A Multi-Scale Lip Segmentation Method Based on Markov Random Field

Yuanyao Lu*, Xiaoshan Zhu and Qingqing Liu

School of Electronic and Information Engineering, North China University of Technology, No. 5, Jinyuanzhuang Road,
Shijingshan District, Beijing, China, 100144

*Corresponding author

***Abstract*—In order to perform lip segmentation in a lip-reading system, we propose a new method based on maximum a posterior and Markov random field (MAP-MRF) framework to statistically model observed images of different texture areas in this paper. First, we establish a multi-scale model to capture the characteristics of each sub-region in the image and use the tree structure in the wavelet domain to calculate the probability of tree nodes at different scales. Thus, the number of layer can be considered as one segment cluster. Then, we utilize MRF to translate the lip segmentation problem into labeling optimization issue. Finally, the Bayesian criteria and the extended expectation maximum (EM) algorithm are applied to estimate child node parameters. The experimental results of this method are more robust than the traditional iterative condition model (ICM).**

***Keywords*—lip segmentation; MAP-MRF framework; multi-scale model; wavelet domain**

I. INTRODUCTION

Lip-reading, also known as visual speech recognition (VSR), is an intelligent and attractive human-machine interaction (HCI) technique that can understand speech content by interpreting the speaker's lip movement when speech is degraded or unavailable due to noise or crosstalk. With a wide range of applications such as speaker recognition, video surveillance, behavior analysis, and virtual face animation, lip reading has drawn the attention of the HCI, pattern recognition (PR), and artificial intelligence (AI) communities [1-5]. A typical lip-reading system includes four stages: lip positioning, lip segmentation, lip feature extraction, and recognition, as shown in Figure 1.

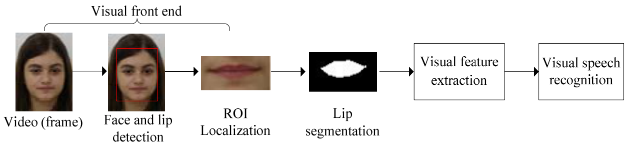

FIGURE I. SCHEMATIC DIAGRAM OF LIP READING SYSTEM

Lip segmentation is an important part of a lip-reading system because the accuracy of the results has a decisive influence on the subsequent recognition rate [6]. Literature [7-9] proposed a method based on shape templates, including several complex methods such as active contour model (ACM), active shape model (ASM) and active appearance model (AAM). In addition, the literature [10] proposed a novel multi-type shape guided fuzzy clustering algorithm that can successfully segment the lip region with teeth and whiskers. Recently, MRF theory is often used in conjunction with statistical decision-making and estimation theory to develop an objective function based on principles to support optimality. The maximum posterior probability [11] is one of the most sensible proofs for optimization and flow selection in MRF modeling.

In this paper, we propose an algorithm for multi-scale analysis of images. Characteristics and multi-scale quad-tree structure model in the wavelet domain by describing MRF for each node over the multi-resolution lattice set. In this model, it is possible to obtain correlation over a set of neighbors of increasing size by multi-resolution representations for the region image associated to each node.

II. LIP LOCALIZATION

In a lip-reading system, the lip area needs to be gained before the lip segmentation. In this paper, we use a face detector that combines the AdaBoost algorithm proposed by Viola and Jones [12] with Haar-like features to detect and locate human faces. This method has good performance and has become the mainstream method of face detection. After detecting the face area, based on the physical structure of the face and the special proportion of the lips on the face, we define the lower 1/3 area of the face as the area of the lip [13]. At this point, we have the result of rough positioning of the lips. The process is shown in Figure 2.

