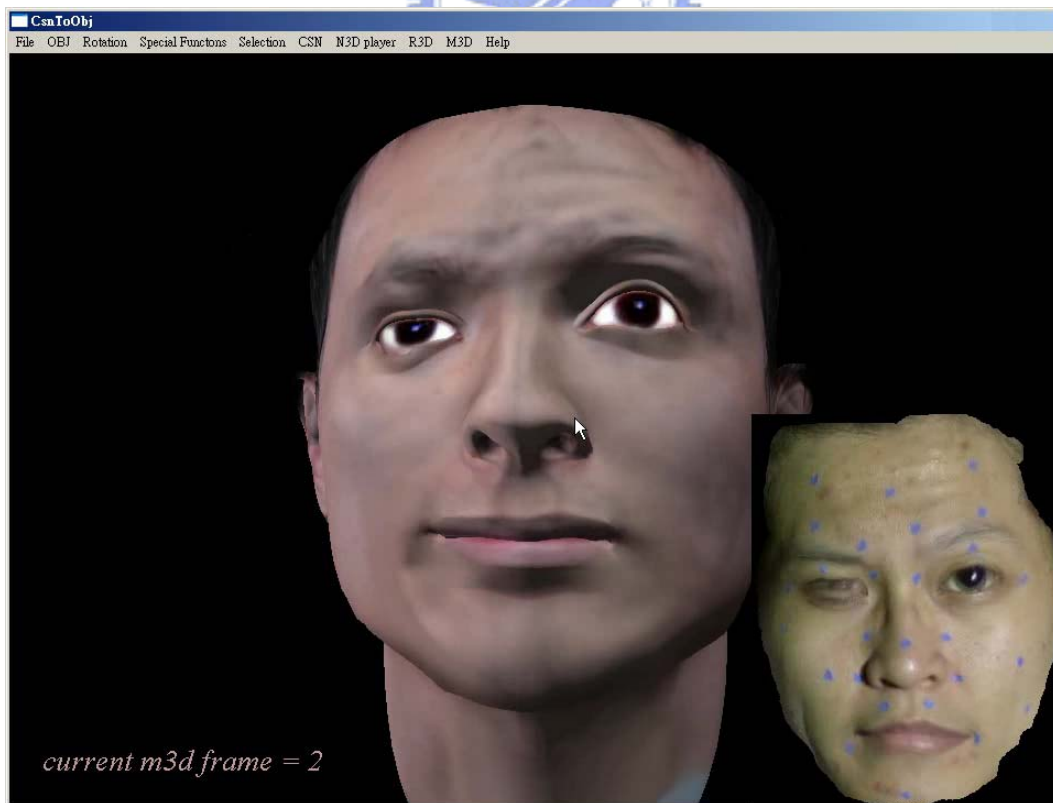


Figure 25: The synthesized head model. The picture at right corner is corresponding texture.

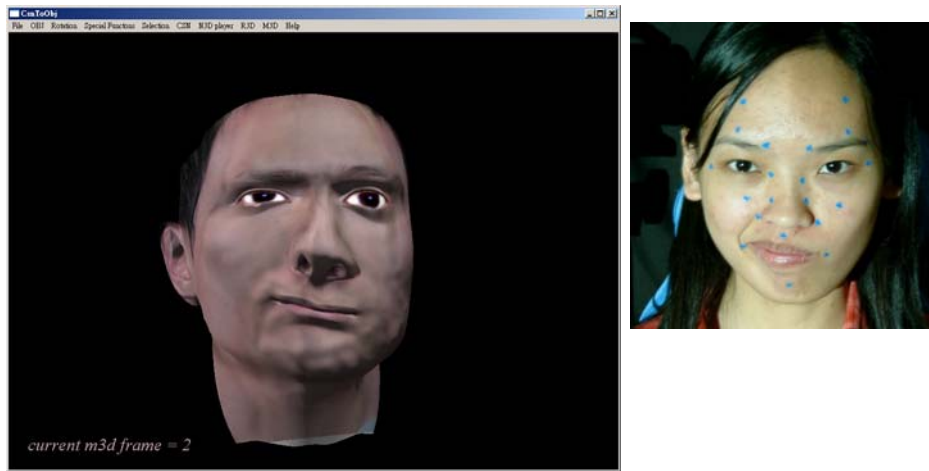


Figure 26: Synthesize a normal difference map of female to a male model.

