Lip Classification Using Fourier Descriptors

Zeliang Zhang, Shaocheng Song* , Ming Ma, Qian Xu

Beihua University, Institute of Information Technology and Media, Jilin, Jilin, 132013, China

***Abstract* —The classification of human lip shapes is very important for lip reading recognition. It provides a definite base in order to be sure of each status during lip reading recognition. This paper presents a method to classify lip shapes based on Fourier Descriptor. AdaBoost classification is used to localize the lip region. Firstly, lip contour is extracted using Canny's edge detection. Then, in order to find the best features of lip's shape, we use Fourier Descriptor to extract important features. Finally, we classify input lip features into one of the lip categories. Experimental results show that the system accuracy rate is 85%. Therefore, this system, which using Fourier Descriptor features, presents good results and advantages for lip classification.**

Keywords - Fourier Edge Descriptor; lip detection; lip classification; canny operator

I. INTRODUCTION

The classification of the lip shapes is very important for lip reading recognition. It provides a definite base in order to be sure of each status during lip reading recognition. At present, there are not many domestic and foreign studies on lip classification. In 1968, Fisher proposed the concept of Viseme to find the minimum division unit of voice in the visual sense. Some scholars put forward that image sequence may be classified directly. For instance, Professor Yao Hongxun of Harbin Institute of Technology proposed a clustering method to classify image sequence [1]; other scholars finished lip classification by using feature vector sequence to match hidden Markov model states directly [2]. This paper presents a method to classify lip shapes based on Fourier Descriptor and realizes a lip classification system, as shown in Fig. 1.

II. FOURIER LIP CLASSIFICATION PRETREATMENT

Before lip classification, we need to do some pretreatments, which mainly include face detection and lip localization.

Face detection has obtained rather good development in recent years. Face detection accuracy is mainly influenced by position, structure, expression and influence of shelter and direction. Detection methods include position-based method, feature invariance method, template matching method and statistical model method [3]. Since face detection is mainly the inspection using geometry texture information of the face, which does not take into account of image color difference, we must make some process to the input color image to avoid misjudgment of the face detection. We adopted Gaussian filter and histogram equalization to smooth the images, which could not only make the image look suppler, but also remove noise in the image [4].

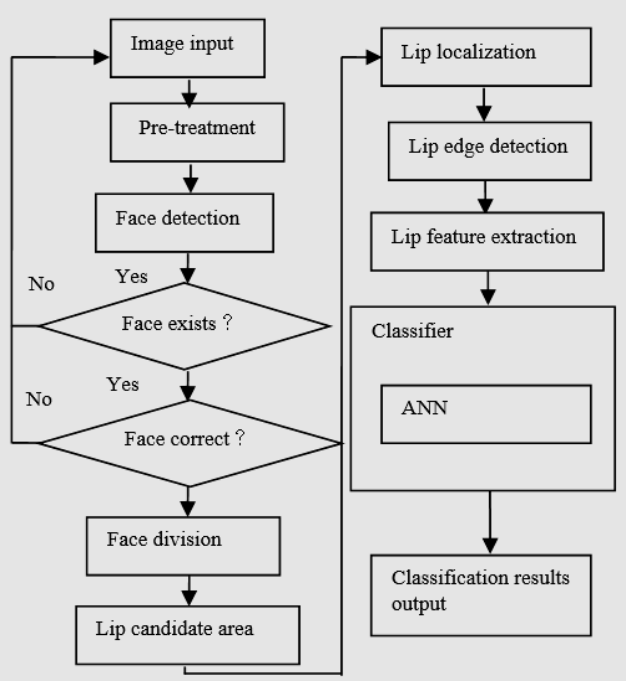

Figure 1. System flow chart

When we input the images, single color image is converted to gray image via color space conversion and noise in the image is removed by Gaussian filter [5]. Gaussian smooth mathematical equation is as Equation (1):

$$G(x,y)=\frac{1}{2\pi\sigma^2}\exp[-\frac{x^2+y^2}{2\sigma^2}] \quad (1)$$

Gaussian filter is a smoothing filter and smoothing procedure is controlled by Gaussian distribution standard σ. The greater of the value σ, the higher of the degree of smoothing. When the mean value of Gaussian distribution equals 0 and σ is 1, we can get the following Gaussian distribution chart as Fig. 2.

