SVR-Based Facial Texture Driving for Realistic Expression Synthesis

Wenhui Zhu[1,2], Yiqiang Chen[2], Yanfeng Sun[1], Baocai Yin[1] and Dalong Jiang[2]
[1]*Multimedia and Intelligent Software Technology Beijing Municipal Key Laboratory,*
Beijing University of Technology, Beijing, China, 100022
[2]*Institute of Computing Technology, CAS, Beijing, China, 100080*
{*whZhu, yqChen, dlJiang*} *@jdl.ac.cn*, {*yfSun, ybc*} *@bjut.edu.cn*

Abstract

The facial texture variation is the key factor for realistic expression synthesis. It always changed with facial motion under some illumination. In this paper, we propose a realistic facial expression texture driving model based on the support vector regression and MPEG-4. It can learn and recall the regression relationship between facial animation parameters and the parameters of expression ratio image through support vector regression method. First, We can get the parameter set of expression ratio image and the eigenERI space by principle component analysis method, who will generate reasonable ratio image. Then, a life-like facial animation can be synthesized quickly and effectively with the support vector regression mapping. In our experiment, it not only captures subtle changes in the variation illumination, but also can synthesis realistic facial expression in bad environment.

1. Introduction

With regard to realistic expression synthesis, dealing with shape and texture, there is one of most interesting yet difficult problems in computer graphics, and it is still a labor-intensive work. For this problem, the key issue is the constrained relationship between facial geometry morphing and expression details changed.

One class of researches about expression is based on the feature points and the 2D transfiguration of texture [1], what is called, expression mapping or performance driven animation. It uses a performer's motion vectors to drive those of others. Thereinto, the new expression texture is generated through geometry-controlled image warping [2, 3]. By all appearances, one short is that does not produce expression details caused by skin deformation such as wrinkles, and only has the 2D transfiguration of texture ignoring the texture depth information (for short ***TDI***), which denote 3D object surface's reflection to illumination.

Another class of researches very adopts implicitly the TDI. The methods using and rebuilding the expression details include the morph-based approaches and their extensions [3, 4] and ERI [5]. The morph-based approaches reasonably combine two or more expressions in a set of example expressions. ERI, however, represents the depth information of 2D texture, namely illumination change. Although both of them generate realistic facial expression details, but they can't be extended better. In our work, a simple ERI's parameterized method is adopted for generating an ERI driving model by the support vector regression and it is a statistic model based on MPEG-4.

The rest of this paper is organized as follow. Some state-of-art researches are mentioned for comparison and our methods are introduced in Section 2. In section 3, we show an overview of the research framework. Section 4 describes the data pre-process. Then, an ERI driving model is described in Section 5. Follow in next, a novel expression synthesis algorithm, based on the ERI driving model, is proposed in Section 6. The results of test and synthesis are shown in Section 7. Finally, we conclude the topic about the paper in the Section 8.

2. Related work

It's been a hot topic about extracting and applying TDI for facial animation. However, it is still an open issue on the constrained relationship between facial shape and its TDI.

The sample-based approach [6, 4] is used to make photorealistic expression with details, adopts stealthily the TDI. In particular, Pighin et. al. [4] used the convex combination of the geometries and textures of the example face models to make realistic facial expression. Later, Zhang [7] implement the facial subdivision and compositing. That makes it more flexible. These methods may work well for particular subject, but hardly to build a general model driving other's expression.

Liu et. al. [5] proposed a technique (about ERI) extracting the TDI of photo to share one person's expression details with others face. Tu [8] and Jiang [9] extended it to 3D facial model. In [8], the gradients of the ratio value at each pixel in ratio images were regarded as changes of a face surface, and a hardware-supported bump mapping was applied to render these detailed normal variations. Moreover, Jiang [9] and Du [10] proposed the ERI-parameterized method. In our work, the ERI's parametric model is generated by PCA.

The basic problem and power of facial animation are the varieties and constrains of facial shape at geometry. It's very important to control texture details variety from sample data. In our work, the FAPs is used to describe the states of facial shape varieties.

In this paper, we present a novel method (facial texture driving, for short ***FTD***) that can generate realistic facial images with different expression details and can dive different people's facial expression model. Our aim is mainly to build an ERI driving model by SVR method. It is worthy of attention that facial texture driving model is a statistical and photographic reality.