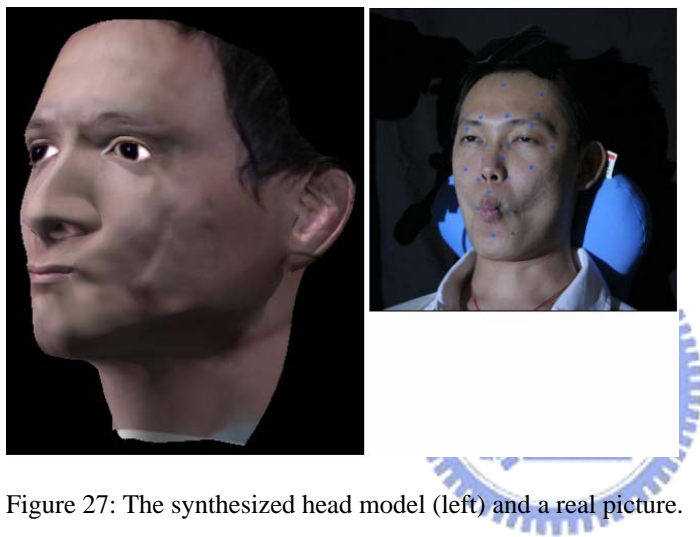


Figure 27: The synthesized head model (left) and a real picture.