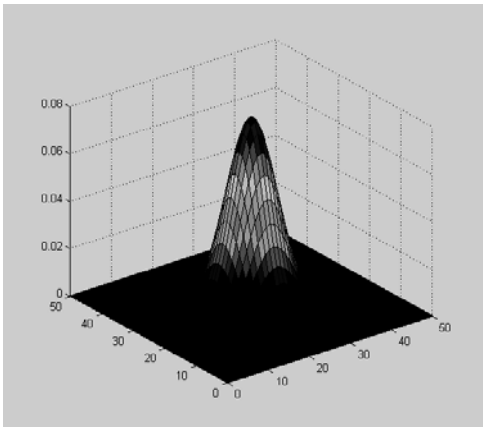

Figure 2. Gaussian distribution chart

The use of histogram equalization makes the overall image distribution relatively uniform, not too bright or too dark. The aim of histogram equalization is to use probability to make the image pixel values uniformly distributed in a certain range, histogram equalization is as Equation (2):

$$S_k = \sum_{j=0}^{k} \frac{n_j}{n_{total}} \tag{2}$$

Wherein, S_k is the gray value after equalization, n_{total} is the total pixel value, n_j is the current total amount of pixels, $k = 0,1,2,...,L-1$, L is the total amount of gray images, usually 256. Fig. 3 is the sketch map of lip image after histogram equalization. As can be seen from the figure, the overall contrast of the image increased a lot and the whole image is also brighter.

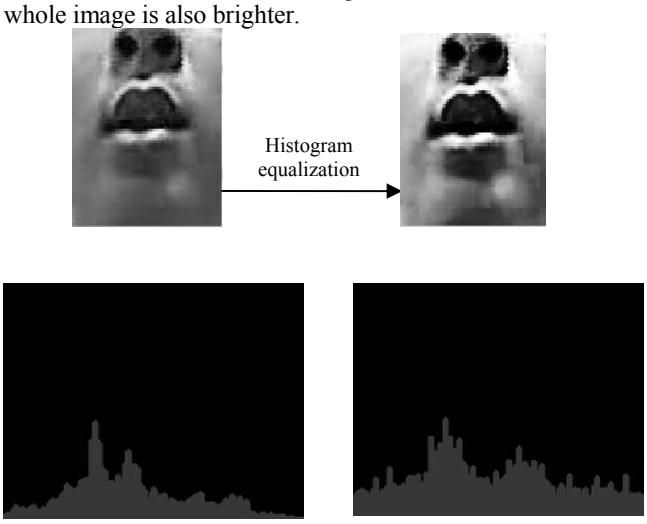

Figure 3. Histogram equalization

After the above pretreatment, we can detect the face. Detection method is based on AdaBoost algorithm and the Haar-Like rectangle feature [6]. Meanwhile this method can be applied to lip detection. The only difference is that the image used in lip detection is the lower part of the image of the frame selected region after face detection, as in reference [7].

III. FOURIER EDGE DESCRIPTOR

Fourier edge descriptor is a method for describing the edge. A series of Fourier coefficients is used to represent the shape feature of closed curve [8], it is only suitable for single closed curve, not composite closed curve, as shown in Fig. 4.

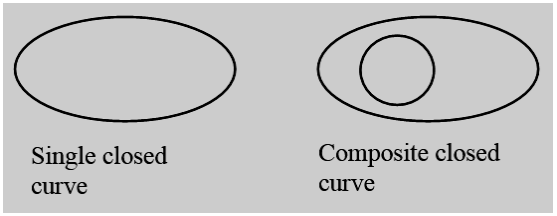

Figure 4. Single closed curve and composite closed curve.

