3D FACIAL EXPRESSION TRACKING AND REGENERATION FROM SINGLE CAMERA IMAGE BASED ON MUSCLE CONSTRAINT FACE MODEL

Shigeo MORISHIMA
Associate Professor, Department of Electrical Engineering and Electronics
SEIKEI UNIVERSITY
3-3-1 Kichijoji-kitamachi, Musashino-shi, Tokyo 180-0001
E-mail: shigeo@ee.seikei.ac.jp
JAPAN

Commission V, Working Group SIG

KEY WORDS: Expression Synthesis, Muscle Constraint Face Model, 3D Parameter Estimation, Neural Network, Optical Flow

ABSTRACT

Muscle based face image synthesis is one of the most realistic approach to generate facial expression in computer graphics or to realize life-like agent for human computer interaction. Our facial muscle model is composed of facial tissue elements and muscles. In this model, forces are calculated effecting facial tissue element by contraction of each muscle strength, so the combination of each muscle parameter can decide a specific facial expression. Now each muscle parameter is decided on trial and error procedure comparing the sample photograph and generated image using our Muscle-Editor to synthesize a specific face image. In this paper, we propose the strategy of automatic estimation of facial muscle parameters from single camera image using neural network.

A neural network which has learned several patterns of facial expressions can convert landmarks or optical flow information into muscle parameters. This neural network can realize an inverse mapping of the image synthesis process from the 3D muscle contraction to the 2D point movement in the display. So this is also 3D motion estimation from 2D point or flow information in captured image under restriction of physics based face model. The facial image is then re-generated from the facial muscle model to estimate the difference from the original image both subjectively and objectively. We also tried to generate animation using the captured data from the image sequence. As a result, we can get and synthesize images which give an impression close to the original.

1. INTRODUCTION

Recently, research into creating friendly human interfaces has flourished remarkably. Such interfaces smooth communication between a computer and a human. One style is to have a virtual human[Badler 1993][Tahlmann 1995] appearing on the computer terminal who can understand and express not only linguistic information but also non-verbal infomation. This is similar to human-to-human communication with a face-to-face style and is sometimes called a Life-like Communication Agent[Morishima 1996]. In the human-human communication system, facial expression is the essential means of transmitting non-verbal information and promoting friendliness between the participants. We have already developed a facial muscle model [Sera 1996] as a method of synthesizing realistic facial animation.

The facial muscle model is composed of facial tissue elements and muscle strings. In this model, forces effecting each facial tissue are calculated by contraction of each muscle. So a combination of muscle strengths decides a specific facial image. Currently, however, we have to manually determine each muscle parameter by trial and error, comparing the synthesized image to a photograph.This paper proposes two methods of automatic estimation of facial muscle parameters from 2D information, i.e., marker movements and optical flow in a frontal face image. In the first approach, small colored circle makers are attached to

a subject's face to measure the quantity of transformation of the face when an expression appears. Then we can find out the difference between any specific expression and a neutral expression.

A neural network which has learned several patterns of facial expressions can convert the marker movements into muscle parameters. This neural network can realized an inverse mapping of the image synthesis process from the 3D muscle contraction to the 2D point movement in the display. So this is also 3D motion estimation from 2D point tracking in captured image under restriction of physics based face model. The facial image is then re-synthesized from the facial muscle model to estimate the difference from the original image both subjectively and objectively. We also tried to generate animation using the captured data from the image sequence. As a result, we can get and synthesize images which give an impression close to the original.

2. FACIAL MUSCLE MODEL

2-1 Layered Dynamic Tissue Model

The human skull is covered by deformable tissue which has five distinct layers. Four layers (epidermis, dermis, subcutaneous connective tissue, and fascia) comprise the skin, and the fifth layer comprises the muscles of facial expression. In accordance with the structure of real skin, we employ a synthetic tissue model constructed from the elements illustrated in Figure 1, consisting of nodes interconnected by deformable springs (the lines in the figure). The epidermal surface is defined by nodes 1, 2, and 3, which are connected by epidermal springs. The epidermal nodes are also connected by dermal-fatty layer springs to nodes 4, 5, and 6, which define the fascia surface.