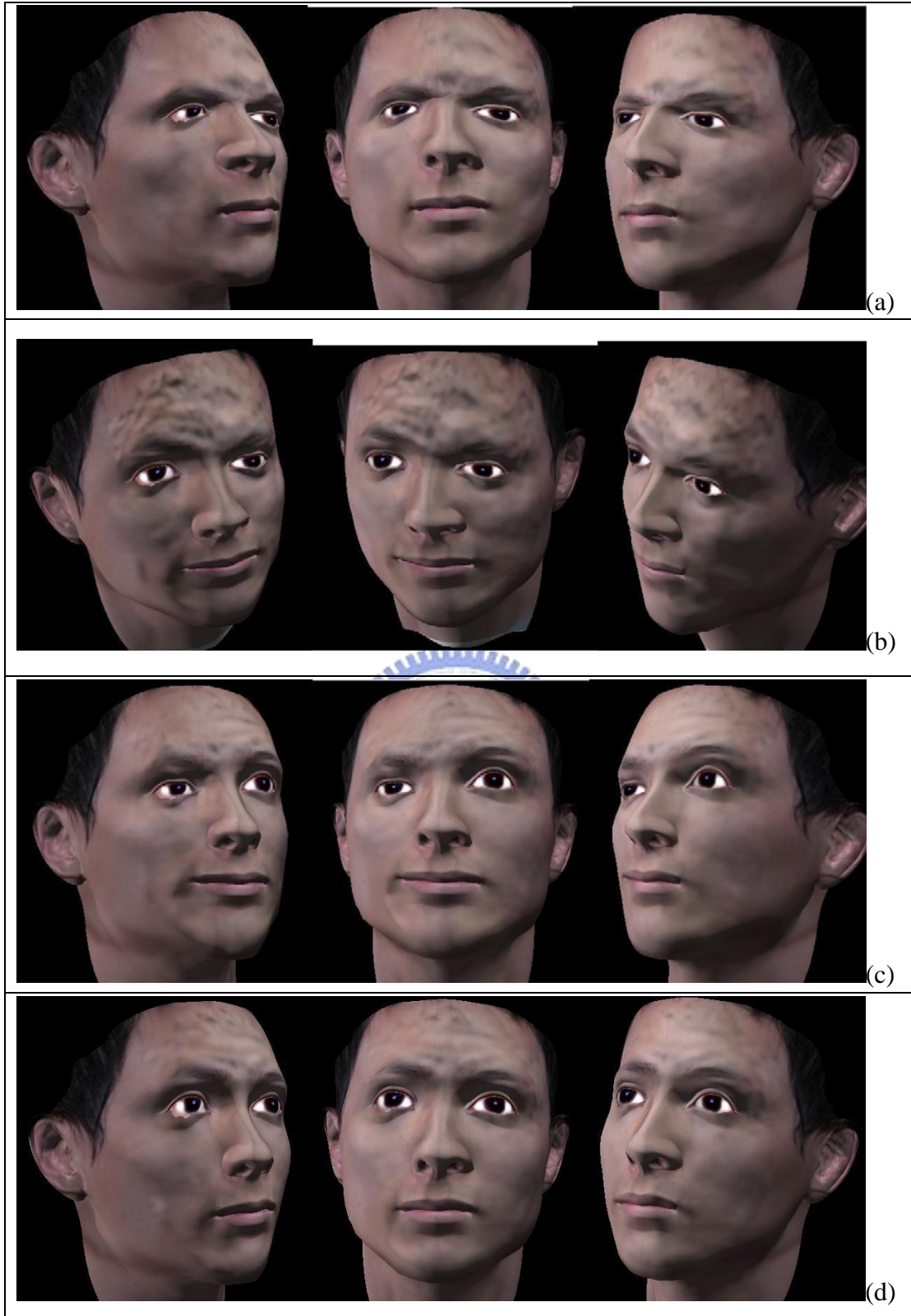


Figure 28: Figure (a),(b),(c),and (d) are four sets of synthesized expressions. The forehead wrinkles in (a), (c) and (d) are thinner because error during interpolation procedure. If we applied Gaussian filter (described in chapter 3.4) less times, there are more noises in the synthesis result (such as (b)).