3. System overview

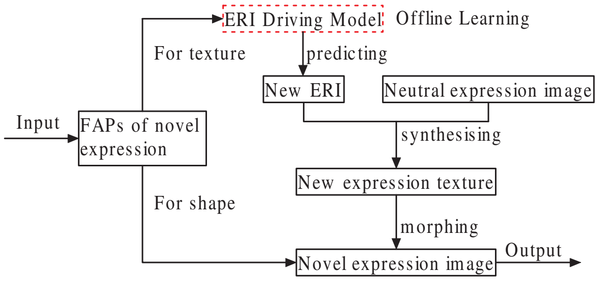

Fig. 1. An overview of the FTD system.

Fig. 1 is an overview of our work, and the details part of offline learning will be showed in Fig. 4. It consists of an offline learning unit and a real time predict and synthesis unit. When it running, the system takes as input the FAPs of a new expression, and synthesizes the final expression image. We have taken a new synthesizing algorithm for decreasing calculation. In our FTD model, there are mainly ERI driving, ERI synthesis and expression texture synthesis. More details will be described in the following sections.

4. Offline processing of the video clips

4.1. Markers and features points

In our trail, six basic expressions types, defined in MPEG-4, are captured by video camera. The key-frames (10 to 20 samples per expression type, as samples trained) and the non-key-frames (one samples per expression type, as samples tested) are extracted with resolution of 200*240 pixels. At the same time, their geometry shape have been hold, too. Fig. 2(a) shows the blue markers for a coarse alignment in the key-frames, and was labeled on the positions hardly without details variety. In Fig. 2(b), there are 92 markers in total, for pinpointing every frame through manually marking its, and that can be done by automatically computing face features [12]. Thereof, the 32 feature points belonging to the FPs of MPEG-4 are adopted. All samples will be done step by step as follow of this section.

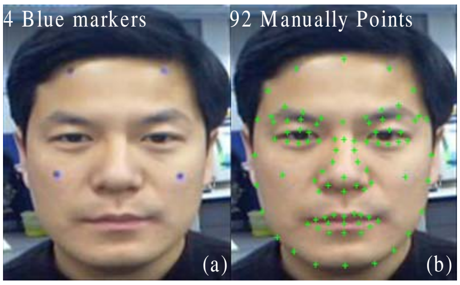

Fig. 2. Markers and Manually points

4.2. Expressions alignment and FAPs extraction

After all key-frames with blue markers are obtained, their scales are adjusted to one of neutral expression frame using our detection tool. Then, all key-frames with manually markers are aligned to the neutral expression frame using a simple triangulation based image warping, although more advanced techniques [3] may be used to obtain better image quality. Finally, we have got all sample images, shown as Fig. 2(b), and their shapes, which is manually markers.

For describing facial geometric data, we easily get the 31 FAPs by extracting feature points of key-frames according to formula (1) in MPEG-4.

$$FAP_i = \frac{FP' - FP}{FAPU} \tag{1}$$

where i=3,...,68; FP and FP' is the coordinate of feature points; FAPU is basic unit and change the real distance to relative scale.

4.3. ERI extraction and ERI-parameterized

For exhibiting or reappearing expression details, in our work, ERI are continuously calculated from all key-frames sequences. The first frame is always used as a neutral face image and is denoted by F_0. The rest of samples trained set are used as expressive face images and are denoted by F_i,

$1 \leq i < n$, where n is the total number of key-frames. Each expressive sample F_i is warped to align with F_0, and all different vectors are expressed by FAPs of key-frames. Then, ERI sequence can be calculated as:

$$R_i(u,v) = \frac{F_i(u,v)}{F_0(u,v)} \tag{2}$$

where i is 1, ..., n, (u,v) are the coordinates of a pixel in the image, F_i and F_0 are the color values of the corresponding pixels.

In order to get the ERI driving model, we adopt ERI-parameterized by PCA in our method. We have selected the maximal 23 variables and the eigenERI space to represent 97% variation in training set of all frames. Fig. 3 shows an original ERI in some frame and one reconstructed by PCA.