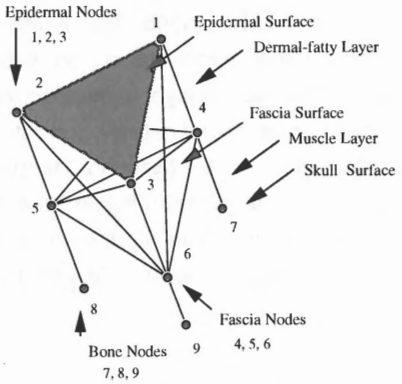


Fig. 1. Triangle Skin Tissue Prism Element

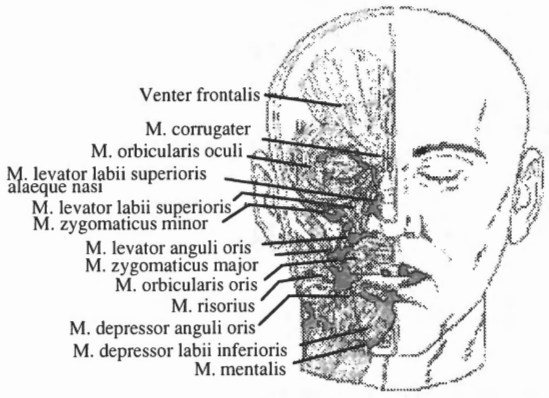
Fascia nodes are interconnected by fascia springs. They are also connected by muscle layer springs to skull surface nodes 7, 8, 9.The facial tissue model is implemented as a collection of node and spring data structures. The node data structure includes variables to represent the nodal mass , position , velocity , acceleration , and net force . The spring data structure comprises the spring stiffness , the natural length of the spring , and pointers to the data structures of the two nodes that are interconnected by the spring. Newton's laws of motion govern the response of the tissue model to force [Lee 1995][Terzopoulos 1993]. This leads to a system of coupled, second-order ordinary differential equations that relate the node positions, velocities, and accelerations to the nodal forces. The equation for a generic node i is as follows:

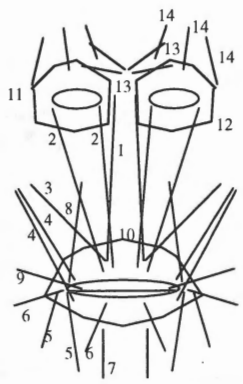
$$m_i \frac{d^2\mathbf{x}_i}{dt^2} + \gamma_i \frac{d\mathbf{x}_i}{dt} + \tilde{\mathbf{g}}_i + \tilde{\mathbf{q}}_i + \tilde{\mathbf{s}}_i + \tilde{\mathbf{h}}_i = \tilde{\mathbf{f}}_i$$

m_i is the nodal mass at node i
γ_i is the damping coefficient
$\tilde{\mathbf{g}}_i$ is the total spring force at node i
$\tilde{\mathbf{q}}_i$ is the total volume preservation force at node i
$\tilde{\mathbf{s}}_i$ is the total skull penetration force at node i
$\tilde{\mathbf{h}}_i$ is the total nodal restoration force at node i
$\tilde{\mathbf{f}}_i$ is the total applied muscle force at node i

2.2 Modified facial muscle model

The facial muscle model had two kinds of muscle models at forehead, namely, Frontaris and Corrugater. Frontaris pulls





1. M. levator labii superioris alaeque nasi
2. M. levator labii superioris
3. M. zygomaticus minor
4. M. zygomaticus major
5. M. depressor anguli oris
6. M. depressor labii inferioris
7. M. mentalis
8. M. risorius
9. M. levator anguli oris
10. M. orbicularis oris
11. M. upper orbicularis oculi
12. M. lower orbicularis oculi
13. M. corrugater
14. M. frontalis

Fig. 2. Simulated Muscle Structure [Sincher 1928]

up the eyebrows and makes wrinkles in the forehead. Corrugater pulls the eyebrow together and makes wrinkles between left and right eyebrows. But those muscles can't pull down the eyebrows and make the eyes thin. So a new "Orbicularis oculi" muscle model is appended[4]. According to the anatomical chart shown in Fig. 2, the Orbicularis oculi is separated into an inside part and an outside part. The inside part makes the eye close softly and the outside part makes eye close firmly. To make muscle control simple around the eye area, the Orbicularis oculi is modeled with a single function in our model. Normally, muscles are located between a born node and a fascia node. But the Orbicularis oculi has an irregular style, whereby it is attached between fascia nodes in a ring configuration; it has 8 linear muscles which approximate a ring muscle. Contraction of the ring muscle makes the eye thin. The final facial muscle model has 12 muscles in the forehead area and 27 muscles in the mouth area.