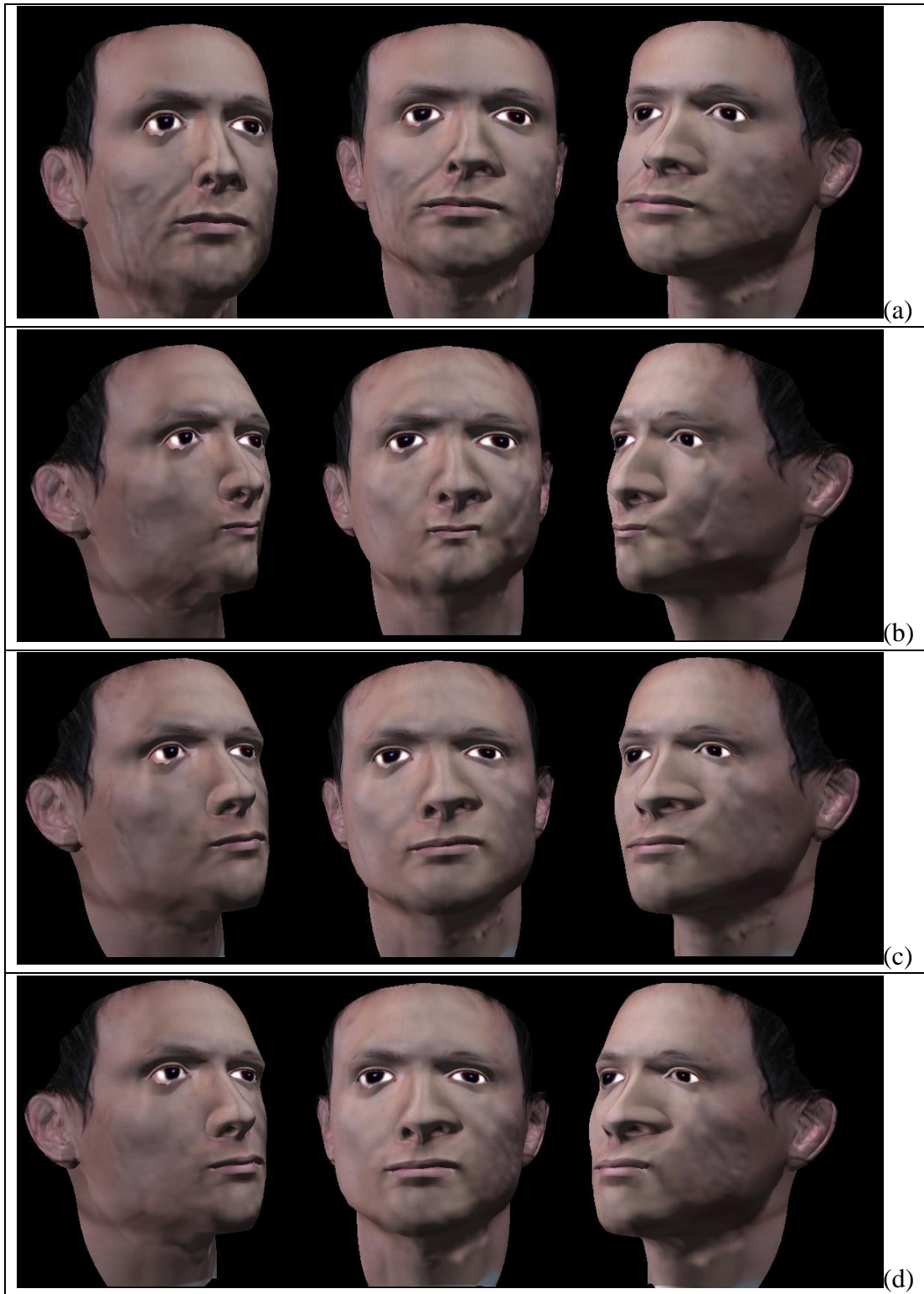


Figure 29: Figure (a),(b),(c),and (d) are four sets of synthesized expressions.

