Expression Recognition Driven Virtual Human Animation

Junghyun Cho[1], Yu-Jin Hong[1,2], Sang C. Ahn[1], and Ig-Jae Kim[1,2]

[1] Imaging Media Research Center, Korea Institute of Science and Technology, Korea
[2] Dept. of HCI & Robotics, Korea University of Science and Technology, Korea
{jhcho,hyj,asc,kij}@imrc.kist.re.kr

Abstract. Since the character expressions are high dimensional, it is not easy to control them intuitively with simple interface. So far, existing controlling and animating methods are mainly based on three dimensional motion capture system for high quality animation. However, using the three dimensional motion capture system is not only unhandy but also quite expensive. In this paper, we therefore present a new control method for 3D facial animation based on expression recognition technique. We simply utilize off-the-shelf a single webcam as a control interface which can easily combine with *blendshape* technique for 3D animation. We measure the user's emotional state by a robust facial feature tracker and facial expression classifier and then transfer the measured probabilities of facial expressions to the domain of *blendshape* basis. We demonstrate our method can be one of efficient interface for virtual human animation through our experiments.

Keywords: 3D facial animation, control interface, blendshape, facial feature tracking, expression recognition.

1 Introduction

Controlling 3D avatar, especially its emotional expression, is widely used in a variety of applications such as teleconference, movies, real-time avatar game, and human computer interaction. Therefore, it is very important to provide intuitive control interface for avatar animation. Due to high complexity of human expressions, however, it is difficult to control over three dimensional facial animation with simple interface.

Traditionally, we should use a facial motion capture system to express the detail emotional variation of virtual avatar. The facial motion capture is the process of electronically converting the movements of a user's face into a digital database using multiple cameras. A facial motion capture database describes the coordinates of reference points on the user's face.

Typical marker based motion capture systems apply up to 100 markers to the users face and track the marker movement with high resolution cameras to get accurate movements of facial parts so that we can express facial expression finely that has high degree of freedom. Unfortunately these systems are expensive, complicated, and

N. Streitz and P. Markopoulos (Eds.): DAPI 2014, LNCS 8530, pp. 89–96, 2014.

time-consuming to use. Marker-based systems are accurate but cumbersome and do not allow full expression for the actor due to the attached markers.

Nowadays, researchers focused on markerless technologies use the features of the face such as the corners of the lips and eyes, and wrinkles and then track them. This technique is much less cumbersome, and allows greater expression for the actor but accuracy is less than the one of marker based.

Two dimensional capture can be achieved using a single camera but this produces less sophisticated tracking, and is unable to fully capture three dimensional motions.

To overcome these drawbacks when we apply two dimensional capture method, Chai et al. [9] proposed the system to extract a small set of animation control parameters from video. Because of the nature of video data, these parameters may be noisy, low-resolution, and contain errors. Their system used the knowledge embedded in motion capture data to translate these low-quality 2D animation control signals into high-quality 3D facial expressions. To adapt the synthesized motion to a new character model, they introduced an efficient expression retargeting technique whose run-time computation is constant independent of the complexity of the character model.

Although this method enabled us to get a real-time facial animation but it needs large 3D motion dataset in advance.

So, we propose a new method which use only a single webcam but doesn't require large 3D motion database. Once we set the animation control interface, we have to choose animation production method. In terms of 3D facial animation production, the most popular approach currently is the blendshape technique, which synthesizes expressions by taking a linear combination of a set of pre-modeled expressions.

A fundamental question in developing a blendshape-based facial animation system is how to form the expression basis. A casual approach is to use an expression basis comprised of manually modeled, intuitively recognizable key expressions. In our approach, we build each member of the expression basis beforehand according to the shapes of emotional expressions that we could classify [3]. Here, we set the number of basis is equal to the number of emotional expressions that we could recognize.

Another fundamental issue that must be solved when developing a blendshape technique is how to assign the weights to each member of the expression basis set in order to produce the desired expression sequences. As we set the same number of expression basis to the emotional expression, we can simply assign the recognition result, in terms of probability, of each pre-trained expression to the weight of each member of basis at every video frame.

The remainder of this paper is organized as follows. Section 2 explains the facial expression recognition technique we used. Section 3 presents the procedure for avatar animation. Section 4 reports the experimental results, and Section 5 concludes the paper.