3. MUSCLE PARAMETERS

3-1 Feature Points

A marker is attached on each feature point of the subject's face to measure and model facial expression. A feature point is chosen for each muscle, from the grid point in face model which gives the biggest movement when contracting the muscle. If the feature point has already been chosen by another muscle, the point which gives second biggest movement is chosen as the feature point. Some feature points are appended and modified manually to make each feature point move more independently. We defined 16 feature points in the forehead area and 26 feature points in the mouth area as shown in Fig. 3.

3-2 Face Area Division

A simpler neural network structure can help speed the convergence process in learning and reduce the calculation cost in mapping, so the face area is divided into the three sub-areas indicated in Fig. 4. They are the mouth area, left-forehead area, and right-forehead area, which each give independent skin motion. Three independent networks are prepared for these three areas.

3-3 Neural Network Structure

A layered neural network finds a mapping from feature point movements to muscle parameters. A four-layer structure is chosen to effectively model the non-linear performance. The first layer is the input layer, which corresponds to 2D marker movement.

The second and third layers are hidden layers. Units of the second layer have a linear function and those of the third layer have a sigmoid function. The fourth layer is the output layer, corresponding to muscle parameters, and it has linear units.

Linear functions in the I/O layers are introduced to maintain the range of I/O values. Feature point movements have 2 dimensions, so the number of input-layer units is double the number of feature points. The number of output-layer units is the number of muscles in each sub-area. The number of units in the hidden layer is decided heuristically. For 2 forehead sub-area, neural network consists of 16 units in input layer, 20 units in hidden (Second, Third) layer and 8 units in output layer. For the mouth sub-area, the neural network consists of 52 units in the input layer, 60 units in the hidden (Second, Third) layer, and 28 (27 muscles + a

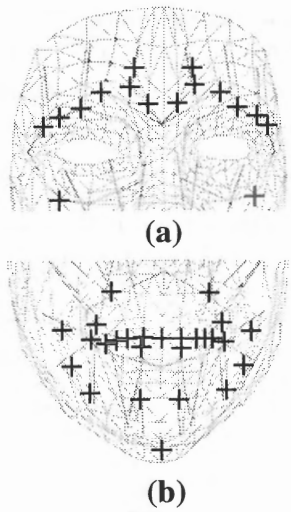


Fig. 3. Feature Points Location
(a) Forehead (b) Mouth

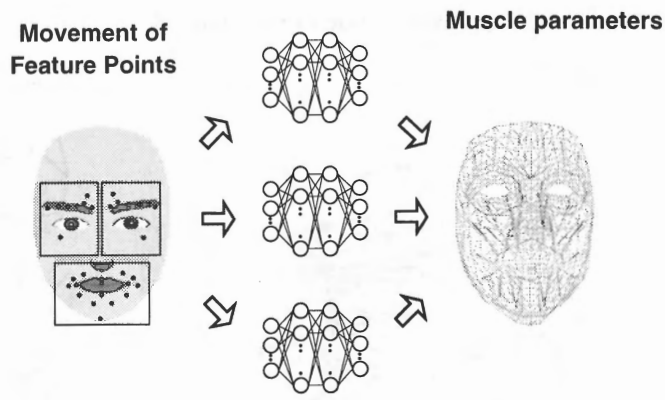


Fig. 4. Estimation of Facial Muscle Parameters by 3 Neural Network

parameter for jaw rotation) units in the output layer.