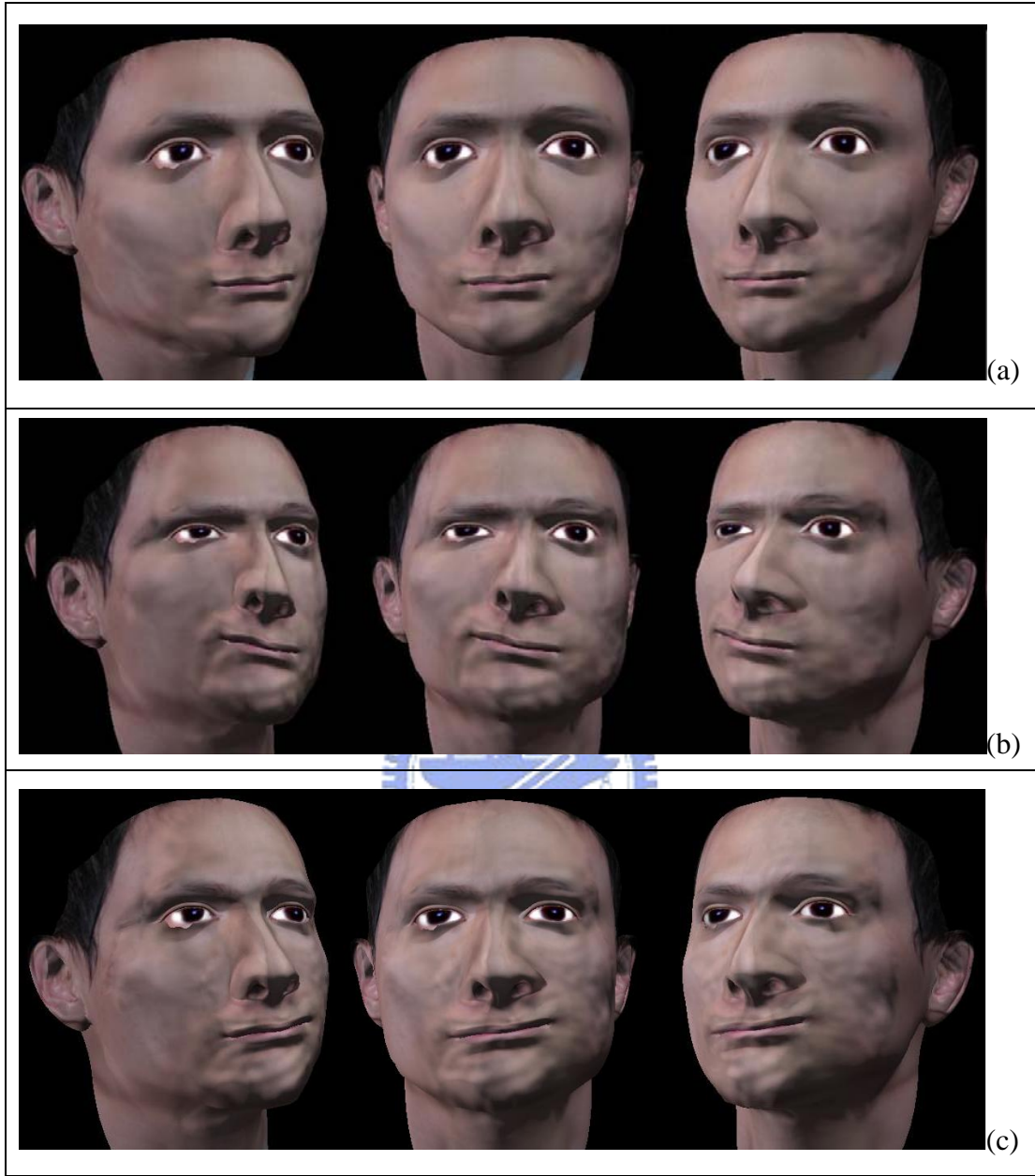


Figure 30: Figure (a),(b),and (c) are three sets of synthesized expressions.



Figure 31: The prototypes which we used to synthesize fig28 and fig22~25.



Figure 32: The prototypes which we used to synthesize fig29 and fig27.



Figure 33: The prototypes which we used to synthesize fig30 and fig26.

6. Conclusion and Future Work

6.1 Conclusion

In this thesis, a 3D detailed facial expression synthesis system is presented. The system consists of three functions — 3D expression editing, 2D texture synthesizing, and acquisition of normal difference maps. By manipulating the positions of feature points, users could modify the expression of a 3D head model. Then, by deriving the relationship between modified geometry and geometry of our prototypes, the corresponding normal difference maps could be synthesized. We demonstrated that a corresponding normal difference map could be evaluated acquired by a low cost method. We also showed that the synthesized normal difference map could be used to generate conjectured facial details on the modified 3D head model.

Our contributions include: 1) an inexpensive method to acquire and synthesize facial details, and 2) a framework for editing and synthesizing 3D detailed expression.

6.2 Future Work

The current system has some drawbacks that can be improved. First, our normal difference map is acquired by a simplified method. We will use more delicate methods (such as [13], [24]) for a more accurate result. Second, the positions of feature points are decided empirically. Further analysis could increase the fidelity. Third, the expression produced could not yet be animated smoothly. Solving the temporal coherence could produce smoother animation.

7. Reference

- [1] Beier, T., and Neely, S. "Feature-based Image Metamorphosis", Proceedings of the 19th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH), Vol. 26, Pages 35-42, 1992.
- [2] Blanz, V., Basso, C., Poggio, T., and Vetter, T. "Reanimating Faces in Images and Video", Computer Graphics Forum 22 (3), EUROGRAPHICS 2003, Pages 641 - 650, 2003.
- [3] Blanz, V., and Vetter, T. "A Morphable Model for the Synthesis of 3D Faces", Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH), Pages 187-194, 1999.
- [4] Brand, M. "Morphable 3D models from video", Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Vol. 2, Pages 456-463, 2001.
- [5] Breger, C., Covell, M., and Slaney, M. "Video rewrite: Driven Visual Speech with Audio", Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH), Page 353-360, 1997.
- [6] Deng, Z., Chiang, P-Y., Fox, P., and Neumann, U. "Animating Blendshape Face by Cross-Mapping Motion Capture Data", Proceedings of the 2006 Symposium on Interactive 3D Graphics and Games, Pages 43-48, 2006.
- [7] Ezzat, T., Geiger, G., and Poggio, T. "Trainable Videorealistic Speech Animation", Proceedings of the 29th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH), Pages 388-398, 2002.
- [8] Fang, H., and Hart, J. C. "Textureshop: Texture Synthesis as a Photograph Editing Tool", ACM Transactions on Graphics (TOG), Volume 23, Issue 3 (August 2004), Special Issue: Proceedings of the 2004 SIGGRAPH Conference., Pages 354-359, 2004.
- [9] Fletcher, R. "Practical Methods of Optimization", Vol. 1, John Wiley & Sons, 1980.
- [10] Forsyth, Ponce "Computer Vision: a Modern Approach", Prentice Hall, 2003.
- [11] Golovinskiy, A., Matusik, W., and Pfister, H. "A Statistical model for Synthesis of Detailed Facial Geometry", ACM Transactions on Graphics (TOG), Volume 25, Issue 3 (July 2006), SESSION: Appearance modeling, Pages: 1025-1034, 2006.
- [12] Guenter, B. K., Grimm, C., Wood, D., Malvar, H. S., and Pighin, F. H., "Making Faces", Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH), Pages 55-66, 1998.
- [13] Horn, B.K. 1990. "Height and Gradient from Shading", International Journal of Computer Vision, Vol. 5(1), Pages 37-75, 1990.
- [14] Lin, I.-C., and Ouhyoung, M. "Mirror Mocap: Automatic and Efficient Capture of Dense 3D Facial Motion Parameters from Video", The Visual Computer, Vol. 12(6), Pages 355-372, 2005.