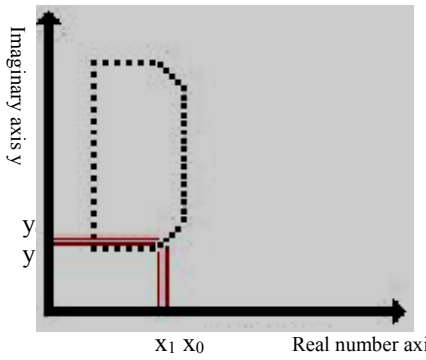

Figure 5. A digital boundary and its complex sequence indicates.

Fig. 5 shows the coordinates of the N points on the xy plane , It begins at any point (x_0, y_0) of the coordinates, and while scanning the border, moves along the curve in a certain direction, for example in a clockwise direction, passes the coordinates $(x_0, y_0), (x_1, y_1), (x_2, y_2), ..., (x_{n-1}, y_{n-1})$, then moves back to the original position. This movement is cyclical. $x(k) = x_k$和$y(k) = y_k$ can be used to express these coordinates, thus these edges can be represented as coordinates sequence s(k)=[x(k), y(k)],k=0,1,2,…,N-1, and each coordinate point can be treated as a complex number, as in Equation (3).

$$s(k) = x(k) + y(k), k = 0,1,2,..., N-1 \tag{3}$$

In Equation (3), the X sequence is the real part of the complex coordinates and Y sequence is the imaginary part of the complex coordinates. The discrete Fourier transformation of $s(k)$ can be expressed as Equation (4).

$$a(u) = \frac{1}{N} \sum_{k=0}^{N-1} s(k) \exp^{-j2\pi uk/N}, u = 0,1,2,..., N-1; j = \sqrt{-1} \tag{4}$$

The complex coefficient $a(u)$ get from Equation (4) is known as Fourier Edge descriptor. The formula for the inverse Fourier transforms $a(u)$ into $s(k)$ is as follows:

$$s(k) = \sum_{u=0}^{N-1} a(u) \exp^{-j2\pi uk/N}, k = 0,1,2,..., N-1; j = \sqrt{-1} \quad (5)$$

Assume that when we make inverse transformation, without using the entire coefficients, we merely use the first M coefficients, that is to say, in Equation (5), when $u > M-1$, and make $a(u) = 0$, the result of the formula is as Equation (6):

$$\hat{s}(k) = \sum_{u=0}^{M-1} a(u) \exp^{-j2\pi uk/N}, k = 0,1,2,..., N-1; j = \sqrt{-1} \quad (6)$$

Although we only use the first M coefficients to reconstruct each component of $\hat{s}(k)$, K is still from 0 to N-1. Therefore, when we redraw coordinate points, the same coordinate points exist. After the Fourier conversion, high-frequency components represent the details of the object, while the low-frequency components represent the overall shape of the object, so when we use less M coefficients to reconstruct the components of $\hat{s}(k)$, the edge loses more details.

Take a square as an example. We use different M coefficients to reconstruct the square, when the value of M is small (M<8), the reconstructed shape is similar to a circle, and when the value of M is large (M>48), the reconstructed shape is close to a square. By using Fourier descriptor we can use only a small amount of coefficients to get a general edge shape, but if angles or more accurate edge shape is needed, more higher order terms must be used.

IV. THE RREALIZATION OF LIP CLASSIFICATION

Fig. 6 is a more refined flow chart of lip classification system. The input is a rectangular area contains lips and the output is the classified result.

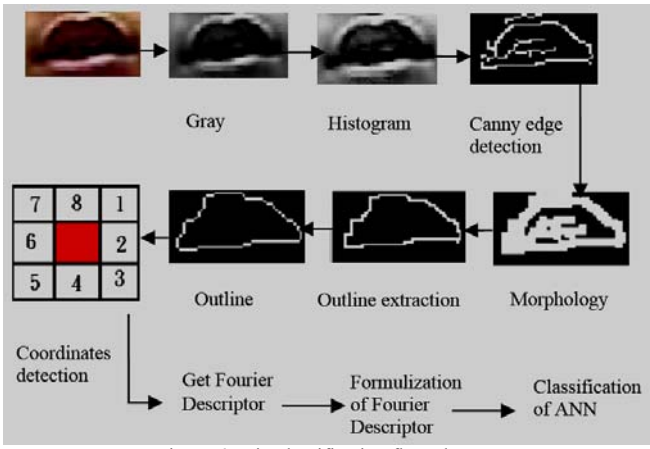

Figure 6. Lip classification flow chart.