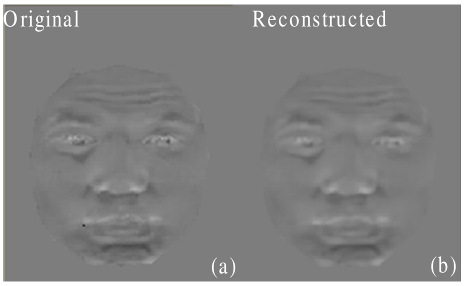

Fig. 3. F_i'**s ERI and one reconstructed.**

5. SVR-based ERI driving model

Next, an ERI driving model for its every parameter is built. Given a set of training data $\{(x_1, y_1), (x_2, y_2), ..., (x_l, y_l)\} \subset \boldsymbol{X} \times \boldsymbol{R}$, where $\boldsymbol{X}$ denotes the space of the feature of FAP, $x_i \in \boldsymbol{R^n}$ is a FAP_i, $y_i \in \boldsymbol{R^1}$ is the ERI's parameter. So, we can regard it as a regressive problem, and represent it by the support vector regression method, including ε-SVR [13] and ν-SVR [14]. Finally, for training, we need solve:

$$f(x) = \sum_{i=1}^{l}(-\alpha_i + \alpha)K(x_i, x) + b \tag{3}$$

where α_i is Lagrange multiplier, $K(x_i,x)$ is kernel function, $f(x)$ is the decision function. For predicting, it's only a dot product operation and cost very low in the real time.

Fig. 4 shows a offline learning flow chart of ERI driving model. In Section 4.3 and 4.4, we have got the samples learning, including FAPs and ERI's parameters. So, we can get a regressive model from the FAPs vectors to the ERI's parameters through SV algorithm. In a statistical sense, FTD model can recall a realistic facial appearance.

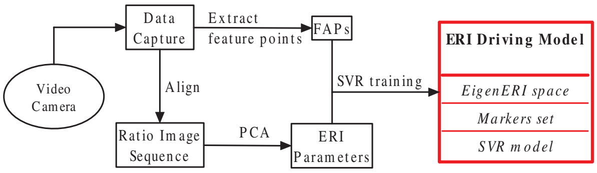

Fig. 4. Offline learning progress

6. SVR-based facial expression synthesizing

For any expression synthesis, there are an interpolation of cosine function [9] and the polynomial regression [8], etc. Here, a new method is used to do this. The ratio image is used to update illumination changes of the novel expression. Given any neutral face B, an ERI driving model and the FAP of any facial expression, there are five steps used to explain how the texture animation synthesis.

Step 1: *According to markers set of ERI Driving Model, neutral face B is morphed to* B_{temp}*. (only once)*

Step 2: *A set of ERI's parameter is predicted by a SVR-based ERI driving model when inputting a set of FAPs. Next, a new ratio image R' in the eigenERI space is synthesized.*

Step 3: *Get a texture B'(u,v) = R'(u,v) * B_{temp}(u,v) through every pixel of B.*

Step 4: *Compute the new feature positions of B by FDP in MPEG-4. Then, B' is warped according to the new feature positions of B.*

Step 5: *Set B' be a new expression image.*

The flow chart is shown in Fig. 1. Because the ERI-predicted is same at the shape, ***Step 1*** only need once for each face. By repeating ***Step 2*** - ***Step 5***, the realistic expression animation is generated under the ERI driving model. In this way, we only perform warping once instead of Liu's [5] twice for every input image.

7. Results

For more robust, we have collected the video data in a plain illumination environment, using the PC digital camera that supports continuous capture.

For choosing better SVR mode, four kinds of ERI driving model have been built, and the MSE and SCC of them are shown as Fig. 5. In evidence, the MSE and SCC of SVR with RBF kernel function are better than those of SVR with linear kernel function and the ν-SVR with RBF kernel function is applied to synthesis test.

In Fig. 6, the first column is a performance's expression images from the video; the second and the third are the textures from different illumination environment, and the neu-

Accuracy SVR		Mean Squared Error	Squared Correlation Coefficient
Epsilon	Linear	0.792391	0.500327
	RBF	0.495087	0.987423
nu	Linear	0.944577	0.516104
	RBF	0.494541	0.987423

Fig. 5. ERI driving model's testing results

tral expression textures denote image B of synthesis algorithm in Section 6. From the first row to the end, the image or the texture denotes respectively neutral, anger, disgust, fear, joy, sadness and surprise. Thereinto, the second column is performance's texture, and the third column is other's that. Expression texture shows the details have been realistic recalled.

8. Conclusions

In this work, we present a SVR-based facial texture driving model for realistic face expression synthesis. It is capable of generating the statistical realistic expression details while only requiring a set of FAPs under the plain illumination environment, and it works better in the network according to MPEG-4.

In the future work, for enhancing combinability of FTD model, we will take into account a kind of localizing method. Another is improving the ERI's model-parameterized. Finally, we need achieve and will catch a telling 3D FTD model.