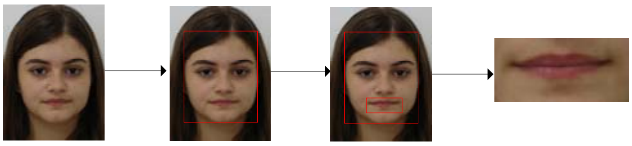

FIGURE II. FACE AND LIP DETECTION

III. THE PROPOSED LIP SEGMENTATION ALGORITHM

In order to capture the characteristics of each sub-region of the image, taking into account the large homogeneous region and the boundary region containing details, the scale characteristics of the image should be comprehensively used. Using a set of different size classification windows, fine-scale accurate segmentation is achieved. Given an original image *x*, the image size is $2^j \times 2^j$ and the number of pixels is $n=2^{2j}$. Divide the image into four sub-images of equal size and follow this pattern to get a binary image data block. Since these four subimage blocks are embedded in their parent blocks at the next coarse scale, these image blocks have a quadtree structure, as shown in Figure 3, and each node on the quadtree corresponds to a binary image data block. A block of data on scale *j* is denoted by d^{j}, then d^{0} is the root of the tree representing the entire image *x*, and d^{j-1} is the leaf representing a single pixel. Given a random field image *X*, the binarized image data block is also a random field, denoted as D^{j}.

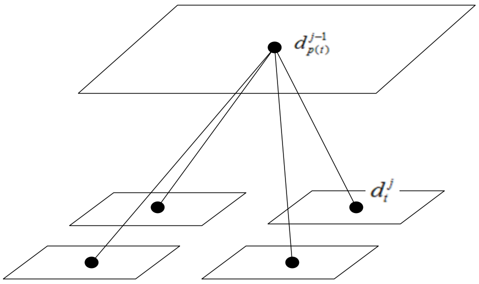

FIGURE III. QUAD-TREE STRUCTURE

Assuming that the three sub-ands of the wavelet transform are independent of each other and the parameters follow Gaussian mixture distribution, then the joint probability density function of the model parameter *M* = {Θ^{LH}, Θ^{HL}, Θ^{HH}}is denoted as follows:

$$f(w \mid M_{model}) = f(w^{LH} \mid \Theta^{LH}) \times f(w^{HL} \mid \Theta^{HL}) f(w^{HH} \mid \Theta^{HH}) \tag{1}$$

where the Θ = {μ, σ2}and the Gaussian distribution is represented as:

$$g(x,\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma}} \exp\{-\frac{(x-\mu)^2}{2\sigma^2}\} \tag{2}$$

Therefore, the edge distribution of wavelet coefficients is:

$$f(w_t) = \sum_m p_{S_t}(m) g(w_t, \mu_{t,m}, \sigma_{t,m}^2) \tag{3}$$

Calculate the joint probability function according to the Bayesian criteria $p(x_s = t, T_{p(s)/s} \mid \Theta)$ for each child node *t* in the scale *j* as a probability mass function of *X$_s$*:

$$p(x_s = t \mid W, \Theta) = \frac{\alpha(x_s = t)\beta(x_s = t)}{\sum_{t=0}^{J} \alpha(x_s = t)\beta(x_s = t)} \tag{4}$$

$\beta(x_s) = p(T_s \mid x_s = t, \Theta)$, $\alpha(x_s) = p(x_s = t, T_{p(s)/s} \mid \Theta)$ is the probability mass function of each node in scale *j*. In addition, the expected joint probability distribution between the hidden state of the child node of the wavelet domain quadtree and the implicit state of the parent node *P(s)* can be given by:

$$p(x_s = t, x_{p(s)} = j \mid W, \Theta) = \frac{\beta(x_s)\varsigma_{x_s=t}^{x_{p(s)}=j} \alpha(x_{p(s)}\beta(x_{p(s)/s}))}{\sum_{t=0}^{J} \alpha(x_s = t)\beta(x_s = t)} \tag{5}$$

In the above formula, $\varsigma_{x_s=t}^{x_{p(s)}=j}$ is state transition probabilities between the hidden states of child and parent nodes. After obtaining the joint distribution of the implicit state of the single node and the implicit expectation of the parent-child node, the class-node assignment that maximizes the expectation is calculated according to the maximum a posteriori (MAP) probability criterion:

$$\Theta^{k+1} = \arg\max_{\Theta} E[\ln p(W, X \mid \Theta), W, \Theta^k] \tag{6}$$

