Lip Segmentation under MAP-MRF Framework with Automatic Selection of Local Observation Scale and Number of Segments

Yiu-ming Cheung, *Senior Member, IEEE*, Meng Li, Xiaochun Cao, and Xinge You

***Abstract*— This paper addresses the problem of segmenting lip region from frontal human face image. Supposing each pixel of the target image has an optimal local scale from the segmentation viewpoint, we treat the lip segmentation problem as a combination of observation scale selection and observed data classification. Accordingly, we propose a hierarchical multiscale Markov random field (MRF) model to represent the membership map of each input pixel to a specific segment and local-scale map simultaneously. Subsequently, lip segmentation can be formulated as an optimal problem in the maximum *a posteriori* (MAP)-MRF framework. Then, we present a rival-penalized iterative algorithm to implement the segmentation, which is independent of the number of predefined segments. The proposed method mainly features two aspects: 1) its performance is independent of the predefined number of segments, and 2) it takes into account the local optimal observation scale for each pixel. Finally, we conduct the experiments on four benchmark databases, i.e. AR, CVL, GTAV, and VidTIMIT. Experimental results show that the proposed method is robust to the segment number that changes with a speaker's appearance, and can enhance the segmentation accuracy by taking advantage of the local optimal observation scale information.**

***Index Terms*— Lip segmentation, MAP-MRF framework, number of segments, local scale selection.**

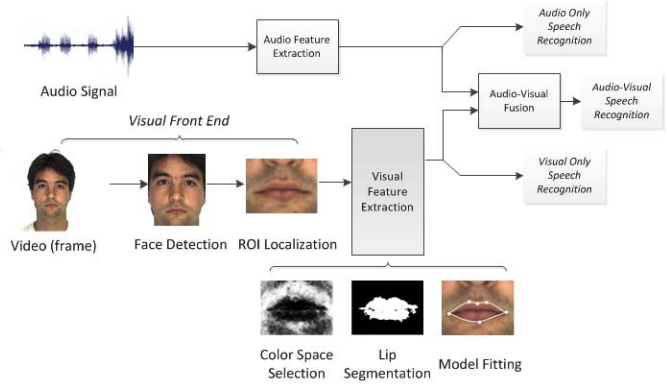


Fig. 1. The main process blocks of audio-visual speech recognition (AVSR) systems.

I. Introduction

ANALYZING the visual cues such as the movement or shape of the human lip, which is formally known as automatic lip-reading [1], has received considerable attention from the community because of its various applications such as audio-visual speech recognition (AVSR) [2], visual only speech recognition (VSR) [3], speaker identification [4], lip synchronisation [5], and so forth. All these applications have a common processing step named visual front end that mainly involves three components: face detection, region of interest (ROI) localization, and visual feature extraction as shown in Fig. 1. In general, the visual feature extraction can be further divided into three steps: color space selection, lip segmentation, and model fitting. Color space selection can be viewed as a pre-processing step that is relatively uncomplicated but necessary. Lip segmentation studies are usually conducted in several common color spaces such as HSV, $CIELUV$, $CIELAB$, and some variants such as Q channel [6], $\hat{H}$ channel [7], and LUX [8]. Subsequently, segmentation is employed to extract the lip region as a segment label map, membership map, and so forth. Then, based on the segmented lip region, the parameterized model is built via model fitting so as to obtain the low dimension parameter set (also known as feature vector in some articles) for the subsequent processes such as audio-video fusion and recognition. In this paper, we will focus on lip segmentation only as its accuracy has a major influence on the performance of the global system [9].

Thus far, a number of state-of-the-art lip segmentation methods have been proposed (Section II will give a review of related works). Nevertheless, lip segmentation is still

Manuscript received April 25, 2012; revised February 15, 2013, June 4, 2013, September 20, 2013, December 31, 2013, and June 2, 2014; accepted June 8, 2014. Date of publication June 17, 2014; date of current version July 1, 2014. This work was supported in part by the Research Grant Council of Hong Kong under Project HKBU 210309, in part by the Faculty Research Grants, Hong Kong Baptist University, Hong Kong, under Project FRG2/11-12/067 and Project FRG2/12-13/082, in part by the National Science Foundation of China under Grant 61272366, Grant 61272203, and Grant 61332012, in part by the 100 Talents Programme, Chinese Academy of Sciences, Beijing, China, in part by the International Scientific and Technological Cooperation Project under Grant 2011DFA12180, in part by the National Science and Technology Research and Development Program under Grant 2012BAK02B06, and in part by the Ph.D. Programs Foundation, Ministry of Education of China, under Grant 20110142110060. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Stefano Tubaro.

Y.-m. Cheung is with the Department of Computer Science, Institute of Computational and Theoretical Studies, Hong Kong Baptist University, Hong Kong, and also with the United International College, Beijing Normal University-Hong Kong Baptist University, Zhuhai, China (e-mail: ymc@comp.hkbu.edu.hk).

M. Li is with the Department of Computer Science, Hong Kong Baptist University, Hong Kong (e-mail: mli@comp.hkbu.edu.hk).

X. Cao is with the State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China (e-mail: caoxiaochun@iie.ac.cn).

X. You is with the Department of Electronics and Information Engineering, Huazhong University of Science and Technology, Wuhan 430074, China (e-mail: youxg@hust.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2014.2331137

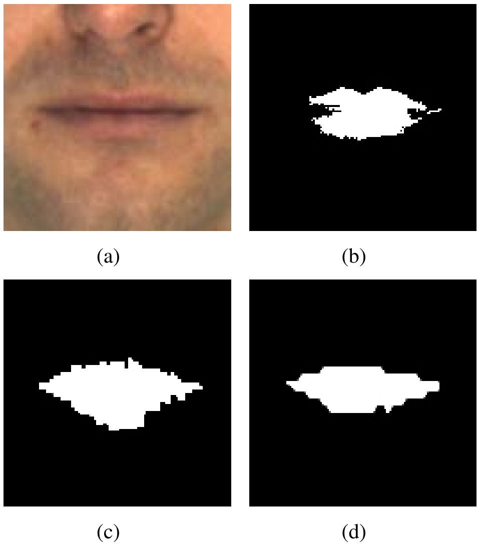


Fig. 2. Segmentation results obtained by the same method under different scales: (a) the source image, (b) the segmentation result based on the source image, (c) the segmentation result based on the sub-sampling images with $\frac{1}{2}$ source size, (d) the segmentation result based on the sub-sampling images with $\frac{1}{4}$ source size. For the purpose of comparison, all sub-sampled images are scaled to the size of source image.