Acknowledgement. This work has been supported by National Natural Science Foundation of China (NO. 60303018). We would like to thank Junfa Liu for helping with the data collection.

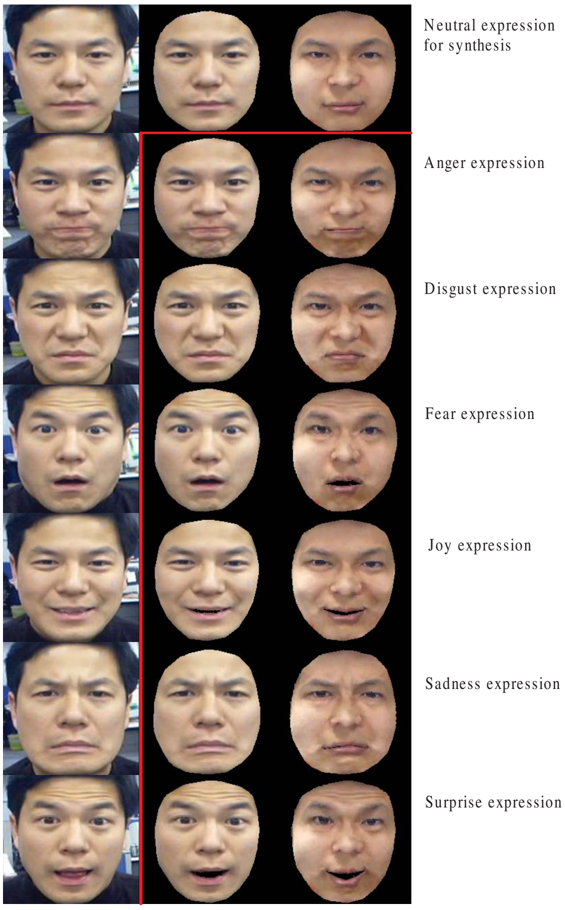

Fig. 6. Expression's contrastive test

References

[1] L. Williams. Performance-driven facial animation. In Computer Graphics, pages 235-242, Siggraph, August, 1990.

[2] G. Wol berg. Digital Image Warping. IEEE Computer Society Press, 1990.

[3] T. Beier and S. Neely. Feature-based image metamorphosis. In Computer Graphics, pages 35-42. Siggraph, July 1992.

[4] F. Pighin, J. Hecker, D. Lischinski, R. Szeliski, and D. H. Salesin. Synthesizing realistic facial expressions from photographs. In Computer Graphics, Annual conference Series, pages 75-84. Siggraph, July 1998.

[5] Z. Liu, Y. Shan, and Z. Zhang. Expressive expression mapping with ratio images. In Computer Graphics, Annual Conference Series, pages 271-276. Siggraph, August 2001.

[6] Ezzat I., Poggio T., 1996. Facial analysis and synthesis using image-based models. In: Proc. of the Second Internat. Conf. on Automatic Face and Gesture Recognition, pp. 116-121

[7] Qingshan Zhang, Zicheng Liu, Baining Guo and Harry Shum. Geometry-driven photorealistic facial expression synthesis. Eurographics / SIGGRAPH Symposium on Computer Animation, 2003.

[8] Pei-Hsuan Tu, I-Chen Lin, Jeng-Sheng Yeh, Rung-huei Liang, Ming Ouhyung. Expression detail for realistic facial animation. Proc. CAD/ Graphics 2003, pp. 20-25, Macau, Oct 28-30, 2003.

[9] Jiang Da-Long, Gao Wen, Wang Zhao-Qi, Chen Yi-Qiang. Realistic 3D facial animations with partial expression. Chinese Journal of Computers, Vol. 27, No. 6, pp. 750 757, June 2004.

[10] Yangzhou Du, Xueyin Lin. Emotional facial expression model building. Pattern Recongition Letters 24(2003) 2923-2934.

[11] K. Water, A muscle model animating three-dimensional facial expressions, Computer Graphics, 21(4):17-24, July 1987.

[12] S.Z. Li and L. Gu. Rea-time multi-view face detection, tracking, pose estimation, alignment, and recognition. In IEEE Conf. on Computer Vision and Pattern Recognition Demo Summary, 2001.

[13] V. Vapnik. Statistical learning theory. Whiley, New York, NY, 1998.

[14] B. Schökopf, A. Smola, R.C. Williamson, and P.L. Bartlett. New support vector algorithms. Neural Computation, 12:1207-1245, 2000.