Automatic Facial Gesture Construction Using Image Warping

Stephen Karungaru, Minoru Fukumi and Norio Akamatsu
Faculty of Engineering, University of Tokushima,
2-1, Minami-Josanjima, Tokushima 770-8506. Japan.
karunga@is.tokushima-u.ac.jp

ABSTRACT

In this paper, we present a method by which different facial expressions can be constructed from an expressionless face image. The user is required to provide only the image and the desired facial expression. The rest of the processes are fully automatic including the selection and detection of the control points necessary to perform the warping. Given a visual scene, our method detects and extracts the position of a face using a neural network based method. Then, from the detected face position, the lips region and hence the warping data is extracted. The warping algorithm (expression creator) used is based on the triangulation method that uses triangles to deform an image. The required facial expression can then be created by selecting it from a list of keywords, for example smiling, angry, surprised etc. This work achieves an average face expression creation accuracy of 95.8%.

Keywords: Face expression, lips extraction Warping and Neural network.

1. INTRODUCTION

There are many types of facial gestures that a human can make. The general classes of facial expressions include smiling, angry, surprised, sad etc. While a person's facial expression can be grouped in one of these classes, each person's expression is unique. In this paper we explore the possibility of artificially creating these facial expressions. For a given expressionless face input, our method creates a facial expression selected by the user from a list of keywords.

This method starts with the detection of a face from a visual scene using a face detector. The lips region is then extracted from the face region using a color threshold method. Once the keyword has been input, the face is then warped to the selected face expression using triangular interpolation. The processes are described in details in this paper.

Face expression detection has been given a lot of attention by many researchers. Most methods use image similarities as cue. Some use only global features such as color [1] and texture histograms [2]. Tieu and Viola [4] present a boosting approach to select a small number of features from a very large set allowing fast and effective online classification. In most methods however, the data complexity is usually a major handle to clear. Pentland et al [4] show how principal Component Analysis (PCA) can reduce the dimensionality while maintaining the systems level of performance. Other methods include those that depend on local information to aid segmentation, [5], [6]. Silapachote et al [7] proposes a classification technique that uses the AdaBoost for learning. Although many face expression detection recognition methods have been proposed, few methods exist that attempt to create the face expressions given an expressionless image. This work concentrates on that area.

The computer simulations in this work were carried out using a Dell Optiplex SX260 Pentium 4 personal computer.

The rest of this paper is organized as follows. Section 2 describes the process of face detection with section 3 describing how to extract the lips region from the face detected in section 2. The control points learning neural network is explained in section 4.1 Section 4.2 describes the warping process and computer simulations are in section 5. Section 6 concludes this work.

2. FACE DETECTION

Initially, from a given image, the position of the face region in the images needs to be determined. This is accomplished using a neural network based face detector. The face detector, referred to as face detection neural network (FDNN) extracts the position of the face from an image [8]. The FDNN consists of a face locator, down sampler and a merger.

The face locator consists of a skin color detector and a face detector neural network. At first, the skin color regions of the given image are detected using the skin color detector.

The skin color detector was implemented using a YIQ color system threshold based method. The neural network face detector is the part of this system that does the actual face detection.

The face detector is chosen to be a three layered back propagation trained neural network. The size of the training samples was set at 20x20 pixels because experiment showed that both the accuracy of the system and the speed were best using this size

The error back propagation method [9] is used to train the neural network. The system is trained to produce an output of 0.95 for a face and 0.05 for a non-face. Structural learning with knowledge [10] is also used during training to reduce the size of the face locator and therefore improve the overall speed.

However, the face locator can only detect faces whose size is about 20x20 pixels. A down sampler provides the face locator with the ability to detect faces that are larger than 20x20 pixels.

Since multiple detections are likely to occur around the same face, a merger is used to combine the multiple detections into one.

The FDNN used independently has an accuracy of 97.4% when tested with images with complex backgrounds and including many people per image. In this work, because of the relatively simple background (uniform) and one subject per scene, the FDNN's accuracy is almost100%..

3. LIPS REGION EXTRACTION

After the position of the face has been detected, the lips region can be extracted using a color threshold method. The lips region can be found relatively fast because search for the lips region is carried out only in the lower half of the face regions.

The threshold method used for this is based on the YIQ color space [11]. Only one threshold is used since we averaged the two color components *Q* and *I* into one.

$$39 \le IQ \le 69 \qquad (1)$$

Where: IQ is the sum of the I and Q color components

The threshold above segments the lips region below the center of the face. The threshold is set to capture the redness of the lips. Note that, it is assumed that the lips are mostly reddish and the female subjects in our test will not wear a different color lipstick.
This method highly depends on the accurate extraction of the lips region. If this is not achieved, then the results of warping to create a selected face expression from an expression face look very unnatural.

However, the simple threshold shown in equation (1), works well because, the lips regions are searched for in a known space, the face. The skin region has already been pre-selected during skin color detection and face detection steps.

4.0 SYSTEM DESIGN

The system designed here to create various face expressions given an expressionless face consists of a neural network and a warping method.