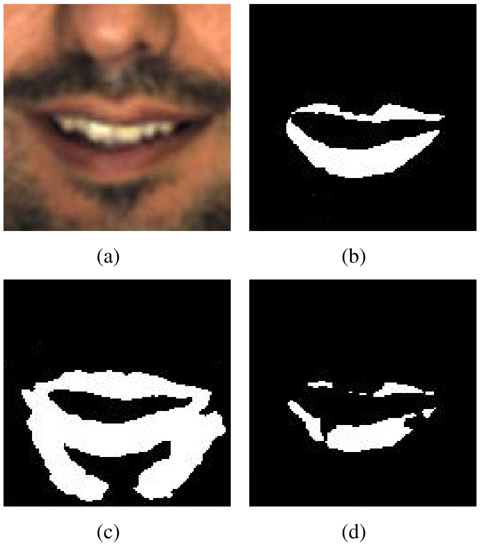


Fig. 3. (a) the source image, and (b), (c), (d) illustrate the corresponding lip segmentation results obtained by the same method with $\hat{m} = m^*$, $\hat{m} < m^*$, and $\hat{m} > m^*$, respectively, where m^* denotes the true number of segments and $\hat{m}$ is the pre-assigned number of segments.

a challenging problem involving two major issues. Firstly, almost all existing methods process the whole input image under an invariant global observation scale. However, in the real world, it is quite common that an image is composed of structures with different optimal local observation scales from the segmentation viewpoint [10]. Fig. 2 illustrates the segmentation results obtained by the method proposed in our previous work [11] under different global observation scales. Under the finest scale (see Fig. 2 (b)), the upper lip can be extracted exactly, however the parts around the mouth corners are clearly over-segmented. Although this problem can be tackled under coarser scales (see Fig. 2 (c) and (d)), the upper lip is under-segmented. Moreover, in some other cases, the over- and under-segmentation may even arise simultaneously under the same global scale. Thus, it is hard to solve the problem by simply adjusting the global observation scale.

The selection of the number of segments is another problem in lip segmentation. Almost all the existing methods require pre-assigning the number of segments (also called *number of clusters*) appropriately, or their performance may result in over- or under-segmentation. An example of the under- and over-segmentation caused by an inappropriately pre-assigned number of segments is shown in Fig. 3. In the task of lip segmentation using color cues, each of the segments is uniform with respect to specific hue characteristics such as lip region, skin region, teeth, and so forth. In practice, it is hard to determine the number of segments exactly beforehand because the appearance of speakers, e.g. mustache, teeth, and tongue, differs. To the best of our knowledge, only the methods proposed in [8] and [12] are somewhat independent of the number of segments. Nevertheless, the lip tracking scheme presented in [8] suggests that the number of segments should be set to two. In [12], the number of clusters is determined by an individual local exhaustive search before segmentation, i.e. there are redundant data traverses which is a typical shortcoming for the real time system.

In this paper, we propose a lip segmentation method to deal with the two problems stated above. We design a hierarchical multi-scale Markov random field (MRF) model, in which the MRF posed on each layer represents the membership map of each input pixel belonging to a specific segment. Then, another MRF is built to represent the local-scale map of the given image in order to introduce the influence of observation into the segmentation result. Based on the two MRFs, we obtain a three-term energy function which involves priori energy of scale, priori energy of label, and likelihood energy. Thus, the lip segmentation task with local scale consideration can be formulated as an optimal problem in the maximum a posteriori (MAP)-MRF framework. An optimal segmentation result can be achieved by minimizing the energy function.

Moreover, considering the possible redundant segments that are caused by assigning the number of segments larger than the ground truth, we will present an iterative rival-penalized algorithm featuring independence from the number of segments when implementing the segmentation. Subsequently, the pixels that fall into the redundant segment can be corrected to the true segment by using the algorithm. We will evaluate the performance of the proposed method in terms of its robustness against the number of segments, segmentation accuracy, and efficiency in comparison with the existing counterparts in four benchmark databases.

The remainder of this paper is organized as follows: Section II provides a review of related works. Section III introduces the MRF model briefly, while Section IV describes the proposed method in detail. Section V shows the experimental results. Lastly, we draw a conclusion in Section VI.

II. A Review of Related Works

Related works in the literature based on different theories and methodologies can be roughly summarized into five categories. The first is the thresholding-based method, which provides a simple way to implement lip segmentation. For instance, the lip segmentation proposed in [13] utilizes two threshold values in Q channel to obtain lip pixel candidates. The works described in [6] and [14] also employ the thresholding-based method to implement lip segmentation. Since the threshold values are determined empirically, such a method is not essentially applicable for cases with complexion difference, appearance of a moustache, and illumination variation. Although thresholding-based lip segmentation has rarely been reported in recent years, it is still commonly used in lip segmentation as a pre- or post-processing step.

The second category is the gradient-based methods that have been utilized widely to extract the lip contour [15]–[17] due to their efficacy in boundary detection. However, the accuracy of the methods is easily affected by false boundary edges.

The third category is the shape template-based methods which include several sophisticated methods such as the active contour model (ACM), the active shape model (ASM), and the active appearance model (AAM). ACM, which is also known widely as the "snake" model, is an energy-minimized deformable spline influenced by constraint (i.e. internal energy) and image forces (i.e. external energy) that pull the spline towards the object contours [18], [19]. ASM-based methods build the point distribution model to represent the shape of a class of objects (i.e. training set), and deform the model to fit an example of the object in a new image. Similarly, AAM-based methods also build a statistical model to represent an object. The model is based not only on the shape, but also on the appearance of the object [2], [20], [21]. Nevertheless, the final segmentation accuracy of such a method depends on the initial template position. Moreover, its performance is somewhat sensitive to the noise boundaries brought by the appearance of moustache and teeth.

The fourth category is the clustering-based segmentation methods. For example, fuzzy C-means (FCM) and K-means clustering-based methods have been employed to perform lip segmentation [22]–[24]. Moreover, the works described in [25]–[27] utilize statistical models, e.g. the Gaussian mixture model (GMM) and FCM, to estimate the lip membership maps as well. Nevertheless, such methods may miscalculate the membership due to similarity and overlap between the lip and non-lip pixels in color space. As a result, lip segmentation methods that depend solely on edge or color information will not deliver satisfactory performance [28]. Liew et al. [29] have therefore proposed a fuzzy clustering algorithm taking spatial restriction into account. This method considers both the distributions of data in feature space and the spatial interactions between neighboring pixels during clustering. Another fuzzy clustering based lip segmentation method proposed in [30] obtains the spatial continuity constraints by using a dissimilarity index that allows spatial interactions between image voxels. Similarly, Leung et al. [31] and Wang et al. [12] deal with lip segmentation using fuzzy clustering with spatial restriction. Although these methods give promising results, their accuracy highly depends on the pre-defined number of segments, whose selection is often a nontrivial task in practice.

The last category is the MRF-based methods that provide an alternative means for lip segmentation by taking into account the spatial restriction. Recently, some architectural modifications on classical MRF have been presented. For instance, Liévin et al. [8], [32], [33] have utilized the multilayer MRF model, i.e. the layers corresponding to adjacent frames in a video sequence, to deal with the lip segmentation and tracking problems. Since each MRF is constructed by the color features of the corresponding frame, this model integrates both hue information and temporal dynamics. Further, Zhang et al. [14], [34] have utilized the MRF model, in which the energy function is composed of both of the color and the edge information to perform the lip segmentation. Moreover, the methods presented in [35] and [36] have also relied on MRF to segment the lip region. Nevertheless, to the best of our knowledge, the two problems stated in Section I, i.e. the local observation scale problem and the dependency problem of the number of segments, have not yet been solved in most of the existing MRF-based lip segmentation methods.

III. An Overview of the MRF Model in Image Segmentation

This section will briefly introduce the MRF model. Interested readers may refer to [37]–[39] for more details.