The lips we get from figure 1 can not be used as features, because of the limited position and size information the lips contain. We need more information, especially the shape of lips. We can get the edge information after the processes of gray-scale equalization, canny operator and binaryzation. The edge information is quite messy, morphology is used to remove the noise and segment independent image or combine adjacent elements in the image. After that the contour is extracted. A too thick contour line is not needed, which will disturb the obtainment of the coordinate point. Therefore in order to ensure maximum pixel connectivity between eight directions is not more than three, refining is needed. Then obtain the leftmost point position on the image, starting from this position, with eight connected way, get all coordinate points clockwise, as shown in Fig. 7.

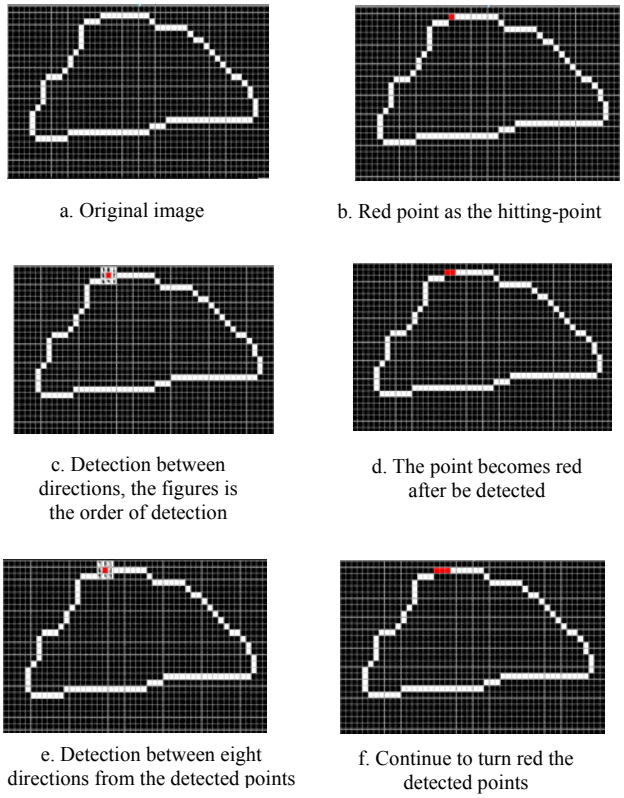

Figure 7. Obtaining lip contour coordinate points

After collecting the ordered coordinate points, the Fourier transformation starts. The numerical obtained from the transformation are called Fourier Descriptors, which are used as features of lip classification.

Figure 8. Lip categories.

The classifier used is artificial neural network (ANN) [9]. ANN is one of the best classifier at present. It is slow in training phase, because it uses gradient descent method to adjust the weights between the neurons to minimize the error, but it is rapid in the phase of classification or recognition. In the experiment, the artificial neural network used belongs to the back propagation neural network, which consists of 2 hidden layers and each layer has 100 nodes. The input layer nodes are decided by the lip feature dimension, and output layer consists of 7 nodes.

V. EXPERIMENT RESULTS AND ANALYSIS

In this experiment, 200 images are chosen, including 160 AVSP images [10] and 40 captured images by the author of this paper. The lips on these images are divided into seven categories: normal lip, closed lip, slightly open lip, open not grinning lip, half open with lower tooth lip, open with lower tooth lip and pout lip. In training phase the 160 AVSP images are used. The changes of the lips are extracted and trained to be the classifier. In the classification phase, 40 captured images are used. We implemented lip classification tests on the 40 images by classifier, and divided the final testing results into 7 classes: the first class (normal), the second class (closed), the third class (slightly open), the fourth class (open not grinning), the fifth class (half open with lower tooth), the sixth class (open with lower tooth) and the seventh class (pout).