Then use the improved maximum expectation (EM) criterion to find parameter estimate, the mathematical expectation of observational random field *W* is:

$$\begin{aligned} W &= -E_w[\ln p(w) \mid x, \Theta)] \\ &= -\sum_w p(w) \ln p(w) \ln p(w \mid x, \Theta) \end{aligned} \tag{7}$$

For the E-step of EM algorithm, the expectation of model parameter is represented as:

$$w_{pk}^{(t)} = \frac{[c_k^{(t)} p(w_p \mid \mu_k^{(t)}, \sigma_k^{2(t)})]^k}{\sum_{k=1}^{k} [c_k^{(t)} p(w_p \mid \mu_k^{(t)}, \sigma_k^{2(t)})]^k} \tag{8}$$

where *k* is the number of iterations. For the M-step of EM algorithm, we update model parameters:

$$c_k^{(k+1)} = \sum_{p=1}^{L} w_{pk}^{(k)} \mu_x(w_p) / mJ \tag{9}$$

In the above formula, m is the set of child nodes with the label m on the scale J, and c(0) is the initial estimate value. The parameters of mean and variance are given by:

$$\mu_k^{(t+1)} = \sum_{p=1}^{L} w_{pk}^{(t)} \mu_x(w_p) w_p \Big/ \sum_{p=1}^{L} w_{pk}^{(t)} \mu_x(w_p) \tag{10}$$

$$\sigma_k^{2(t+1)} = \frac{\sum_{p=1}^{L} w_{pk}^{(t)} \mu_x(w_p)[w_p - \mu_k^{(t+1)}]^2}{\sum_{p=1}^{L} w_{pk}^{(t)} \mu_x(w_p)} \qquad (11)$$

where

$$\mu_x(w_p) = \sum_{v \in s} \Delta(x_v, w_p) \qquad (12)$$

$$\Delta(x_v, w_p) = \begin{cases} 1 & if\,(x_v = w_p) \\ 0 & otherwise \end{cases} \qquad (13)$$

The maximum expectation algorithm is iterated until $\| \Theta^{k+1} - \Theta^k \| \leq \xi$. ξ represents the iterative termination conditions of maximum expected iterative algorithm, and $\| \cdot \|$ is even-Liard-norm.

IV. Experimental Results

In this paper, we test the proposed method using 200 images from AR Face Database [14] in the MATLAB R2013a environment. And we compare the proposed method with ICM, the results reveal that our algorithm is more accurate, as shown in Figure 4.

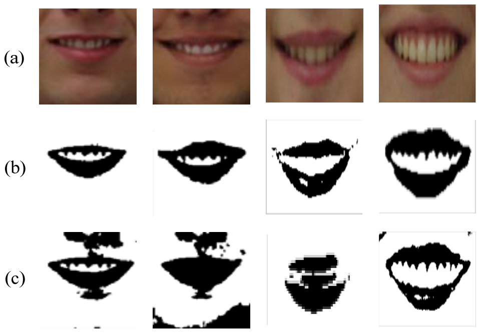

FIGURE IV. LIP SEGMENTATION RESULTS (A)THE ORIGINAL IMAGES (B) RESULTS OF THE PROPOSED METHOD (C) RESULTS OF THE ICM METHOD

To evaluate the performance of the algorithms, we use two criterions quoted in [15]. The first measure determines the percentage of overlap (OL) between the segmented lip region L_1 and the ground truth L_2:

$$OL = \frac{2(L_1 \cap L_2)}{L_1 + L_2} \times 100\% \qquad (14)$$

The second criterion is segmentation error (SE) defined as

$$SE = \frac{OLE + ILE}{2 \times TL} \times 100\% \qquad (15)$$

where *OLE* is the number of non-lip pixels classified as lip pixels, *ILE* is the number of lip-pixels classified as non-lip ones, and *TL* denotes the number of lip-pixels in the ground truth. We show the mean values of OL and SE of each group in Table 1. The OL of our method is higher than that of the traditional method, and SE is lower.