Let $S = \{1, 2, \dots, s\}$ be the lattice site set of a given image with r rows and c columns in pixel, where $s = r \times c$, and O is the set of observed data in a space, e.g. gray-scale space, or RGB color space. Thus, we can utilize a random field $\mathcal{X} = \{\mathcal{X}_i = x_i \mid i \in S,\ x_i \in O\}$ to represent the image, where $\mathcal{X}_i$ denotes the random variable posed in site i. Moreover, since MRF-based image segmentation problems are usually formulated as the labeling problems, we define a label MRF $\mathcal{F} = \{\mathcal{F}_i = f_i \mid i \in S,\ f_i \in L\}$ to represent the segmentation result, where the label set $L = \{0, 1, \dots, m-1\}$ is composed of the segment index, and m is the number of segments. For the sake of description hereinafter, the probability of event $\mathcal{F}_i = f_i$ is denoted as $P(f_i)$, while the joint probability $P(\mathcal{F}_1 = f_1, \mathcal{F}_2 = f_2, \dots, \mathcal{F}_s = f_s)$ is denoted by $P(f)$. Based on the definition of MRF, $\mathcal{F}$ satisfies the Markovianity

$$P(f_i | f_{S-\{i\}}) = P(f_i | f_{N_i}), \tag{1}$$

where N_i is the set of sites neighboring i.

Within the MAP framework, the optimal label configuration, i.e. the optimal segmentation result, $f^* = \{f_i^* \mid i \in S\}$ can be obtained by solving the following equation:

$$f^* = \underset{f \in \Omega^L}{\arg\max} P(f \mid x), \tag{2}$$

where the Cartesian product $\Omega^L = L \times L \times \cdots \times L = L^s$ is the configuration space of $\mathcal{F}$ on the site set S, f and x denote a configuration of $\mathcal{F}$ and $\mathcal{X}$, respectively. Eq. (2) can be further computed by the following equation:

$$f^* = \underset{f \in \Omega^L}{\arg\max} P(f) P(x \mid f). \tag{3}$$

According to the Hammersley-Clifford theorem [37], the prior probability $P(f)$ can be specified as

$$P(f) = \frac{1}{Z} e^{-U(f)/\tilde{T}}, \tag{4}$$

where Z is a normalizing constant, and $U(f)$ is the prior energy function. The temperature $\tilde{T}$ is a constant which will be simply set at 1 hereinafter.

As for the likelihood term $P(x \mid f)$, similar to Eq. (4), we let

$$P(x \mid f) = \frac{1}{Z'} e^{-U(x|f)} \tag{5}$$

provided that $P(x \mid f)$ is Gaussian distributed, where Z' is a normalizing constant and $U(x \mid f)$ denotes the likelihood energy function.

We define the energy function:

$$\mathcal{E}(f; x) = U(f) + \alpha U(x \mid f), \tag{6}$$

where the weighting parameter α is introduced to determine the proportion of $U(f)$ to $U(x \mid f)$ in the entire energy function. Thus, Eq. (3) can be rewritten as

$$f^* = \arg\min_{f \in \Omega^L} \mathcal{E}(f; x). \tag{7}$$

Besag et al. [37] provide a further specified form of the prior energy function:

$$U(f) = \sum_{c \in C} V_c(f), \tag{8}$$

which is the sum of clique potentials $V_c(f)$ over the set C of all possible cliques. Thus far, a number of optimization methods, e.g. simulated annealing (SA), and iterated conditional modes (ICM), have been utilized to solve Eq. (7).

IV. MRF-Based Image Segmentation in Various Scales

In this section, we firstly propose a hierarchical multi-scale MRF model. Then, an iterative algorithm to optimize the proposed objective function is introduced, which can perform lip segmentation without knowing the true number of segments beforehand.

A. The Hierarchical Multi-Scale MRF Model

As stated in Section III, a given image can be represented by x, which denotes a configuration of the observation random field $\mathcal{X} = \{\mathcal{X}_i = x_i \mid i \in S,\ x_i \in O\}$. For convenience of description, the sites set S is mapped into a 2D pixel coordinate $\{(p_x, p_y) \mid 1 < p_x \le c, 1 < p_y \le r\}$ temporarily. Thus, we have $x_i = x_{p_x,p_y}$ if $i = (p_y - 1) \cdot r + p_x$.

Based on the scale-space representation [40], an image pyramid can be obtained by:

$$\tilde{x}^{(k)}_{p_x,p_y} = g_k(p_x, p_y) * x_{p_x,p_y} \tag{9}$$

with

$$g_k(p_x, p_y) = \frac{1}{2\pi \cdot 2^k} \cdot exp(-\frac{p_x^2 + p_y^2}{2^{k+1}}), \tag{10}$$

where $\tilde{x}^{(k)}_{p_x,p_y}$ denotes the corresponding observed data of x_{p_x,p_y} in layer k of the pyramid.

We utilize the notation $\tilde{x}^{(k)}$ to represent the image posed in layer k of the pyramid. Since the size of each $\tilde{x}^{(k)}$ is different, we further employ interpolation to normalize them to have the same size as the finest layer. The normalized configuration is denoted by $x^{(k)} = \{x_i^{(k)} \mid i \in S\}$. Specifically, let $x^{(0)}$ be equal to x. An illustration of $x^{(k)}(k = 0, 1, 2)$ is shown in Fig. 4. Since the values in neighboring interpolation nodes are not changed abruptly, the differences among the results obtained by various interpolation methods are small. In this research, the linear interpolation is utilized because of its low computational complexity.

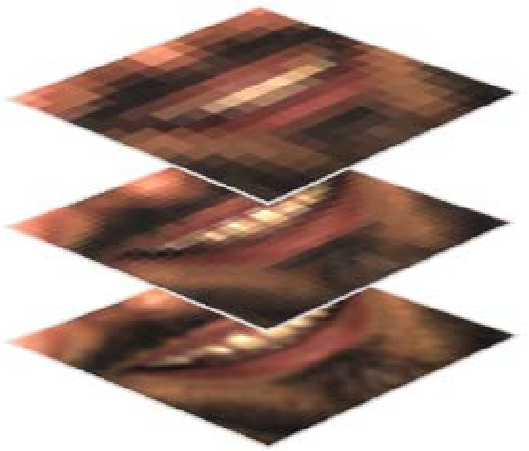


Fig. 4. An illustration of the normalized 3-layer pyramid, comprising $x^{(0)}$ (i.e. the source image x), $x^{(1)}$, and $x^{(2)}$ from bottom to top.

Thus, the traditional MRF-based image representation can be extended into multi-scale space by using a hierarchical model. For arbitrary observation scale k (i.e. the layer k of the image pyramid), there is an observation random field $\mathcal{X}^{(k)} = \{\mathcal{X}_i^{(k)} = x_i^{(k)} \mid i \in S,\ x_i^{(k)} \in O\}$. For each $\mathcal{X}^{(k)}$, there is a corresponding hidden label MRF named $\mathcal{F}^{(k)} = \{\mathcal{F}_i^{(k)} = f_i^{(k)} \mid i \in S,\ f_i^{(k)} \in L\}$ to represent the segmentation result.

Suppose each pixel of a given image has an optimal local scale from the segmentation viewpoint. Given an image, we further assume that the local scale values compose an MRF

$$\mathcal{T} = \{\mathcal{T}_i = t_i \mid i \in S,\ t_i \in D\}, \tag{11}$$

where $D = \{0, 1, \ldots, d - 1\}$ is the scale value set, in which 0 refers to the finest scale while $d-1$ means the coarsest scale. $t = \{t_i \mid i \in S, t_i \in D\}$ is the configuration of $\mathcal{T}$. Thus, the observation random filed and the corresponding hidden label MRF in the hierarchical model can be further rewritten as

$$\mathcal{X} = \{\mathcal{X}_i^{(t_i)} = x_i^{(t_i)} \mid i \in S, t_i \in D, x_i^{(t_i)} \in O\} \tag{12}$$

and

$$\mathcal{F} = \{\mathcal{F}_i^{(t_i)} = f_i^{(t_i)} \mid i \in S, t_i \in D, f_i^{(t_i)} \in L\}, \tag{13}$$

respectively.

