FACIAL EXPRESSION SEQUENCE SYNTHESIS BASED ON SHAPE AND TEXTURE FUSION MODEL

Lei Xiong[1,2], Nanning Zheng[1], Qubo You[1], Jianyi Liu[1]

[1](School of Electronics and Information Engineering, Xi'an Jiaotong University, Xi'an 710049 China)
[2](The Air Force Engineering University, Xi'an 710038 China)

ABSTRACT

Shape and texture are two aspects in facial expression synthesis. For the requirements of synthesized shape and texture are different, it is unsuitable to use same method to synthesize shape and texture. In this paper, we propose a Statistical Shape and Radio Texture Fusion Model for facial expression sequence synthesis. In this framework, facial shape and texture had been processed separately, then they had been fused together to synthesize final result. The main contributes of this paper are: First propose statistical shape and radio texture fusion model, process shape and texture separately. Second introduce ASM model into facial shape synthesis, and construct expression model of ASM parameters. Third put forward a shape based sample select mechanism, fusion shape and texture process together. Experiment results demonstrate that the proposed model is more expressive and realistic than traditional methods, and can process batch input well.

Index Terms— Machine vision; face representation; ASM; ERI; facial expression synthesis

1. INTRODUCTION

Facial expressions play a major role in how people communicate information. Research in psychology [1] showed that facial expressions play a major role in human conversation coordination and have a greater influence on auditors than the textual content of a spoken message. So that facial expression analysis and synthesis is a very important research topic in human-computer interaction and multimedia communication [7].

Generally, facial expression synthesis methods can be divided into two classes. The first class is pixel-based method. Liu [3] propose ERI (Expression Ration Image) method to process texture information. This method can capture subtle but visually important details of facial expressions. But this method is not robust and relies so much on precise label feature points. The second class is statistical-model-based method [4]. B.Abboud[2] use AAM to synthesize facial expression. Experiments result show that it is linear relationship between neutral and expression face in AAM parameter space. The main drawback of this kind of method is that represent ability is restricted by train set, so the synthesized expression face can't keep the special facial feature of input face.

These two kinds of method have obvious shortcoming, and use only one method can't get satisfactory result. Consider the different properties between shape and texture, we propose a new fusion model. For shape information, the speciality of target facial shape can be dropped, but the generality of human facial shape must be kept. So we use ASM to process it. For texture information, the speciality of target facial texture sometimes is more important than generality of human facial texture, so we use ERI to process it. And finally, we fusion them together to synthesize target facial expression sequence.

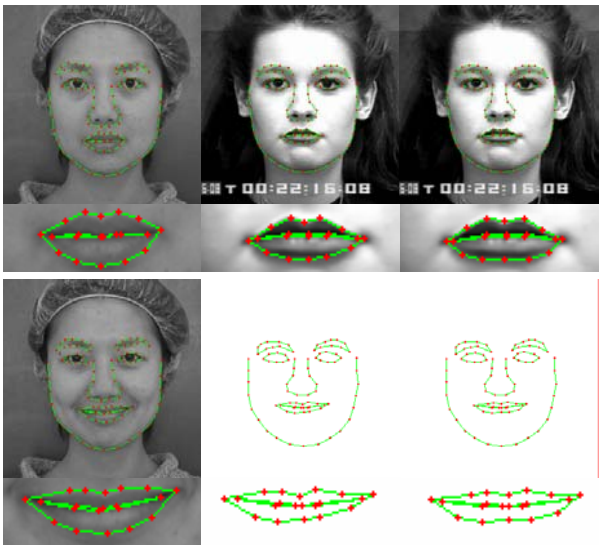
2. SHAPE AND TEXTURE FUSION MOSDEL

As we state above, shape and texture are two aspect of expression synthesis. For a successful synthesis, there are different requirements to shape and texture. The way a particular expression is displayed on a face is very person specific and thus cannot be specifically reconstructed. So the most important standard for expression synthesis is reasonability. The common features of certain expression shape must been kept, and permit to drop some shape individuality. But texture individuality, such as nevus, is very important in expression synthesis, in some sense it is more important than common features of face texture, we must keep these specific texture feature in facial expression synthesized. Because of these different requirements, we consider use different method to synthesize shape and texture of expression. The structure of fusion model is illustrated in Fig.2.