After the lip contour is extracted from the image, the next phase is the training movement by Fourier Descriptor [11]. In the artificial neural network, the input and output are known. The main task is training the weight value. When the weight value is determined, lip classifier is completed and can be applied to other unclassified images.

We use the amount 10, 20, 30 Fourier descriptors for classification and comparison [12]. Table 1 is classification result of training samples taking 10, 20, 30 Fourier descriptors. Table 2, 3 and 4 are the classification results of training samples taking 10, 20, 30 Fourier descriptors by using different classifiers [13].

TABLE 1. THE RESULTS (TRAINING PHASE) TO A FOURIER DESCRIPTORS 10,20,30 CLASSIFICATION

	The 1st class	The 2nd class	The 3rd class	The 4th class	The 5th class	The 6th class	The 7th class
A normal	38	0	0	0	0	0	0
B closed	0	19	0	0	0	0	0
C slightly open	0	0	19	0	0	0	0
D open not grinning	0	0	0	19	0	0	0
E half open with lower tooth	0	0	0	0	19	0	0
F open with lower tooth	0	0	0	0	0	27	0
G pout	0	0	0	0	0	0	19

TABLE 2. THE TEST RESULTS (CLASSIFICATION PHASE) TO A FOURIER DESCRIPTORS 10 CLASSIFICATION

	The 1st class	The 2nd class	The 3rd class	The 4th class	The 5th class	The 6th class	The 7th class
A normal	8	0	0	0	0	0	0
B closed	0	5	0	0	0	0	0
C slightly open	0	0	5	0	0	0	0
D open not grinning	0	0	0	5	0	0	0
E half open with lower tooth	0	0	0	0	6	0	0
F open with lower tooth	2	0	0	0	0	3	1
G pout	0	0	0	0	0	0	5

In Table 2, we can see that the misjudgment rate reached 50% in the open lower tooth category. Through careful analysis we find the main reason. In the morphological process, due to the edges on both sides are thin, when the edge is removed, only the upper and lower lips are left, therefore, this category may be mistaken for normal lip and pout lip.

In Table 3, we can also see that the open with lower tooth category exists classification error. After the analysis, we find that this is because the error of morphology. However, Fourier descriptors increase from 10 to 20, which makes the parameters of neurons in the network be corrected to some extent, so the misjudgment is less than that in table 2.

TABLE 3. THE TEST RESULTS (CLASSIFICATION PHASE) TO A FOURIER DESCRIPTORS 20 CLASSIFICATION

	The 1st class	The 2nd class	The 3rd class	The 4th class	The 5th class	The 6th class	The 7th class
A normal	8	0	0	0	0	0	0
B closed	0	5	0	0	0	0	0
C slightly open	0	0	5	0	0	0	0
D open not grinning	0	0	0	5	0	0	0
E half open with lower tooth	0	0	0	0	6	0	0
F open with lower tooth	0	0	0	0	0	4	2
G pout	0	0	0	0	0	0	5

TABLE 4. THE TEST RESULTS (CLASSIFICATION PHASE) TO A FOURIER DESCRIPTORS 30 CLASSIFICATION

	The 1st class	The 2nd class	The 3rd class	The 4th class	The 5th class	The 6th class	The 7th class
A normal	7	0	0	0	1	0	0
B closed	0	5	0	0	0	0	0
C slightly open	0	0	5	0	0	0	0
D open not grinning	0	0	0	5	0	0	0
E half open with lower tooth	0	0	0	0	6	0	0
F open with lower tooth	4	0	0	0	1	1	0
G pout	0	0	0	0	0	0	5

In Table 4, there is one normal lip category be misidentified to the half open with lower tooth category. It is not a serious problem; the most serious problem is that more open with lower teeth lip category be misidentified to be normal lip category. After careful analysis, we can find that this is the sample problem. Chins in some samples were cut off, thus in the morphological process, the lower lips disappeared. As a result these samples can only be classified to normal lip category.

The experimental results showed that the accuracy rate of 10, 20, 30 Fourier descriptors for classification were: 92.5%, 95% and 85%. Open with lower teeth lip category exposed a lot of problems. The reason is that the images underwent the same treatment procedure, which ignored the fact that some images required different treatments. Thus some necessary information of the open with lower teeth lip category was removed before recognition.