Accordingly, the optimization problem shown in Eq. (2) can be extended by adding the scale configuration as another variable directly:

$$\{f^*, t^*\} = \arg\max_{\substack{f \in \Omega^L \\ t \in \Omega^D}} P(f, t \mid x), \tag{14}$$

where $\Omega^D = D \times D \times \cdots \times D = D^s$ is the configuration space of the scale MRF $\mathcal{T}$ on the site set S, t denotes the configuration of scale, and t^* denotes the optimal scale selection result.

The corresponding energy function can be denoted by

$$\mathcal{E}(f, t; x) = U(t) + \beta U(f^{(t)}) + \gamma U(x^{(t)} \mid f^{(t)}) \tag{15}$$

where β and γ are weighting parameters. The details of the derivation process are given in Appendix I.

We suppose that the observed data, which belong to segment j and layer k, follow a specific Gaussian distribution with the parameter $\theta^{(k)}[j] = \{\mu^{(k)}[j], \Sigma^{(k)}[j]\}$, where $\mu^{(k)}[j]$ and $\Sigma^{(k)}[j]$ denote the mean vector and covariance matrix, respectively. Since the content changes usually take place continuously in a video stream, we assume that the difference between $\theta^{(0)}[j], \theta^{(1)}[j], \ldots, \theta^{(d-1)}[j]$ is neglectable. Thus, $\theta^{(k)}[j]$ can degenerate to $\theta[j] = \{\mu[j], \Sigma[j]\}$, which is shared by all observed data that fall into the segmentation class j. Thus, the notation of the likelihood energy term, $U(x^{(t)} \mid f^{(t)})$, can be replaced by $U(x^{(t)} \mid f^{(t)}; \Theta)$, where $\Theta = \{\theta[j] \mid j \in L\}$. Eq. (15) can be rewritten as

$$\mathcal{E}(f, t; x, \Theta) = U(t) + \beta U(f^{(t)}) + \gamma U(x^{(t)} \mid f^{(t)}; \Theta). \tag{16}$$

Hereinafter, Eq. (16) is called the objective function. Since $\mathcal{T}$ is an MRF, the prior energy term $U(t)$ in Eq. (16) can be specified as

$$U(t) = \sum_{c \in C} V_c(t) = \sum_{i \in S} \sum_{i' \in N_i} [1 - \delta(t_i - t_{i'})] \tag{17}$$

where $\delta(\cdot)$ is the Kronecker delta function.

To specify the second term of Eq. (16), we introduce a 3D neighborhood system. Recall the definition of the 2D neighborhood system $N_i = \{i' \in S \mid |i - i'| \le rad, i \ne i'\}$, which represents the spatial restriction between the neighboring sites in a 2D MRF. Since the change of label in specific pixel caused by the minor observation scale variation cannot be abrupt or frequent, it is feasible to assume that the label MRFs on different pyramid layers $\{\mathcal{F}^{(k)} \mid k \in D\}$ can be treated as a 3D MRF defined on S. In this MRF, the value at site i is not only dependent on the values at sites belonging to N_i, but also the values at neighboring sites in the next scale layers. The values in the neighboring sites of i is $\{f_{i'}^{(k')} \mid i' \in N_i,\ k \in D,\ k' = k, k+1, k-1\}$. Fig. 5 depicts a 3-D neighborhood system.

Thus, the second term in Eq. (15) can be expressed by

$$\begin{aligned} U(f^{(t)}) &= \sum_{c \in C} V_c(f) \\ &= \sum_{i \in S} \sum_{i' \in N_i} \sum_{t_i' \in \{t_i - 1, t_i, t_i + 1\}} \Big[1 - \delta(f_i^{(t_i)} - f_{i'}^{(t_i')})\Big]. \end{aligned} \tag{18}$$

The likelihood term is specified as

$$\begin{aligned} &U(x^{(t)} \mid f^{(t)}; \Theta) \\ &= \sum_{i \in S} \Big[(x_i^{(t_i)} - \mu[f_i^{(t_i)}])^T (\Sigma[f_i^{(t_i)}])^{-1} (x_i^{(t_i)} - \mu[f_i^{(t_i)}])\Big], \end{aligned} \tag{19}$$

where T denotes the matrix transpose operation.

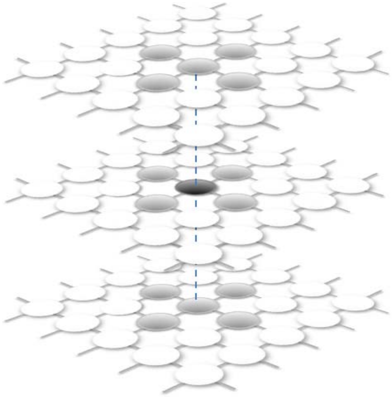


Fig. 5. Illustration of the 3D first-order neighborhood system of a given site, in which the black circle is the given site and gray circles denote the corresponding neighborhood system.

According to MAP theory, the optimal segment label map f^* and optimal local-scale map t^* can be obtained by solving the following equation:

$$\{f^*, t^*\} = \mathop{\arg\min}_{\substack{f \in \Omega^L \\ t \in \Omega^D}} \mathcal{E}(f, t; x, \Theta). \tag{20}$$

B. The Proposed Algorithm Without Knowing the Number of Segments

From the practical viewpoint, the true number of segments m^* in lip segmentation usually varies due to the appearance (e.g. moustache, teeth, and tongue) of speakers, and the estimation bias of m^* may induce segmentation error. Therefore, we present an algorithm featuring automatic selection of the number of segments to perform lip segmentation within the proposed model.

Suppose each pixel belongs to multiple segments with a different degree of membership when the corresponding observation scale is determined, the membership can be formulated as

$$\mathfrak{f}_i^{(k)}[j] = \frac{G(x_i^{(k)}; \theta[j])}{\sum_{j'=0}^{m-1} G(x_i^{(k)}; \theta[j'])},\ i \in S,\ j \in L,\ k \in D \tag{21}$$

where $G(\cdot)$ denotes the Gaussian probability density function (p.d.f.), and k denotes the scale of the given observation. Given k, there are m membership sets named $\{\mathfrak{f}_i^{(k)}[0] \mid i \in S\}$, $\{\mathfrak{f}_i^{(k)}[1] \mid i \in S\}$, ..., $\{\mathfrak{f}_i^{(k)}[m-1] \mid i \in S\}$. Moreover, as $k \in \{0, 1, \ldots, d-1\}$, the number of all possible membership sets is $m \times d$. When the observation scale at site i, denoted as t_i, is determined, there are only m membership sets left and denoted by

$$\mathfrak{f}_i^{(t_i)}[j] = \frac{G(x_i^{(t_i)}; \theta[j])}{\sum_{j'=0}^{m-1} G(x_i^{(t_i)}; \theta[j'])},\ i \in S,\ j \in L. \tag{22}$$

When each membership set is viewed as a random field defined on S, they compose a hierarchical model. There are m layers in this model, and the random field posed in the jth layer can be represented by $\{\mathfrak{f}_i^{(t_i)}[j] \mid i \in S\}$. Then, the value