2 Facial Expression Recognition

An automatic facial expression recognition system needs to solve the following problems: detection and location of faces in a cluttered scene, facial feature extraction, and facial expression recognition.

2.1 Face Detection

Face detection has been regarded as the most complex and challenging problem in the field of computer vision, due to the large intra-class variations caused by the changes in facial appearance, lighting, and expression. Such variations result in the face distribution to be highly nonlinear and complex in any space which is linear to the original image space.

Face detection techniques have been researched for years and much progress has been proposed in literature. Most of the face detection methods focus on detecting frontal faces with good lighting conditions. These methods can be categorized into four types, such as, knowledge-based, feature invariant, template matching and appearance-based.

Any of the methods can involve color segmentation, pattern matching, statistical analysis and complex transforms, where the common goal is classification with least amount of error. Bounds on the classification accuracy change from method to method yet the best techniques are found in areas where the models or rules for classification are dynamic and produced from machine learning processes.

Viola and Jones [2] presented a fast and robust method for face detection which is 15 times quicker than any technique at the time of release with 95% accuracy. The technique relies on the use of simple Haar-like features that are evaluated quickly through the use of a new image representation. We used this technique in this work.

2.2 Facial Feature Tracking

In recent years, research for facial feature extraction and tracking methods became popular among computer vision society, such as SDM [6], CLM [7, 8], Shape Regression [10], and so on. Owing to these developments, we could get very robust facial feature tracker under various situations. Among them, we developed our algorithm to locate facial features based on tree-regression concept. We simply extract facial feature candidates based on intensity difference of two pixels in the image which are randomly selected. The selected pixels are indexed by the same local coordinate have the same semantic meaning. This enables the tracker to work better under geometric distortion. To select the effective features from a large number of features, we used *correlation based feature selection* method, then simplify the formula which decrease the cost of computations. For every node of the tree in every level one feature would be selected. Maximum $L^2 - 1$, here L is the number of maximum level of tree, features will be selected for every tree. Discriminative features are highly correlated to the regression target. The target $\Delta\hat{S}$ is vectorial delta shape which is the difference between the ground truth shape and current estimated shape, $\Delta\hat{S} = \hat{S} - S^{t-1}$. Good features would have highest correlation value. The following steps to find the features for each tree node which we want to build:

- Project regression target ΔS to a random direction to produce a scalar.
- Select the feature with highest correlation to the projection.
- Repeat step 1 and 2 for every node to obtain new features.
- Select optimum thresholds to generate tree.

Once we build the tree, we search facial features as follows. Face candidate area in an image will be detected by a Viola-Jones face detector, and then a mean shape will be located at the area. High correlated features that had been calculated in training phase will be extracted from the area for every tree. These features will lead to a leaf of the tree and they are led to predict a delta shape ($\Delta\hat{S}$) that will be added to the mean shape. This updates the facial feature shape in every regressor. Fig. 1 shows our search process.

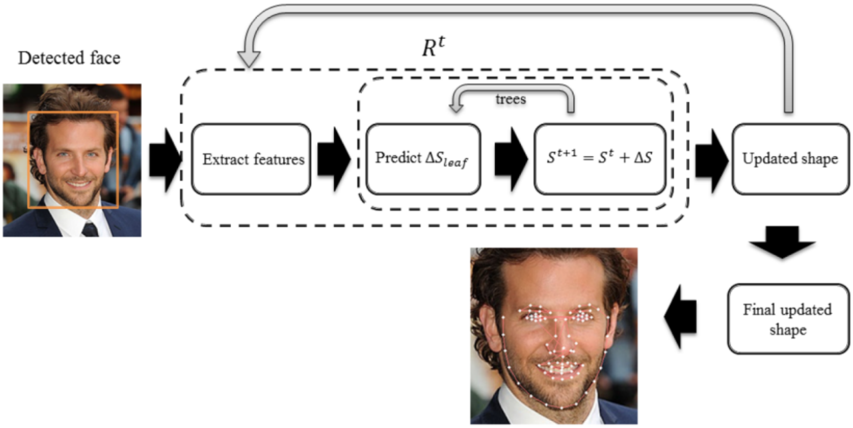

Fig. 1. Search block diagram of tree based facial feature extractor

2.3 Facial Expression Classification

We used a Gaussian Mixture Model (GMM) for our facial expression recognition. A GMM is a parametric probability density function represented as a weighted sum of M component Gaussian densities [1] as given by the equation,