2.1. Shape process

Traditional shape synthesis method [3] is easy to understand and simply to implement, but it is too sensitive to feature point labeling, and easy to generate unrealistic result. We can see from Fig.1. Column (b) and (c) is the same person with different labeled shape of mouth. All of this two labeled mouth shape is perfect, but synthesized

shapes of smile mouth are different. Obviously, mouth shape in (b) is worse than (c), because bottom lip is not symmetrical.



(a) (b) (c)

Fig.1. Shape synthesis result with traditional method. (a) sample smile shape; (b) failed synthesized smile shape; (c)successful synthesized smile shape;

In expression shape synthesis, we need not synthesize the expression shape accurately. The most important thing is exhibit the common feature of facial shape in this expression. We can drop some personal specific in order to keep the reasonability of the synthesized shape. From this point of view, we choose statistical model to represent expression shape. Statistical model can keep common feature of this class of object, and it will drop specific feature of certain samples. It is very suitable to process shape information. Here we use ASM [6] as an example, and c is the parameters of the ASM model.

Now we need to find out the relationships between ASM model parameters of neutral face and expression face. The basic assume of Abboud [2] is a linear model can correlate the appearance parameters c to facial expression intensity according to:

$$c = \alpha_{e0} + \alpha_{e1}\varphi + \varepsilon, \quad (1)$$

Where φ is a scalar and $\varphi \in [0,1]$. $\varphi = 0$ indicate neutral expression and $\varphi = 1$ indicate high magnitude expression. ε is the approximation error, and is the start point. α_{e0} and α_{e1} are coefficient vectors learned for each facial expression e (joy, fear, disgust, surprise, fear, sadness and neutral) by linear regression over the training set.

But this assume has a limitation that α_{e0} and α_{e1} are same no matter which target face input. This result the change of synthesized shapes is similar with each other. In order to produce more realistic result we propose a two step model.

1) We believe that different faces should have different linear models. So, different input face will has different α_{e0} and α_{e1}. We use α_{e0}^{i} and α_{e1}^{i} to indicate it. So we have

$$c(\varphi) = \alpha_{e0}^{i} + \alpha_{e1}^{i}\varphi, \quad (2)$$

c is the parameter with facial expression intensity φ. Here we didn't need ε any more, because different input will have different α_{e0}^{i} and α_{e1}^{i}, and approximation error will include in the α_{e0}^{i}.

2) In order to get α_{e0}^{i} and α_{e1}^{i}, we use quadratic polynomial to learn $\beta_{e2}, \beta_{e1}, \beta_{e0}$ over train set.

$$c_{ie} = \beta_{e2} \cdot c_{in}^{2} + \beta_{e1} \cdot c_{in} + \beta_{e0}, \quad (3)$$

where c_{ie} is the parameter for each facial expression e and $\varphi = 1$, and input face number is i. c_{in} is the neutral face parameter of input face i. $\beta_{e2}, \beta_{e1}, \beta_{e0}$ are coefficient vectors learned for each facial expression by quadratic

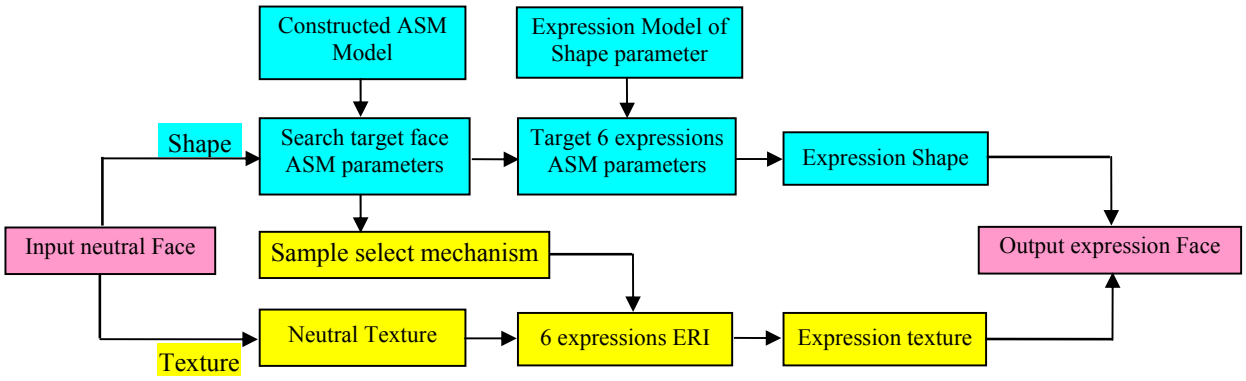


Fig.2. System structure of Shape and Texture Fusion Model

polynomial over the train set. After learn $\beta_{e2}, \beta_{e1}, \beta_{e0}$ from train set, we can calculate c_{ie} for every input face i, so we have two points c_{ie} and c_{in} to perform linear regression to get α_{e0}^{i} and α_{e1}^{i}.