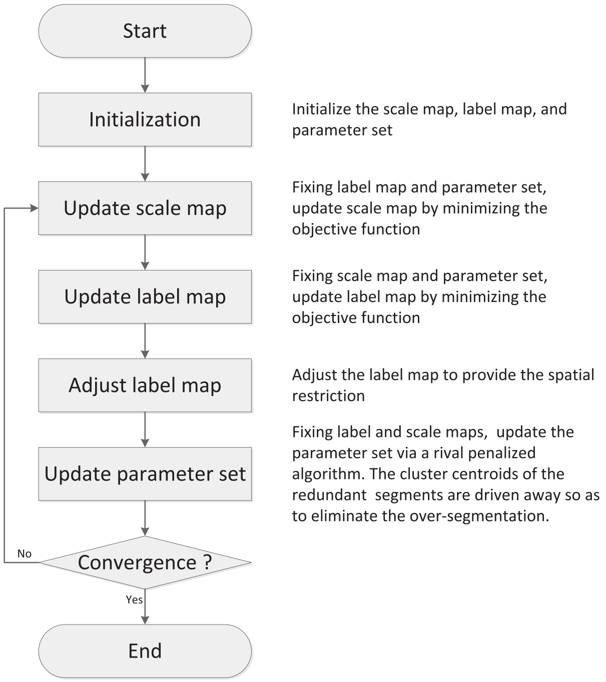

Start
Initialization
Initialize the scale map, label map, and parameter set
Update scale map
Fixing label map and parameter set, update scale map by minimizing the objective function
Update label map
Fixing scale map and parameter set, update label map by minimizing the objective function
Adjust label map
Adjust the label map to provide the spatial restriction
Update parameter set
Fixing label and scale maps, update the parameter set via a rival penalized algorithm. The cluster centroids of the redundant segments are driven away so as to eliminate the over-segmentation.
Convergence ?
No
Yes
End

Fig. 6. The work flow of the proposed method.

of label MRF $\mathcal{F}$ is used for representing the segmentation result, which can be specified by

$$f_i = \arg\max_{j\in L} \mathfrak{f}_i^{(t_i)}[j]. \tag{23}$$

According to Eq. (23), f_i can be regarded as a function with respect to $\mathfrak{f}_i^{(t_i)}[j]$. Thus, Eq. (18) can be regarded as a function with respect to the configuration $\{\mathfrak{f}_i^{(t_i)}[j] \mid i \in S, j \in L\}$. For simplicity, we utilize the notation N_i^+ to represent the site set $N_i \cup \{i\}$. We employ the following equation to approximate the map between $U(f^{(t)})$ and $\{\mathfrak{f}_i^{(t_i)}[j] \mid i \in S, j \in L\}$:

$$\hat{U}(f^{(t)})=\sum_{i\in S}\sum_{j\in L}\sum_{i'\in N_i^+}\frac{\mathfrak{f}_{i'}^{(t_{i'})}[j]}{\sum_{i''\in N_i^+}\mathfrak{f}_{i''}^{(t_{i''})}[j]}\ln\frac{\mathfrak{f}_{i'}^{(t_{i'})}[j]}{\sum_{i''\in N_i^+}\mathfrak{f}_{i''}^{(t_{i''})}[j]}. \tag{24}$$

Unlike the Kronecker delta function, Eq. (24) utilizes the information entropy-based disorder measurement in the local object system to represent the spatial coherence restriction.

We first initialize the parameter set $\Theta = \{\theta[j] \mid j \in L\}$, the configuration of label MRF $f = \{f_i \mid i \in S, f_i \in L\}$, and the configuration of observation scale MRF $t = \{t_i \mid i \in S, t_i \in D\}$. For each observed data $x_i^{(t_i)}$, we utilize an iterative method to update Θ in order to minimize the objective function shown in Eq. (15) while eliminating the redundant segment cluster(s). The proposed method is summarized in Fig. 6, and the detailed implementation is given as follows:

1. The proposed objective function shown in Eq. (15) becomes:

$$\mathcal{E}(t; f, x, \Theta)=U(t)+\beta U(f^{(t)})+\gamma U(x^{(t)} \mid f^{(t)}; \Theta). \tag{25}$$

when Θ and f are fixed.

We find the scale configuration $\hat{t} = \{\hat{t}_i \mid i \in S\}$ which makes Eq. (25) minimum by ICM. Let $t = \hat{t}$.

2. Fixing Θ and t, calculate $\mathfrak{f}_i^{(t_i)}[j]$ for each site via Eq. (22).
3. Considering the prior energy term approximated by Eq. (24), we adjust $\mathfrak{f}_i^{(t_i)}[j]$ to make the local systems $\{\mathfrak{f}_{i'}^{(t_i')}[j] \mid i' \in N_i^+\}$ reach the highest uncertainty, which provides the spatial restriction, i.e.

$$\hat{\mathfrak{f}}_i^{(t_i)}[j] = \exp(-\frac{\sum_{i'\in N_i}\mathfrak{f}_{i'}^{(t_{i'})}[j]\ln\mathfrak{f}_{i'}^{(t_{i'})}[j]}{\sum_{i'\in N_i}\mathfrak{f}_{i'}^{(t_{i'})}[j]}). \tag{26}$$

Let $\mathfrak{f}_i^{(t_i)}[j] = \hat{\mathfrak{f}}_i^{(t_i)}[j]$.

4. Update f_i for each site i via Eq. (23).
5. Fixing f and t, we rewrite Eq. (15) as

$$\mathcal{E}(\Theta; f, t, x)=U(t)+\beta U(f^{(t)})+\gamma U(\Theta; x^{(t)}, f^{(t)}). \tag{27}$$

Then, we update Θ for each x_i via

$$\hat{\theta}[j_w] = \theta[j_w] - \eta_w\frac{d\mathcal{E}(\Theta; f, t, x)}{d\Theta}|_{\theta[j_w]}, \tag{28}$$

and

$$\hat{\theta}[j_r]=\theta[j_r]+\eta_r\cdot\delta E_i^{(t)}[j_w, j_r]\cdot 1_{[0,+\infty)}(\delta E_i^{(t)}[j_w, j_r]) \cdot\frac{d\mathcal{E}(\Theta; f, t, x)}{d\Theta}|_{\theta[j_r]}, \tag{29}$$

where $j_w = \arg\max_j \mathfrak{f}_i^{(t_i)}[j]$, $j_r \in L$ but $j_r \neq j_w$, η_w and η_r are two positive numbers which denote the learning rate and penalty rate, respectively. Let $\theta[j_w] = \hat{\theta}[j_w]$, $\theta[j_r] = \hat{\theta}[j_r]$, and $\delta E_i^{(t)}[j_w, j_r]$ represent the penalty strength, which is an entropy difference-based measurement of the similarity defined by:

$$\delta E_i^{(t)}[j, j'] = E_i^{(t)}[j] - E_i^{(t)}[j, j'] \tag{30}$$

where

$$E_i^{(t)}[j] = -\frac{1}{n}\sum_{i'\in N_i^+}\frac{\mathfrak{f}_{i'}^{(t_{i'})}[j]}{\sum_{i''\in N_i^+}\mathfrak{f}_{i''}^{(t_{i''})}[j]} \times\ln\frac{\mathfrak{f}_{i'}^{(t_{i'})}[j]}{\sum_{i''\in N_i^+}\mathfrak{f}_{i''}^{(t_{i''})}[j]}, \quad i\in S, j\in L, t_i\in D, \tag{31}$$

and

$$E_i^{(t)}[j, j'] = -\frac{1}{2n}\sum_{i'\in N_i^+}\left\{\frac{\mathfrak{f}_{i'}^{(t_{i'})}[j]}{F_i^{(t_i)}[j, j']}\ln\frac{\mathfrak{f}_{i'}^{(t_{i'})}[j]}{F_i^{(t_i)}[j, j']} + \frac{\mathfrak{f}_{i'}^{(t_{i'})}[j']}{\sum_{i''\in N_i^+}\left[\mathfrak{f}_{i''}^{(t_{i''})}[j]+\mathfrak{f}_{i''}^{(t_{i''})}[j']\right]} \cdot\ln\frac{\mathfrak{f}_{i'}^{(t_{i'})}[j']}{\sum_{i''\in N_i^+}\left[\mathfrak{f}_{i''}^{(t_{i''})}[j]+\mathfrak{f}_{i''}^{(t_{i''})}[j']\right]}\right\}, \quad i \in S, j \in L, j' \in L, t_i \in D \textit{ and } j \neq j'. \tag{32}$$