$$p(\mathbf{x}|\lambda) = \sum_{i=1}^{M} \omega_i \, g(\mathbf{x}| \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \tag{1}$$

where $\mathbf{x}$ is a D-dimensional continuous-valued data vector, ω_i, i=1,….,M, are the mixture weights, and $g(\mathbf{x}| \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, i=1,….,M, are the component Gaussian densities. Each component density is a D-variable Gaussian function of the form. In our experiment, we extract 68 facial feature points and classify six expressions that are defined in FACS [11], such as happiness, sadness, surprise, anger, disgust and neutral, so D and M are 136 and 6, respectively.

Given training data and a GMM configuration, we estimate the parameters of the GMM, which matches the distribution of the training feature vectors. There are several techniques available for estimating the parameters of a GMM. Among them, we used maximum likelihood estimation. We capture T frames per each expression, here we set T=90, for learning each one. For a sequence of T training vectors $X = \{\mathbf{x}_1, \dots, \mathbf{x}_T\}$, the GMM likelihood, assuming independence between the vectors, can be written as,

$$p(X|\lambda) = \frac{1}{T}\sum_{t=1}^{T} \Pr(i|\, \mathbf{x}_t, \lambda) \tag{2}$$

To solve a non-linear function of the parameter λ, we estimate ML parameter iteratively using a special case of the expectation-maximization(EM) algorithm. On each EM iteration, the following re-estimation formulas are used which guarantee a monotonic increase in the model's likelihood value,

Mixture Weights

$$\bar{\omega}_i = \frac{1}{T}\sum_{t=1}^{T} \Pr(i|\, \mathbf{x}_t, \lambda) \tag{3}$$

Means

$$\bar{\mu}_i = \frac{\sum_{t=1}^{T} \Pr(i|\, \mathbf{x}_t,\lambda)\, \mathbf{x}_t}{\sum_{t=1}^{T} \Pr(i|\, \mathbf{x}_t,\lambda)} \tag{4}$$

Variances (diagonal covariance)

$$\bar{\sigma}^2{}_i = \frac{\sum_{t=1}^{T} \Pr(i|\, \mathbf{x}_t,\lambda)\, x^2{}_t}{\sum_{t=1}^{T} \Pr(i|\, \mathbf{x}_t,\lambda)} - \bar{\mu}^2{}_i \tag{5}$$

3 Avatar Animation

The principle of blendshape interpolation is similar to basis member interpolation. In this case, more than 2 base members can be used at a time, and the interpolation is for a single static expression, rather than across time. Each blendshape can be modeled using a variety of different methods. The amount of detail in each expression can also vary, as long as the resulting faces can be 'combined' in some manner. While this method is popular for specifying facial animation, it requires manual specification, and designing a complete animation can be quite time consuming [4].

3.1 Building Expression Basis

Building appropriate key shapes is an important part of shape decomposition. Each key shape adds flexibility and expressiveness to the model, suggesting that many key shapes should be used. However, the user must create a target model for each key shape. In order to reduce user burden the number of key shapes should be kept small. An ideal method would balance these requirements to find the minimal set of key shapes that maintains the desired animation expressiveness [4]. Here, we propose simple but very efficient way for building our expression basis. We select six representative expressions which our classifier can tell more separately and build the same number of expression basis with the shape of corresponding expression of the subject.

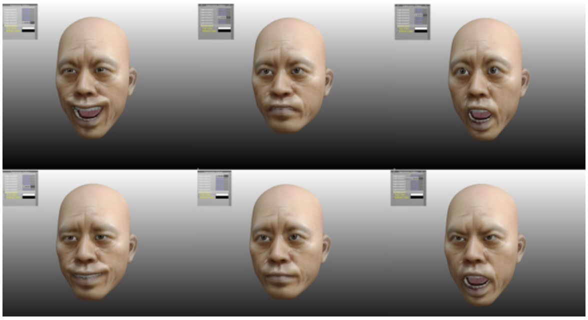

Fig. 2. Six expression basis used in our experiment

4 Experimental Result

For our experiment, we let two test subjects train pre-defined six expressions using a webcam which is installed in front of the monitor. For this, we allow each subject to capture image sequences during three seconds (T=90 frames) for each expression and input them the training engine to build our Gaussian mixture model. We modeled two different avatars which have six different expression bases like Fig. 2. Then we evaluate our proposed system can make the virtual avatar visualize natural expression according to the subject's performance.