- [15] Lin, I.-C., Yeh, J.-S., and Ouhyoung, M. "Extracting 3D Facial Animation Parameters from Multiview Video Clips", IEEE Computer Graphics and Applications (special issue on tracking), Vol. 22(6), Pages 72-80, 2002.
- [16] Liu, Z., Shan, Y., and Zhang, Z. "Expressive Expression Mapping with Ratio Images", Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques, Pages 271-276, 2001.
- [17] Nielson, G. M. "Scattered Data Modeling" IEEE Computer Graphics and Applications, Vol.13, Issue 1, Pages 60-70, 1993.
- [18] Noh, J.-Y., and Neumann, U. "Expression Cloning", Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH), Pages 277-288, 2001.
- [19] Pighin, F., Hecker, J., Lischinski, D., Szeliski, R., and Salesin, D. H. "Synthesizing Realistic Facial Expressions from Photographs", Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH), Pages 75-84, 1998.
- [20] Seitz, S. M., and Dyer, C. R. "View Morphing". Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH), Pages 21-30, 1996.
- [21] Sifakis, E., Neverov, I. and Fedkiw, R., "Automatic Determination of Facial Muscle Activations from Sparse Motion Capture Marker Data", ACM Transactions on Graphics (TOG), Volume 24, Issue 3 (July 2005), Proceedings of ACM SIGGRAPH 2005, Pages 417-425, 2005.
- [22] Tu, P.-H., Lin, I.-C., Yeh, J.-S., Liang, R.-H., and Ouhyoung, M. "Surface Detail Capturing for Realistic Facial Animation", Journal of Computer Science and Technology, Special Issue on Advances in Computer Graphics, Vol. 19(5), Pages 618-625, 2004.
- [23] Waters, K. "A Muscle Model for Animating Three-Dimensional Facial Expression", Proceedings of the 14th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH), Pages 17-24, 1987.
- [24] Wenger, A., Gardner, A., Tchou, C., Unger, J., Hawkins, T., and Debevec, P. "Performance Relighting and Reflectance Transformation with Time-Multiplexed Illumination", ACM Transactions on Graphics (TOG), Volume 24, Issue 3 (July 2005), Proceedings of ACM SIGGRAPH 2005, Pages 756-764, 2005.
- [25] Zhang, L., Snavely, N., Curless, B., and Seitz, S. M. "Spacetime Faces: High-Resolution Capture for Modeling and Animation", ACM Transactions on Graphics (TOG), Volume 23, Issue 3 (August 2004), Special Issue: Proceedings of the 2004 SIGGRAPH Conference, Pages 548-558, 2004.
- [26] Zhang, Q., Liu, Z., Guo, B., and Shum, H. "Geometry-Driven Photorealistic Facial Expression Synthesis", Proceedings of the 2003 ACM SIGGRAPH/Eurographics Symposium on Computer Animation, Pages 177-186, 2003.
- [27] Zhang, Q., Liu, Z., Guo, B., Terzopoulos, D., and Shum, H. "Geometry-Driven Photorealistic Facial Expression Synthesis", IEEE Trans. On Visualization and Computer Graphics, Vol. 12(1), Pages 48-60, 2006.

- [28] Zhang, R., Tsai, P.-S., Cryer, J., and Shah, M. "Shape from Shading: A Survey", IEEE Trans. On Pattern Analysis and Machine Intelligence, Vol. 21(8), Pages 690-706, 1999.