Moreover, $1_{[0,+\infty)}(\delta E_i^{(t)}[j_w, j_r])$ denotes the step function given by

$$1_{[0,+\infty)}(\delta E_i^{(t)}[j_w, j_r]) = \begin{cases} 1, & \delta E_i^{(t)}[j_w, j_r] \in [0,+\infty), \\ 0, & otherwise. \end{cases} \tag{33}$$

The above steps 1-5 are implemented iteratively until f and t converge.

C. Postprocessing

Let the index of lip segment layer be denoted by j^*, and the lip membership set be:

$$lip = \{\mathfrak{f}_i[j^*] \mid i \in S\}. \tag{34}$$

For convenience of description, we map the sites set S into a 2D coordinate $\{(p_x, p_y) \mid 1 < p_x \leq c, 1 < p_y \leq r\}$ again. Thus, lip can be rewritten as

$$lip = \{\mathfrak{f}_{(p_x,p_y)}[j^*] \mid 1 < p_x \leq c, 1 < p_y \leq r\}. \tag{35}$$

Suppose that the lip region is not connected to the border of image, the morphological reconstruction-based method in [41] is employed to suppress border connected noise structures. For the positive elements in the reconstruction result, a threshold selection method proposed in [42] is utilized to bring a binary version of lip. Furthermore, we utilize the morphological opening with 3 × 3 structuring element and denote the result as bin.

For the foreground elements in bin, the corresponding positions

$$\{(p_x, p_y) \mid bin_{(p_x,p_y)} = 1, 1 < p_x \leq c, 1 < p_y \leq r\} \tag{36}$$

are organized by a matrix $\mathbb{M}$ as follows:

$$\mathbb{M} = \begin{bmatrix} p_{x1} & p_{y1} \\ p_{x2} & p_{y2} \\ \vdots & \vdots \\ p_{xl} & p_{yl} \end{bmatrix} \tag{37}$$

where l is the number of foreground elements in bin, and $bin_{(p_x,p_y)}$ denotes the binary value of bin posed in (p_x, p_y).

Then, the eigenvectors and eigenvalues of the covariance matrix of $\mathbb{M}$ can be calculated. We can further obtain an ellipse whose position and inclination are defined by the eigenvectors. Furthermore, the length of major and minor axes are defined by the 1.5 times square root of their corresponding eigenvalues, respectively. Consequently, two horizontal lines crossing the highest and lowest points of the ellipse are obtained. The continued objects on the outside of the two lines are masked out then.

V. Experiments

A. Databases, Counterparts and Performance Measurement

To show the performance of the proposed approach under the different capture environments, we utilized four benchmark face databases: (1) AR [43], (2) CVL [44], (3) GTAV [45], and (4) VidTIMIT [46]. For each image, the part of the face between nostril and chin is clipped by a 128 × 128 (AR, CVL and VidTIMIT) or 64 × 64 (GTAV) pixel window as the source of segmentation experiment. Fig. 7 shows some samples of the clipped source images used in the experiments. It can be seen that the images from different databases have different illumination condition, color temperature, skin color, and moustache.

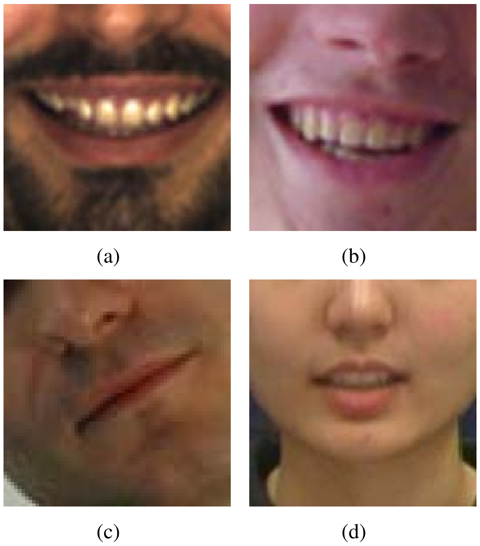


Fig. 7. A sample of the source images utilized in our experiment: (a) AR, (b) CVL, (c) GTAV, and (d) VidTIMIT databases.

In our experiments, we compared the proposed method with three existing counterparts, denoted as Lievin04 [8], Liang06 [47] and Wang07 [12], respectively. Lievin04 is an MRF-based lip segmentation and tracking method, which is robust against the number of segments. This method is designed to deal with the image sequence. Thus, it was utilized directly in the VidTIMIT database in our experiments. However, for the databases AR, CVL and GTAV, composed by separate images, it was utilized with a slight modification. It is believed that segmenting the lip from one separate image can be regarded as a special case of lip tracking from two adjacent frames. Thus, when Lievin04 was employed in the separate image databases, a given image was assigned to frame t and $t-1$ simultaneously in Lievin04's model.

Liang06 is a multi-scale MRF-based image segmentation with automatic selection of the number of segments. Although this method is designed for texture image segmentation rather than lip segmentation, its characteristics, e.g. multi-MRF-based and the independence from the number of segments, are similar to the proposed method. Wang07 is a typical spatial fuzzy clustering (SFC)-based lip segmentation with automatic selection of the number of segments. Analogous to the proposed method, it also consider the spatial restriction.

To evaluate the performance of the algorithms, two measures defined in [29] were used. The first measure

$$OL = \frac{2(A_1 \cap A_2)}{A_1 + A_2} \times 100\% \tag{38}$$

determines the percentage of overlap between the segmented lip region A_1 and the ground truth A_2. The second measure is the segmentation error (SE) defined as

$$SE = \frac{OLE + ILE}{2 \times TL} \times 100\%, \tag{39}$$

where OLE is the number of non-lip pixels classified as lip pixels (i.e. outer lip error), ILE is the number of lip-pixels classified as non-lip ones (inner lip error), and TL denotes the number of lip-pixels in the ground truth.

The ground truth of each image was obtained by manual segmentation. Due to the resolution and illustration condition, the lip region near boundary is often hard to classify, even by a human being. Under the circumstances, two volunteers were invited to segment the ground truth independently and each experiment was evaluated using the two corresponding ground truths.