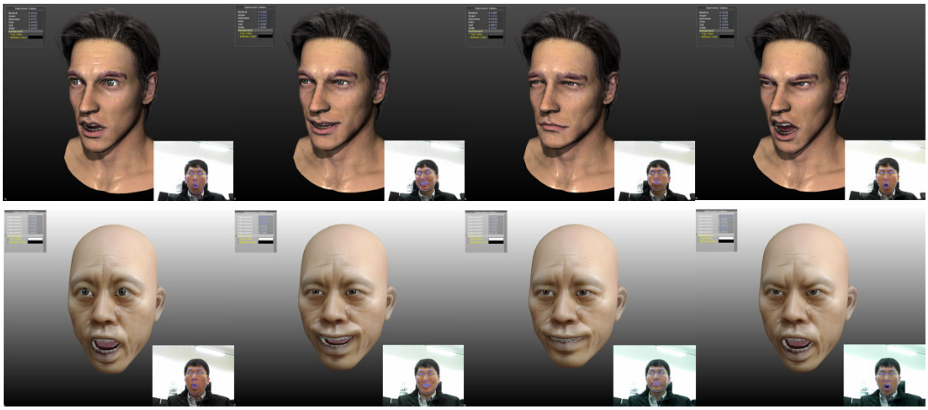

Fig. 3. Example animation sequences with different expressions

Once each subject trained his/her expressions, our classifier showed high performance of recognition over all so we could get avatar's natural expressive animation along with the performance of each subject. Fig. 3 shows two different characters make same expressions following the subject's performance. This example also shows our method can easily control any other different characters without modifying interface. If there's a falling off in recognition quality due to some noises, abrupt expression changes may be occurred. As a consequence of that, we might have weird avatar's expression. Just in case of that, we applied Kalman filter [5] which enables us to get smooth transition. We could get realistic expression of the avatar even though we applied different virtual character to the subject simply by transferring the basis weights. Through our test, we could confirm our proposed system can be efficient and useful interface for controlling a virtual avatar.