The neural network is used to learn the relationships between the expressionless face and the other face expressions. It also provides a method to select the desired face expression during testing. Note that the neural network is trained using the lips region positions and not the image data.

Warping is the method used to actually create the desired face expression selected in the neural network section. In this work, the warping method is the triangular based method. This method was selected because of its simplicity and high speed.

4.1 Neural Network

A neural network is trained to determine the warping control points required for a given facial expression. Therefore, the neural network training data consists of the control points extracted from an expressionless face and the desired face expression keyword.

A lips region can be warped using four control points. These control points are the side corners of the lips and the center of the top and bottom edges.

Using face images from the AR face database [12] the warping control points for this method are determined. The database contains over 4000 color images corresponding to 126 peoples faces. Images feature frontal views of the subjects making several facial expressions, in different illumination conditions and with or without occlusions.

40 subjects were then randomly selected from the database to train the neural network. For each subject,

expressionless, smiling, sad, angry and surprised images were used. We then manually extracted the position of the lips from all the images and also extracted the four control points selected for training and warping. The positions were then averaged for each expression. This gave us the average relationship between the expressionless and other expressions, for example the smiling face, Fig. 1.

(a)

(b)

Fig. 1. Average lips position for 40 persons. (a) Expressionless face and (b) Smiling face

The neural network has three layers. The input layer has seventeen nodes. The first one represents the desired facial expression and the other sixteen, eight (each point has an x and y value) each to represent the control points of the present image and the average image constructed earlier. Note that the input consists of data extracted from the expressionless face only.

The hidden layer has 25 nodes determined using trial and error method.

The output consists of eight nodes. The data used is similar to the input data but it is extracted from the desired expression specified by the input. The neural network is shown in Fig 2.

The positions of the lips are normalized before being input into the neural network. Note that before normalization, these positions, x and y values, are usually more than 1. Therefore, they are all normalized to values between 0 and 1 by dividing them by 1000. Similarly, during warping the outputs of the neural network must be reconverted into their original format by multiplying them by 1000.

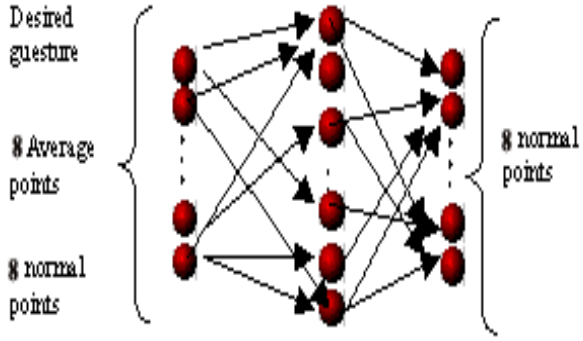

Fig.2. Neural network Structure

Since the inputs and outputs are image positions, they are normalized to the range zero to one before being used to train the neural network.

4.2 Warping

The warping method used in this work is triangles based interpolation. The images are divided into several triangles using the control points already found. For two corresponding triangles, one from the source image and the other from the target image, the warping transformation from one to the other can be performed.

Let points P_1, P_2, P_3 on the source image be located at $\mathbf{x}_1=(u_1, v_1)$, $\mathbf{x}_2=(u_2, v_2)$ and $\mathbf{x}_3=(u_3, v_3)$. Also let points Q_1, Q_2, Q_3 on the target image be located at $\mathbf{y}_1=(x_1, y_1)$, $\mathbf{y}_2=(x_2, y_2)$ and $\mathbf{y}_3=(x_3, y_3)$. The points on the source image can be mapped to those on the target image using eqs. 1 and 2.

$$x = a_{11}u + a_{21}v + a_{31} \quad (2)$$

$$y = a_{12}u + a_{22}v + a_{32} \quad (3)$$

In matrix form, these two equations can be rewritten as,

$$\begin{bmatrix} x_1 & y_1 & 1 \\ x_2 & y_2 & 1 \\ x_3 & y_3 & 1 \end{bmatrix} = \begin{bmatrix} u_1 & v_1 & 1 \\ u_2 & v_2 & 1 \\ u_3 & v_3 & 1 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & 0 \\ a_{21} & a_{22} & 0 \\ a_{31} & a_{32} & 1 \end{bmatrix} \quad (4)$$

The coefficients a_{11}, a_{12}, a_{21}, a_{22}, a_{31} and a_{32} can be found by solving the equation below.

$$\begin{bmatrix} a_{11} & a_{12} & 0 \\ a_{21} & a_{22} & 0 \\ a_{31} & a_{32} & 1 \end{bmatrix} = \begin{bmatrix} x_1 & y_1 & 1 \\ x_2 & y_2 & 1 \\ x_3 & y_3 & 1 \end{bmatrix} \begin{bmatrix} u_1 & v_1 & 1 \\ u_2 & v_2 & 1 \\ u_3 & v_3 & 1 \end{bmatrix}^{-1} \quad (5)$$

Warping can then be done in two ways. One, source to target mapping, that is, for each source pixel, calculate the target pixel and two, target to source mapping in which case the source pixel is calculated for every target pixel.

Target to source is more desirable because no pixels on the target image are missed and therefore no color interpolation is necessary as in source to target mapping. This saves on the total computation time.