B. Model Parameter Initialization

In the first two experiments in Sub-section V.C, the label set was estimated by $\hat{L} = \{0, 1, \ldots, 9\}$, i.e. the pre-assigned number of segments $\hat{m}$ was set at 10, which could provide enough margin for the lip segmentation. Then, we randomly selected $\hat{m}$ pixels in the source image and utilized their observed values to initialize the mean vectors, i.e.

$$\mu[j] = x_{rand(S)}, \quad j \in \hat{L}, \tag{40}$$

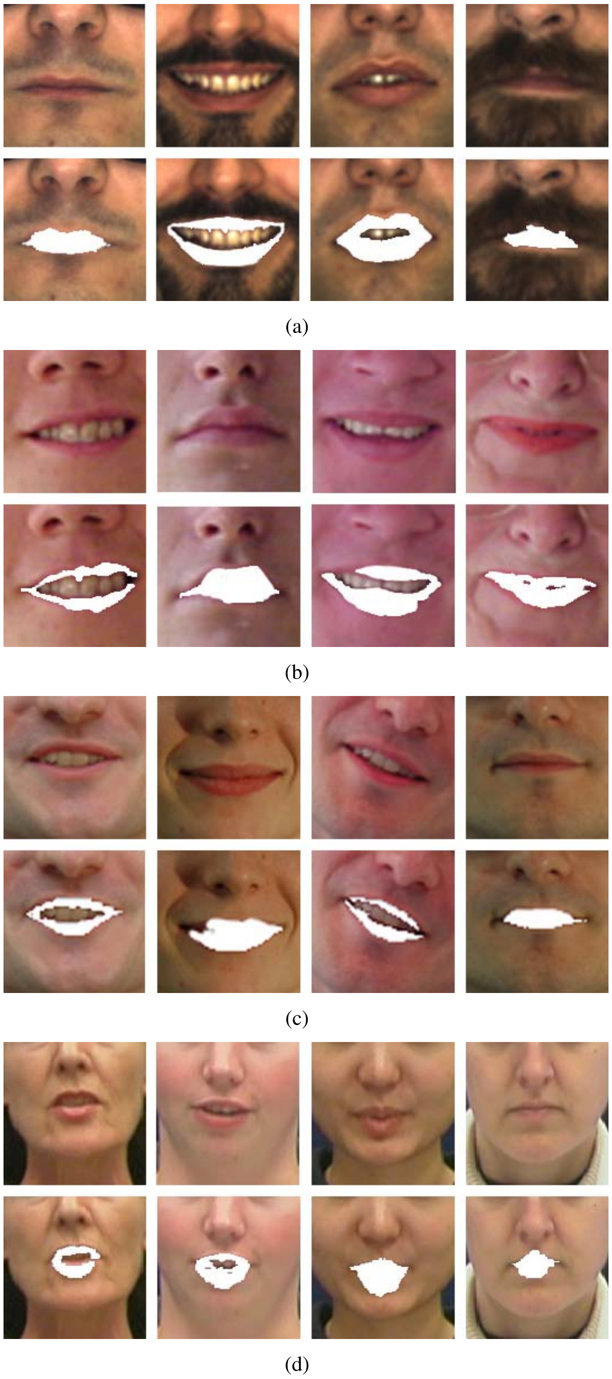
where $rand(S)$ denotes a random element in set S.

Meanwhile, the covariance matrixes were initialized at

$$\Sigma[j] = \begin{bmatrix} 100 & 0 \\ 0 & 100 \end{bmatrix}, \quad j \in \hat{L}. \tag{41}$$

The configuration of the local-scale MRF $\mathcal{T}$ was pre-assigned by $\{\mathcal{T}_i = 0 \mid i \in S\}$. All pixels were observed initially under the finest scale. In addition, we utilized a Cartesian Coordinate-based version of HSV color space as our observation space O because HSV color space resembles the way human beings perceive color information [48]. Specifically, we first transformed the original image into the HSV color space. The *hue*, *saturation*, and *value* component for site i were denoted by h_i, s_i and v_i, respectively. Note that the v_i component only conveys the luminance information, which may be affected by illumination. Moreover, it is hard to discriminate between some regions with deep color in the v_i channel (e.g. the hues of lip and moustache are evidently different, but the luminance values are similar). Thus, we utilized h_i and s_i components only. Since the calculation of the proposed method is based on Euclidean distance, we transformed the h_i and s_i components from the Polar Coordinate system to the Cartesian Coordinate system accordingly. That is, for each site, we let the observed data be:

$$x_i = [s_i \cdot \cos(2\pi \cdot h_i), s_i \cdot \sin(2\pi \cdot h_i)]^T. \tag{42}$$


(a)

(b)

(c)

(d)

Fig. 8. A snapshot of lip segmentation results obtained by the proposed method. The images in (a), (b), (c) and (d) are from AR, CVL, GTAV and VidTIMIT databases, respectively. In each sub-figure, the first row shows the source images, and the second row is the corresponding results.

There are also four other parameters whose values should be determined: β, γ (see Eq. (16)), η_w (see Eq. (28)) and η_r (see Eq. (29)). According to usual practice, the learning rates η_w and η_r are set at 0.01 and 0.001 by a rule of thumb, respectively. For the two weighting parameters β and γ, we utilize the following method to select their optimal values:

1. Let $W = \{0.1 \times k \mid k \in \{1, 2, \ldots, 100\}\}$. The Cartesian product $W \times W$ can be regarded as the value space for (β, γ).
2. We randomly select $\bar{N}$ images from four databases as a training set. Note that the training set is disjoint to the testing set utilized in Sub-section V.C. Fixing β and γ,

TABLE I

AVERAGE OVERLAP AND SEGMENTATION ERROR FOR THE IMAGES FROM AR, CVL, GTAV AND VIDTIMIT DATABASES OBTAINED BY LIEVIN04, LIANG06, WANG07 AND THE PROPOSED METHOD, RESPECTIVELY

Measure	Lievin04		Liang06		Wang07		Proposed Method	
Database	**OL**	**SE**	**OL**	**SE**	**OL**	**SE**	**OL**	**SE**
AR	79.5%	21.3%	85.0%	19.3%	93.1%	8.0%	***91.3%***	***8.1%***
CVL	91.8%	7.8%	89.8%	12.1%	92.5%	7.0%	***92.5%***	***7.9%***
GTAV	72.4%	33.9%	81.3%	22.3%	84.3%	22.7%	***91.7%***	***7.2%***
VidTIMIT	89.5%	13.2%	90.0%	11.5%	90.4%	11.2%	***91.2%***	***7.4%***

TABLE II

MEAN VALUE AND STANDARD DEVIATION OF PROCESSING TIME FOR THE IMAGES FROM AR, CVL, GTAV AND VIDTIMIT DATABASES OBTAINED BY LIEVIN04, LIANG06, WANG07 AND THE PROPOSED METHOD, RESPECTIVELY

Measure	Lievin04		Liang06		Wang07		Proposed Method	
Database	**Mean**	**Std**	**Mean**	**Std**	**Mean**	**Std**	**Mean**	**Std**
AR	4.1s	0.3s	1.5s	0.3s	9.1s	1.3s	***2.8s***	***1.1s***
CVL	4.5s	0.2s	1.5s	0.3s	11.3s	2.0s	***2.5s***	***0.7s***
GTAV	1.0s	0.1s	0.5s	0.1s	7.7s	2.1s	***1.0s***	***0.5s***
VidTIMIT	3.9s	0.7s	1.4s	0.2s	9.1s	1.9s	***2.2s***	***1.3s***