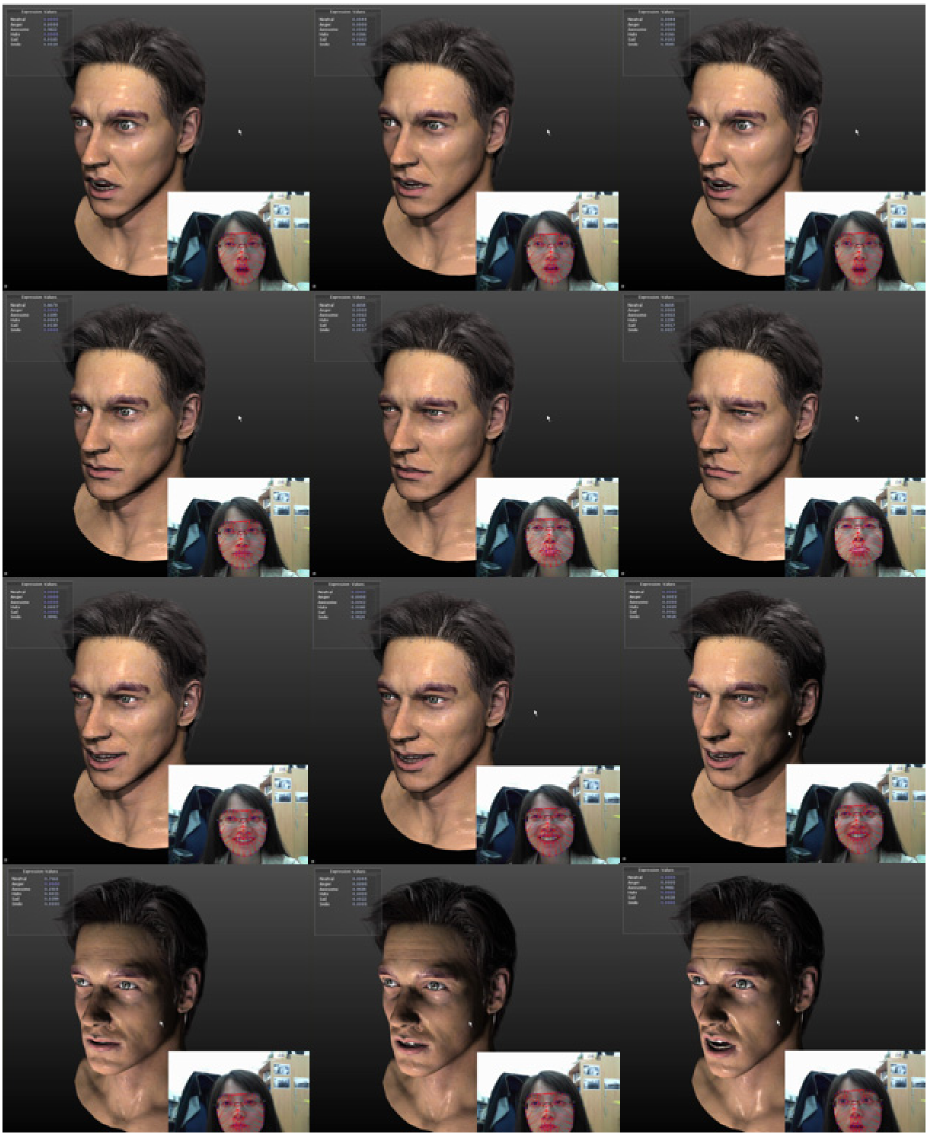

Fig.4. Another example sequences following female subject's expression. Top row shows surprise, second rows show the transition from neutral to sad, third row shows happy state. Especially, bottom row sequences show the stressed expression results by real-time rendering technique.

5 Conclusion

In this paper we propose a new efficient method for facial animation of virtual avatar. By combining the facial feature tracking and facial expression classifying methods with the blendshape interpolation technique, we can obtain the real-time facial animation control without the aid of large 3D facial database and facial tracking devices. In here, our control interface is based on only six expression clusters. Since these preselected expressions are not enough to span all the expression space, some possible expressions may not be shown. If we could improve our classifier's performance so that we could tell subtle expression change more precisely, we may expand the coverage of expression space. This may increase the quality of our control interface.

Generally, the controlling of virtual avatar, especially its expression, is not easy due to its high complexity. Our system classified emotional state by a single webcam and can visualize almost possible expression by simple transferring the measured probability of each expression to the weight of each blendshape basis. Through our experiment, we believe our expression based interface can be one of efficient solutions for controlling virtual avatar animation with inexpensive equipment.

Acknowledgement. This work was supported by the Global Frontier R&D Program on <Human-centered Interaction for Coexistence> funded by the National Research Foundation of Korea grant funded by the Korean Government (MEST) (2012M3A6A3055705) and the KIST Institutional Program (Project No. 2E24790).

References

1. Reynolds, D.A.: A Gaussian Mixture Modeling Approach to Text-Independent Speaker Identification., PhD thesis, Georgia Institute of Technology (1992)
2. Viola, Jones: Rapid object detection using a boosted cascade of simple features. In: Computer Vision and Pattern Recognition (2001)
3. Kim, I.J., Ko, H.: Intuitive Quasi-Eigen Faces. ACM GRAPHITE (2007)
4. Chuang, E., Bregler, C.: Performance driven facial animation using blendshape interpolation, Stanford Technical Report CS-TR-200202 (2002)
5. Kalman, R.E.: A new approach to linear filtering and prediction problems. Journal of Basic Engineering 82(1), 35–45 (1960)
6. Xiong, X., Torre, F.: Supervised Descent Method and its Applications to Face Alignment. IEEE Computer Vision and Pattern Recognition (2013)
7. Cristinacce, D., Cootes, T.: Feature detection and tracking with constrained local models. In: BMVC (2006)
8. Saragih, J.M., Lucey, S., Cohn, J.F.: Face Alignment through Subspace Constrained Mean-Shifts. In: IEEE International Conference on Computer Vision (2009)
9. Chai, J., Xiao, J., Hodgins, J.: Vision-based Control of 3D Facial Animation. In: Eurographics/SIGGRAPH Symposium on Computer Animation (2003)
10. Cao, X., Wei, Y., Wen, F., Sun, J.: Face alignment by explicit shape regression. IEEE Computer Vision and Pattern Recognition (2012)
11. Ekman, P., Friesen, W.: Facial Action Coding System: A Technique for the Measurement of Facial Movement. Consulting Psychologists Press, Palo Alto (1978)