5. COMPUTER SIMULATION AND RESULTS

Computer simulations were then carried out to prove the effectiveness of this method. To calculate the accuracy of this system, two human observers were asked to classify the results of this work based on their perception of what a particular face expression ought to look like.

5.1 Test Data

The test database was constructed using AR face database [12] also used to train the neural network. 30 subjects were selected to perform the test. For each subject, expressionless, smiling, sad, angry and surprised images were used. Note that, the image selected for testing are different from the ones used for training the neural network.

5.2 Simulation Procedure

For a given expressionless face image, faces of other expressions can be created using the procedure described below.

(i) Using the face detector, extract the position of the face and calculate its center position.
(ii) Search for the lips region in the extracted face's lower area using the threshold method. Normalize the results by dividing each value by 1000.
(iii) Enter the required expression's keyword in to the neural network and run it.
(iv) Calculate the warping positions. Multiplying the resulting neural network outputs by 1000 does this.
(v) Warp the face using the triangles base method described earlier.

5.3 Results

The results of this method were verified using two human observers. This means that, the expressions created were presented to two people who were asked to classify the expressions. The two people were labeled Verifier A and B. Table 1 shows the result.

Table 1. Verification of the face expressions created using this system by two human observers.

Verifier	Face expression created % Accuracy			
	Smile	Sad	Angry	Surprised
A	93.3	96.6	100	93.3
B	96.6	96.6	96.6	93.3
Average	95.0	96.6	98.3	93.3

From table 1, the average results of verification by the two human observers were 95%, 96.6%, 98.3% and 93.3% for creation of smiling, sad, angry and surprised

face expressions respectively. The average system accuracy achieved was 95.8%.

The main reason for the failures and sometimes some false detects (about 1 in 30 or 3.3 %) was the human observers comments that the face expression created was not natural looking. This problem can be thought to be due to inaccurate lips region extraction leading to wrong values being fed into the warping algorithm.

Fig 3 shows a selected result of this method for the case where the keyword smile was entered into the system. Fig 3 (a) shows the original image and Fig 3 (b) the smiling image that was created.

Fig. 3 (c) shows the result in the case where angry (disgusted) was entered as the keyword.

6. CONCLUSION

In this paper, a method to create artificial face expressions given an expressionless face was proposed. This method first estimated the average position of several lips positions in expressionless and other facial expressions and used a neural network to learn the relationship between them. The warping method was employed to create the desired expressions from the neural network output data.

The average results of verification by the two human observers were 95%, 96.6%, 98.3% and 93.3% for creation of smiling, sad, angry and surprised face expressions respectively. The average system accuracy achieved was 95.8%. The expressions created looked “very close” to the real ones.

In future, a new verification method, for example template matching, is required to determine the accuracy of the face expressions created. This will eliminate the reliance on the human observer to verify the results. It is also hoped that other facial features will be included in the creation of the expressions. One such facial feature is the eyes.

To further improve the speed of this work, a new lips region extraction method that does not include searching for the face position should be used.

(a)

(b)

(c)

Fig 3. (a) Expressionless face and (b) resulting smiling face created (c) Disgusted face

REFERENCES

[1] M. J. Swain and D. H. Ballard, Color indexing, International Journal of Computer Vision, Vol. 7, pp. 11-32. 1991.

[2] S. Ravela and A. Hanson, On multi-scale differential features for face recognition, Vision Interface, 2001.
[3] K. Tieu and P. Viola, Boosting image retrieval, IEEE Conference on Computer Vision and Pattern Recognition, pp. 228-235, 2000.
[4] A. Pentland, B. Moghaddam, and T. Starner, View-based and modular eigenspaces for face recognition, IEEE Conference on Computer Vision and Pattern Recognition, pp. 84-91, 1994.
[5] C. Carson, M. Thomas, S. Belongie, J. M. Hellerstein, and J. Malik, Blobworld: A system for region-based image indexing and retrieval, Third Intl. Conf. on Visual Information Systems, pp. 509-516, 1999.
[6] J. Jeon, V. Lavrenko, and R. Manmatha, Automatic image annotation and retrieval using cross-media relevance models, 26th Intl. ACM SIGIR Conf., 2003.
[7] P. Silapachote, D. R. Karuppiah, and A. R. Hanson, Feature Selection Using Adaboost For Face Expression Recognition, Proceedings of the Fourth IASTED International Conference on Visualization, Imaging, and Image Processing, pp. 84-89, 2004.
[8] Kah-Kay Sung: Learning and example selection for object and pattern recognition, PhD Thesis, MIT AI Lab, 1996.
[9] M. Ishikawa: Structure learning with forgetting, neural networks, Vol. 9, No. 3, pp 509- 521, 1993.
[10] S. Karungaru, M. Fukumi and N. Akamatsu, “Detection of human face in visual scenes.” Proc of ANZIIS, pp.165-170, 2001.
[11] K. Plataniotis and A. Venetsanopoulos, Color image processing and applications. Springer, Ch.1, 2000.
[12] A.M. Martinez and R. Benavente. “The AR Face Database.” CVC Technical Report No.24, June 1998.