3-4 Learning Patterns

Learning patterns are composed of the individual contraction of each muscle and their combination. In the case of individual motion, contraction of each muscle between maximum strength and neutral is quantized into 11 steps. In the combination case, we create 6 basic facial expressions consisting of anger, disgust, fear, happiness, sadness and surprise, and quantize the difference between neutral and each of these also into 11 steps. The number of learning patterns in the individual muscle case is 77 (7 muscles x 11 steps) for each forehead sub-area, and that in the combination case is 66 (6 expressions x 11 steps). So the total number of learning patterns is 143 in each forehead sub-area. In the mouth area, each muscle contraction does not happen individually. So all learning patterns are composed of combinations. Learning patterns have basic mouth shapes for vowels "a", "i", "u", "e" and "o", and a closed mouth shape for nasal consonant "n". Also 6 basic expressions are appended as in the forehead area, and jaw rotation is specially introduced. Thus a total of 13 actions are selected for training, and they are also quantized into 11 steps. The number of learning patterns is 143 for the mouth sub-area. Each pattern is composed of a data pair: muscle parameter vector and feature point movement vector. Neural networks were trained using Back Propagation. The learning pattern was increased gradually according to strength of muscle parameter.

3-5 Normalization

In order to absorb individual variations in human faces, each feature point movement is normalized by a standard length to decide the facial geometrical feature. The forehead area and mouth area each have local axes normalized by the local standard length. In the forehead area, Dot A is inside the left eye. Dot B is inside the right eye. Dot O is halfway between Dots A and B. Thex-axis goes horizontally through Dot O, and the y-axis goes vertically through Dot O. Dot E is positioned at the hair-line, on the y-axis.

The standard lengths for the forehead area are distance AB for the x-axis and height OE for the y-axis. In the mouth area, Dot A is left edge of lip. Dot B is right edge of lip. Dot O is halfway between Dots A and B. The x-axis goes horizontally through Dot O and the y-axis goes vertically through Dot O. Dot E is the top of the nose, on the y-axis. Dots C and D are where the x-axis intersects the edge of face. The standard length for the mouth area is distance CD for the x-axis and height OE for the y-axis.

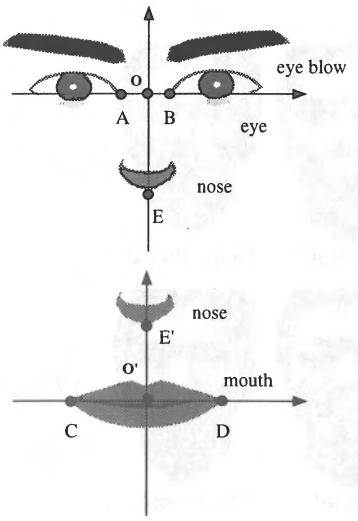


Fig.5 Standard for Normalization

4. EVALUATION

4-1 Closed Test

To confirm whether the learning process of the neural network is successfully completed or not, we input the same data used in learning into the input layer of neural network and resynthesize the facial image using muscle parameters from the output layer of the neural network. Table.1 shows the error between the original image and synthesized image averaged in marker positions for each 6 basic expressions. In all cased, error is negligible and it is appearing the original marker positions can be recovered from estimated muscle parameters.

Example results are shown in Figure. 6. By the evaluation in an impression level, facial features and expressions are almost as same as the original image. So the mapping rules work well for the training data.

4-2 Open Test

We attach markers on a real human's face and get movements of the markers from 2D images captured by a camera when any arbitrary expression is appearing. After

	Difference
	pixels
Anger	0.05
Disgust	0.12
Fear	0.03
Happiness	0.07
Sadness	0.07
Surprise	0.04

Table 1. Estimated Error for Closed Test

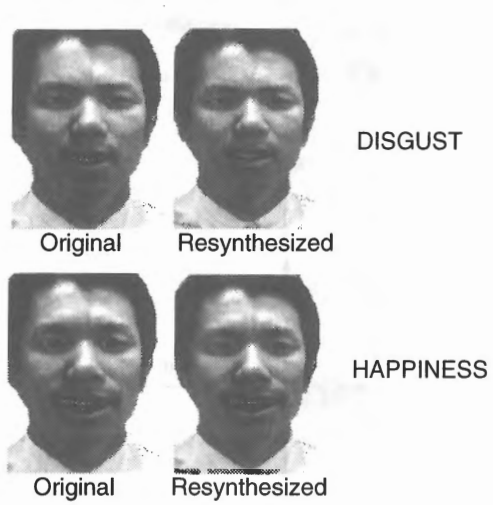


Fig.6 Example of Closed Test Result

normalization, the movement values of the markers are given to the neural network and a facial image is re-generated using the facial muscle model on the parameters from the neural network output. Example images are shown in Fig. 7. Muscle parameters are decided only from the 2D image, but the 3D facial image is well regenerated.