In this 2 step model, we can get a continuous shape series when the facial expression intensity φ changes from 0 to 1. And for different input face, the liner model is different. So we can synthesize a shape sequence of any expression with an intensity parameter φ.

2.2. Texture process

Pixels-based method is a suitable solution for texture synthesis. One of the successful pixels- methods is ERI[3]. The biggest virtue of ERI method is that it can keep all personal facial features of target face when we transfer illumination change of certain expression. In order to get expression texture sequence, we use facial expression intensity φ same as used in shape model. $\varphi = 0$ indicate neutral expression and $\varphi = 1$ indicate high magnitude expression. The ERI with different expression intensity φ can be got use the method below:

$$B' = \left[(R_g - I)\varphi + I \right] B_g \qquad (4)$$

B' is target expression face, R_g is the ERI, B_g is target nature face. With equation (4), we can synthesize a texture sequence of any expression with an intensity parameter φ.

2.3. Shape and texture fusion

ERI didn't consider how to select proper sample to generate radio images. For every expression, if we select one sample to calculate ERI, no matter which target images input, generated target expression texture are similar with each other, it is so stiff.

Here we use shape information to select most suitable sample to calculate ERI. We calculate the ASM parameter distance between target face and all sample faces, which indicate the shape similarity between them.

$$d_i = \left\| c_{target} - c_{sample-i} \right\| \qquad (5)$$

We select the sample with minimum d_i, which means the most similar sample, to calculate ERI. So that different input target images have different ERI and synthesize different expression texture, this is more reasonable when a batch images input.

3. EXPERIMENT RESULTS

Here, we tested the proposed method on the Cohn-Kanade Facial Expression Database [5] and AIAR facial expression database. We select 517 images totally. In order to get rid of the influence of irrelevant, images are translated, rotated and scaled to size 256*256 with centers of the eyes are placed on specific pixels. And we label the image with 87 feature points, and keep 42 biggest eigenvector to build ASM model.

First we compare our model with AAM result in literature [2]. It is show in Fig.3.

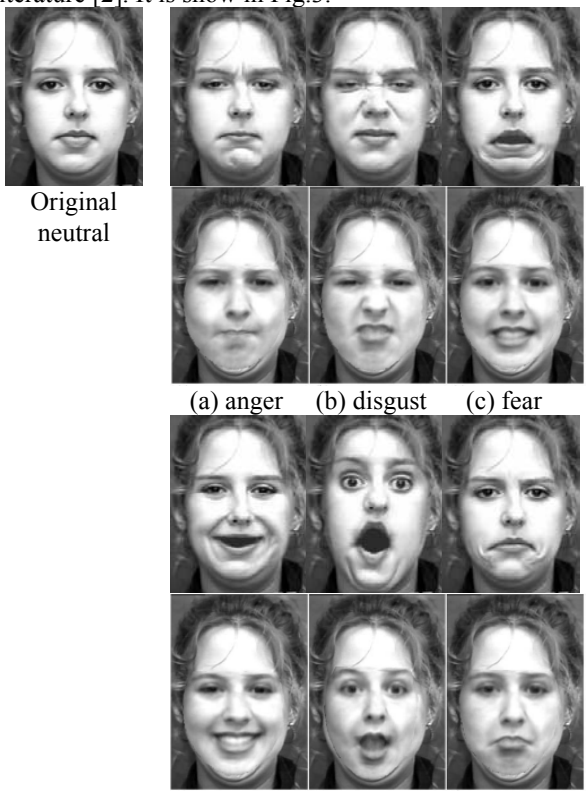


Fig.3. six basic expressions compare between fusion model and AAM result [2]. The upper images are result of our fusion model, the bottom are AAM result.

From the Fig.3, we have 2 conclusions: 1. our fusion model is more expressive than AAM. In all 6 expressions, it can exhibit more facial details than statistical models. 2. our fusion model is more realistic than AAM. It can keep more personal texture features of target face. For example, the nose, eyebrow of the synthesized expression faces are more similar with target neutral face.