TABLE I. OL AND SE OF THE PROPOSED METHOD AND TRADITIONAL ICM METHOD

	Data	1	2	3	4
Proposed method	OL (%)	87.80	81.20	88.60	84.12
	SE (%)	11.70	34.71	11.60	21.01
Traditional ICM	OL (%)	84.21	79.50	87.33	83.41
	SE (%)	15.23	41.20	16.10	23.50

V. Conclusion

In this paper, we proposed a new method which is based on MAP-MRF framework to implement lip segmentation. We formulate the segmentation as a labeling optimization problem. The EM algorithm was used to estimate the node parameters. We have compared the proposed method with traditional ICM method, it can be concluded from the experiment results that the proposed method has better segmentation accuracy.

Acknowledgment

The research was supported by the National Natural Science Foundation of China (61571013), by the Beijing Natural Science Foundation of China (4143061), by the Science and Technology Development Program of Beijing Municipal Education Commission (KM201710009003) and by the Great Wall Scholar Reserved Talent Program of North China University of Technology (NCUT2017XN018013). The authors thank all the partners and the participants in the experiment for their help.

References

[1] B. S., Lin, Y. H., Yao, and C. F., Liu, “Development of novel lip-reading recognition algorithm”, IEEE Journals & magazines, vol. 5, pp.794-801, 2017.

[2] V. Gatteschi, F. Lamberti, and P. Montuschi, “Semantics-based intelligent human-computer interaction”, IEEE Intelligent Systems, vol. 31, pp.11-21, 2016.

[3] A. H. Abdelaziz, S. Zeiler, and D. Kolossa, “Learning dynamic stream weights for coupled-HMM-based Audio-visual speech recognition”, IEEE/ACM Transactions on Audio, speech, and Language Processing, vol. 23, pp. 863-876, 2015.

[4] O. Costilla-Reyes, P. Scully, and K. B. Ozanyan, “Temporal pattern recognition in gait activities recorded with a footprint imaging sensor system”, IEEE Sensors Journal, vol.16, pp. 8815-8822, 2016.

[5] R. H. Kulkarni, and P. Padmanabham, “Integration of artificial intelligence activities in software development processes and measuring effectiveness of integration”, IET Journals & Magazines, vol.11, pp.18-26, 2017.

[6] N. Eveno, A. Caplier, and P. Y. Coulon, “Accurate and quasi-automatic lip tracking”, IEEE Transactions on Circuits and Systems for Video Technology, vol. 14, pp. 706-715, 2004.

[7] N. Eveno, A. Caplier, and P. Y. Coulon, “Jumping snakes and parametric model for lip segmentation”, in Proc. Int. Conf. Image Process., Barcelona, Spain, pp. 867–870, Sep. 2003.

[8] H. Seyedarabi, W. Lee, and A. Aghagolzadeh, “Automatic lip tracking and action units classification using two-step active contours and probabilistic neural networks”, in Proc. Can. Conf. Elect. Comput. Eng., Ottawa, Canada, pp. 2021–2024, May 2006.

[9] Z. Zheng, J. Jiong, D. Chunjiang, X. H. Liu, and J. Yang, “Facial feature localization based on an improved active shape model”, Inform. Sci., vol. 178, pp. 2215–2223, May 2008.

[10] A. W. C. Liew, S. H. Leung, and W. H. Lau, “Segmentation of color lip images by spatial fuzzy clustering”, IEEE Trans. Fuzzy Syst., vol. 11, pp. 542–549, Aug 2003.

[11] X. Dong, and Y. Zhang, “SAR image reconstruction from undersampled raw data using maximum a posteriori estimation”. IEEE Journal of Selected Topics in Applied Earth Observation and Remote Sensing, vol. 8, pp. 1651-1664, 2015.

[12] M. J. Viola and P. Jones, “Robust real-time object detection”, IEEE Transactions on Computer Vision, vol.57, pp. 137-154, 2004.

[13] D. Wu, and Q. Ruan, “Lip reading based on cascade feature extraction and HMM”, ICSP Proceedings, pp. 1306-1310. 2014.

[14] A.M. Martinez and R. Benavente. The AR Face Database. CVC Technical Report #24, June 1998.

[15] A. Liew, S. Leung, and W. Lau, “Segmentation of color lip images by spatial fuzzy clustering”, IEEE Trans. Fuzzy Syst., vol. 11, 2003.