Also, some exceptions occur when a facial feature which cannot be generated by any muscle combination in our model is given as the test sample. In the example image in Fig. 7 "ANGER", the upper and lower lips are being pulled up at the same time. In our current muscle model, the lower lip does not move up beyond the standard position because the muscle action is only limited to contraction. It is necessary to improve the muscle model by including expansion and relocation of muscles to solve this problem.

4-3 Objective Evaluation

Table 2 shows the error beween captured face and synthesized one. 1st column means an averaged error in 42 marker coordinate in the display. Second column is also error in marker location but it cosiders human's sensitivity. At first, 7 people create the face whose impression is very close to the original face using Muscle Editor by trial and error procedure. And these faces are averaged in muscle parameters and true face is defined. Fig.8 shows an averaged face for Fear. Standard deviation(σ) is also calculated for each expression and it means sensitivity of human for expression shift. 2nd column is error in marker position. For example, in case of Anger, error is 5.01σ and σ is 3.41 pixels. 3rd column is error in muscle strength.

These result shows a quantitative standard for evaluation of impression in the synthesized face. Anger is worst one in all cases.

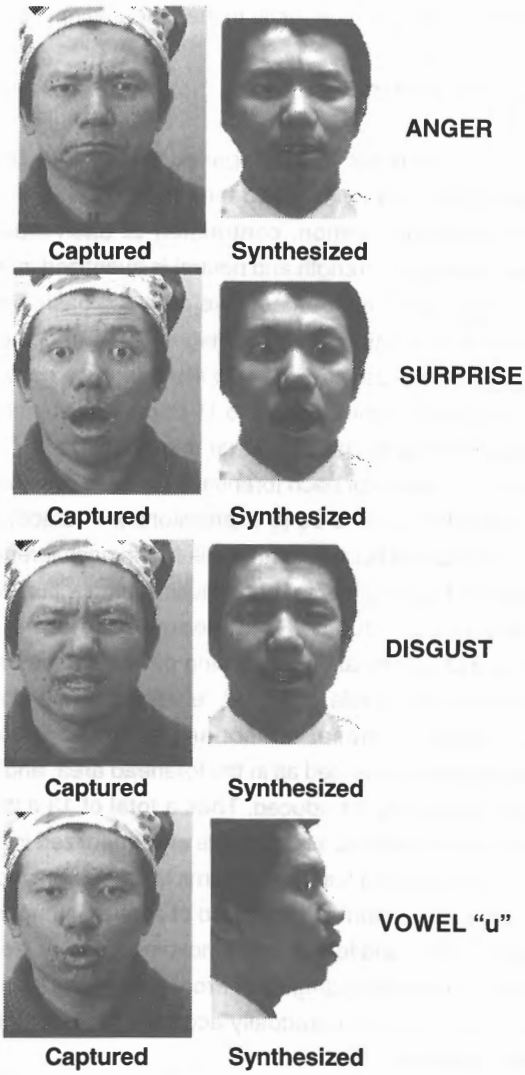


Fig.7 Open Test Result

	Marker Position Error	Marker Position Error	Sigma	Muscle Parameter Error	Sigma
	pixels	σ	pixels	σ	pixels
Anger	9.79	5.01	3.41	2.37	24.6
Disgust	5.25	1.19	5.33	1.73	23.0
Fear	6.46	1.68	5.42	1.38	17.8
Happiness	3.91	0.91	5.29	1.44	21.1
Sadness	6.91	1.75	4.51	1.47	18.5
Surprise	6.48	1.48	6.38	1.59	15.4