Then we compare our method with ERI and results are illustrated in Fig.4. For 4 different target neutral face, the ERI use same happy radio image to synthesize target happy texture. They have similar smile with each other, especially the wrinkle around the mouth. Our fusion model can select suitable sample to generate happy radio image for different

target face. Different target face generate different smile. So our model is more reliable in batch input.

Finally, we input a new target face into the fusion model, and synthesize expression sequence. The result is show in Fig.5. The synthesized expression sequences not only keep the facial feature of target face, but also exhibits dynamic wrinkle of facial expression. Further more, with the control of intensity parameter φ, we can easily synthesize shape and texture in different intensity.

Fig.4. Compare between our fusion model with ERI. (a) original neutral face; (b) result of our model; (c) result of ERI.

4. CONCLUSIONS

In facial expression synthesis, the requirements for shape and texture are different. For shape synthesis, we need to keep general property of expression shape, and permitted to drop some specific property. On the contrast, for texture synthesis, we must keep specific texture feature, such as nevus, and permitted to drop some general feature. Based on above analysis, we propose a shape and texture fusion model. We use ASM to process shape, and ERI arithmetic to process texture, and a shape based sample select mechanism was propose to fusion two parts together. Experiment results demonstrate this fusion model is more expressive and realistic than traditional statistical model, and more robust and easy to process batch input than ERI method. The model not only keeps the basic facial feature of target face, exhibits dynamic wrinkle of facial expression, but also can synthesize expression sequence easily.

REFERENCES

[1] E. Boyle, A.H. Anderson. A. Newlands, "The effects of visibility on dialogue performance in a cooperative problemsolving task," *Language Speech* 37 (1), pp. 1–20, 1994.

[2] Bouchra Abboud, Franck Davoine, Mo Dang, "Facial expression recognition and synthesis based on an appearance model," *Signal Processing: Image Communication* 19, pp. 723–740, 2004.

[3] Z. Liu, Y. Shan, Z. Zhang, "Expressive Expression Mapping with Ratio Images," *SIGGRAPH 2001*, pp 271-276, 2001.

[4] John Ghent, "*A Computational Model of Facial Expression*," Doctor Thesis of Philosophy National University of Ireland, July 2005.

[5] Kanade. T, Cohn, J.F, & Tian, Y, "Comprehensive Database for Facial Expression Analysis," *Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition (FG'00) Grenoble*, France, March 2000.

[6] T.F.Cootes, C.J.Taylor, Active Shape Models-Their Training and Application, *Computer Vision and Image Understanding* Vol.61, No.1, January, pp. 38~59, 1995.

[7] Cosatto, E. and Graf, HP: Photo-realistic talking-heads from image samples. IEEE Transactions on Multimedia (3) ,2000.

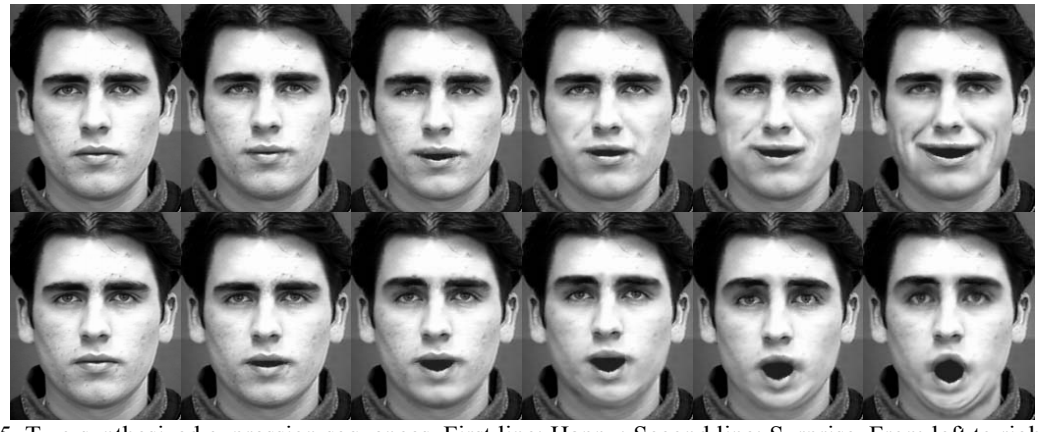


Fig.5. Two synthesized expression sequences. First line: Happy; Second line: Surprise. From left to right the frames corresponding to expression intensity parameter φ equals to 0, 0.2, 0.4, 0.6, 0.8 and 1 respectively.