In the study, AVSP image trained classifier is used to the images captured, the accuracy rate is not high, but the accuracy rate of 95% is fairly good, which proves that the application of Fourier descriptor to lip classification is feasible.

VI. CONCLUSIONS

This paper presents a lip classification method based on Fourier descriptor. The input images undergo face detection and lip localization, then the spatial domain is converted into frequency domain by using Fourier descriptor, finally lip classifier is established by using artificial neural network. The accuracy rate of more than 30 Fourier descriptor is generally low, therefore it is not discussed here. When Fourier descriptor is 20, the accuracy rate is better, but the data amount of training and classification still need to be increased. For future studies, put multiple lip classification results together by combining time sequence and form meaningful lip reading action will be needed.

ACKNOWLEDGMENTS

This work was financially supported by Jilin Scientific and Technological Development Program (20140101185JC, 20140101206JC-16), Jilin City Scientific and Technological Development Program (201464048), Scientific and Technological Research Program of Jilin Educational Committee during the "12th Five-Year Plan" (2015-149, 2015-137) and National Social Science Foundation (15BTQ02). The authors wish to thank my colleagues, they are Sun Yan, Bi Xinwen.

REFERENCES

[1] Chai Xiujuan, Yao Hongxun, Gao Wen, "Basic Lip Classification in Lip-reading Recognition," Computer Science, vol. 29, pp. 130-133, 2002.

[2] Zeliang Zhang, Wenliang Qu, "Review of the Lip-reading Recognition," ICSESS 2014, pp. 593-596, 2014.

[3] Tomoaki Yoshinaga, Satoshi Tamura, "Audio-Visual Speech Recognition Using Lip Information Extracted from Side-Face Images," EURASIP Journal on Audio, Speech, and Music Processing, 2007.

[4] Zhang Zeliang, Li Xiongfei, Yang Chengjia, "An effective parameter estimation algorithm of the visual language features," International Journal of Digital Content Technology and its Applications, vol. 6, pp. 69-76, 2012.

[5] Xu GuoQing, Mu ZhiChun, "Fourier shape descriptor based on multi-level triangular area functions," Journal of University of Science and Technology Beijing, vol. 35, pp. 1201-1206, 2013.

[6] Bear, Helen L., et al, "Which phoneme-to-viseme maps best improve visual-only computer lip-reading," Advances in Visual Computing: Proceedings, International Symposium, vol. 8888, pp. 230-239, 2014.

[7] F. J. Huang, T. Chen, "Real-Time Lip-Synch Face Animation driven by human voice," IEEE Second Workshop on Multimedia Signal Processing, pp. 352-357, 1998.

[8] Zeliang Zhang, Fumei Liu, "Review of the Visual Feature Extraction Research," ICSESS 2014, pp. 449-452, 2014.

[9] Kavousi-Fard, Abdollah, "A new fuzzy-based feature selection and hybrid TLA-ANN modelling for short-term load forecasting," Journal of Experimental and Theoretical Artificial Intelligence, vol. 25, pp. 543-557, 2013.

[10] Zhang Zeliang, Li Xiongfei, Yang Chengjia, "Visual Speech Recognition based on Improved type of Hidden Markov Model," Journal of Convergence Information Technology, vol. 7, pp. 119-126, 2012.

[11] Keetels, M, M. Pecoraro, J. Vroomen, "Recalibration of auditory phonemes by lipread speech is ear-specific," Cognition, vol. 141, pp. 121-126, 2015.

[12] Strand, J., Cooperman, A., Rowe, "Individual differences in susceptibility to the mcgurk effect: links with lipreading and detecting audiovisual incongruity," Journal of Speech Language & Hearing Research, vol. 57, pp. 2322-2331, 2014.

[13] MZ Ibrahim, DJ Mulvaney, "Geometrical-based lip-reading using template probabilistic multi-dimension dynamic time warping," Journal of Visual Communication & Image Representation, vol. 30, pp. 219-233, 2015.