Table 2. Estimated Error for Open Test

Fig.8 Averaged Face with 7 Peoples Impression

5. TO OMIT THE MARKER LOCATION

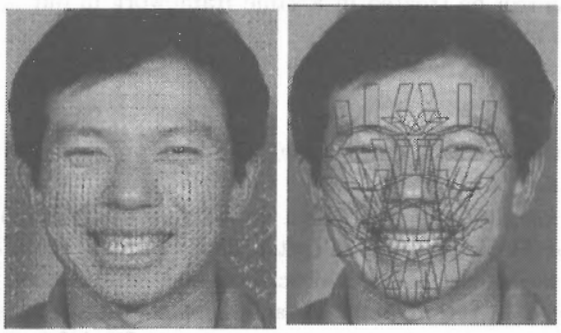
In the next step, the marker has to be omitted to make the motion capture process easy and solve the problem of strong dependance on initially position of markers. Here new method is introduced. This method uses Optical Flow to get measurement of facial expression. Flow is calculated at each flame, and Finally flow is sumed by each flow. And averaged in the mask on the face. Fig.9(a) is result of optical flow calculation and Fig.9(b) show the masks for averaging movement. Facial expression is resymthesized using the facial muscle model on the parameter from the neural network output. This neural network had learned couple of optical flow of original facial image and muscle parameter corresponding with original facial image. As a result, we can get images which are almost same as original images. Fig.10 shows an example of synthesized image. More delicate feature of expression can be captured by optical flow than marker tracking method. This evaluation is future subject.

6. CONCLUSION AND DISCUSSION

A method of automatically estimating 3D facial muscle parameters from 2D marker movements is presented in this paper. Parameter conversion from 2D to 3D works well when the model is fitted to the target person's face precisely in the 2D image and the expression variation is within the range of combinations of basic expressions and vowel pronounciations. We currently fit the model to the facial image by manual operation. But it's impossible to decide the location of all grid points precisely in the real facial image. Of course, a marker's movement strongly depends on its initial location and on the target person, and parameter conversion is very sensitive to its effects.

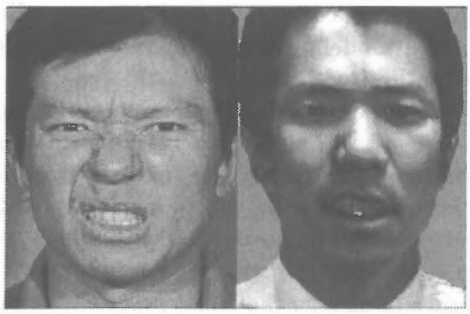
Thus the correspondence between a real face and the model is the next problem to be solved. Now, our facial muscle model requires long computation time. As a result, this metod is not real-time proccessing. Furthermore, from our evaluation test we can see that there are limits to generating any arbitrary expressions with our model, so relocation of the muscles and a new definition of the physics for the new muscles are under examination.

We had also introduced method with optical flow without markers. More delicate facial expression can be captured by optical flow. This evaluation is next subject.



(a) Optical Flow **(b) Masks for Averaging**

Fig.9 Face Feature from Optical Flow



Captured **Synthesized**

Fig.10 Synthesis by Optical Flow

REFERENCE

Badler, I. N., 1993, Simulating Humans: Computer Graphics Animation and Control, Oxford Univ. Press.

Lee, Y., 1995. Realistic Modeling for Facial Animation, Proceedings of SIGGRAPH '95, pp.55-62.

Morishima, S., 1991, A Media Conversion from Speech to Facial Image for Intelligent Man-Machine Interface, IEEE JSAC, Vol.9, No.4, pp.594-600.

Morishima, S., 1996, Life-Like, Believable Communication Agents, SIGGRAPH96 Course Notes #25.

Sera, H., 1996. Physics-based Muscle Model for Mouth Shape Control, Proc. IEEE RO-MAN '96, pp. 207-212.

Sincher, 1928, Anatomie Zahnarzte, Sprimer.

Terzopoulos, D., 1993, Analysis and Synthesis of Facial Image Sequences Using Physical and Anatomical Models, IEEE Trans. on PAMI, vol.15, No.6, pp.569-579.

Thalmann, N., 1995, The simulation of a virtual TV presenter, Computer Graphics and Applications, pp.9-21.

Waters, K., 1995, A Coordinated Muscle Model for Speech Animation, Proc. Graphics Interface '95, pp163-170.