TABLE III

OVERLAP RATE AND SEGMENT ERROR OF SEGMENTED LIPS OVER $\hat{m}$ IN THE IMAGES FROM FOUR DATABASES

Database	**AR**		**CVL**		**GTAV**		**VidTIMIT**	
$\hat{m}$	**OL**	**SE**	**OL**	**SE**	**OL**	**SE**	**OL**	**SE**
3	91.1%	7.7%	92.5%	8.1%	92.4%	7.1%	91.0%	7.8%
4	91.1%	7.7%	92.0%	7.6%	92.3%	7.1%	91.3%	7.1%
5	91.5%	7.8%	92.0%	7.7%	91.5%	7.9%	91.2%	7.2%
6	91.4%	7.7%	91.7%	7.5%	92.3%	7.0%	91.2%	7.3%
7	91.2%	8.0%	92.0%	8.1%	91.3%	7.9%	91.3%	7.3%
8	91.1%	7.9%	92.2%	7.3%	92.0%	7.1%	91.5%	7.2%
9	91.5%	7.8%	91.8%	8.0%	91.7%	7.2%	91.7%	7.2%
10	91.3%	7.5%	92.1%	7.7%	92.2%	6.8%	91.5%	7.2%
11	91.2%	7.3%	91.5%	7.9%	92.1%	7.7%	91.5%	7.2%
12	91.2%	7.3%	91.7%	7.7%	92.3%	7.4%	91.6%	7.2%

for the ith image, we calculate its OL value which is denoted by $OL_i(\beta, \gamma)$.

3. The optimal selection of β and γ can be obtained by

$$(\beta^*, \gamma^*) = \underset{(\beta,\gamma)}{\arg\min} \sum_{(\beta,\gamma)\in W\times W} \sum_{i=1}^{\bar{N}} ||OL_i(\beta, \gamma) - 1||_2. \tag{43}$$

In this paper, we arbitrarily set $\bar{N}$ at 50, under which the values of β^* and γ^* we obtained were 2.2 and 1.9, respectively. Subsequently, we will fix and utilize them to evaluate the proposed approach in Sub-section V.C.

C. Experimental Results

To evaluate the performance of the proposed method under the different capture environments, i.e. illumination condition, color temperature, and complexion, We conducted three experiments. In Experiment 1, we selected 200 images in total from AR, CVL, GTAV and VidTIMIT databases (i.e. 50 from each). For the images from the same database, the average OL and SE were calculated. A snapshot of lip segmentation results is shown in Fig. 8. We also compared three existing counterparts stated in the previous sub-section with the proposed method. Table I lists the average OL and SE values obtained by these four methods.

Furthermore, in Experiment 2, we randomly selected 50 images from all four databases to evaluate the actual running time of these four methods. All experiments were conducted under MATLAB R2010a on a computer with the configuration as follows:

- CPU: Intel(R) Core(TM)2 Duo CPU E7500 @2.93GHz;
- RAM: 4.00GB;
- OS: Microsoft Windows 7 with 64-bit Version.

The mean values and standard deviations of their running times are shown in Table II. It can be seen that the proposed method runs much faster than Wang07 in all cases we have tried so far. Furthermore, the running time of the proposed method is faster than Lievin04 in most cases, except the GTAV database, although it is slightly slower than Liang06.

In addition, to further investigate the robustness of the proposed method against the number of segments, we repeated the first experiment with $\hat{m} = 3, 4, \ldots, 12$ in Experiment 3. Table III lists the average OL and SE on the different image groups and $\hat{m}$. It can be seen that the segmentation performance of the proposed approach is robust against $\hat{m}$ in all cases we have tried so far.

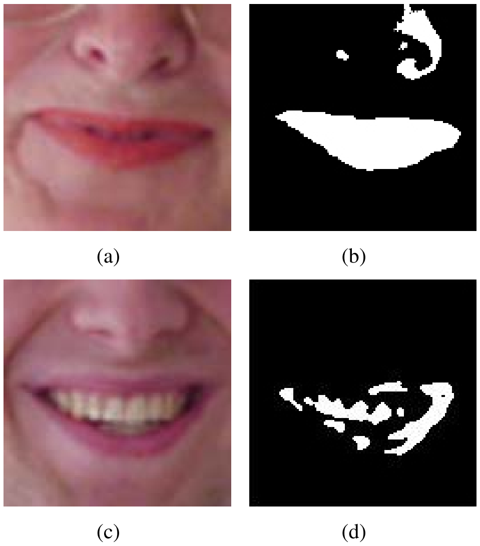


Fig. 9. Segmentation results obtained by Liang04 with the parameter $\gamma = 165$ based on the clips of source images: (a) $026 - MVC - 007F$ and (c) $038 - MVC - 007F$ in CVL, respectively, while (b) and (d) are the corresponding segmentation results.

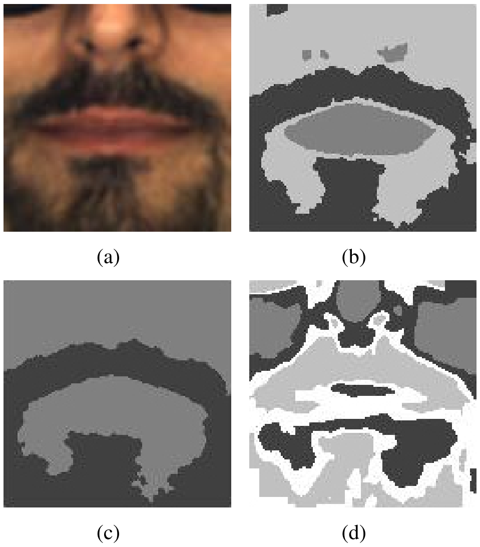


Fig. 10. Segmentation results obtained by Liang06 with different pre-defined numbers of segments. (a) is a clip of source image $m-004-8$ in AR, (b), (c), and (d) are the segmentation results obtained by Liang06 with three (i.e. an appropriate number of segments), two and four segments, respectively.

D. Discussion

In the original model of Lievin04, the threshold parameter γ is a constant value, which, however, often leads to over- or under-segmentation in the datasets we have utilized. Under the circumstances, we have adaptively selected the value to match different images when implementing Lievin04. For example, for the clip of $026 - MVC - 007F$ in CVL, which is shown in Fig. 9 (a), the lip region can be extracted accurately by Lievin04 with $\gamma = 165$ as shown in Fig. 9 (b). However, the clip of $038 - MVC - 007F$ in CVL (see Fig. 9 (c)) is over-segmented as shown in Fig. 9 (d) if we leave γ unchanged.

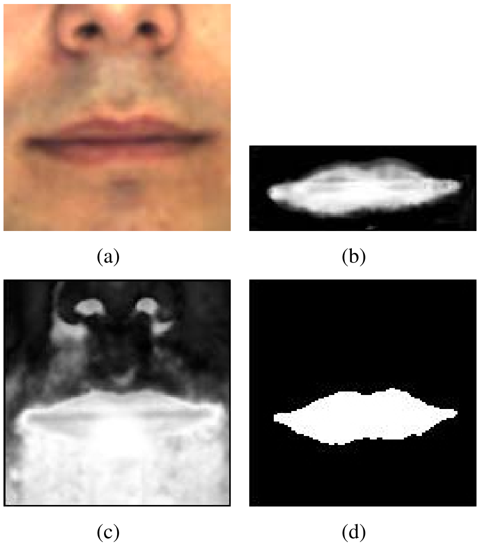


Fig. 11. Segmentation results obtained by Wang07 with different image sizes and by the proposed method. (a) is a 128 × 128 clip of source image $m - 030 - 1$ image in AR, (b) and (c) are the segmentation results obtained by Wang07 with an appropriate size and original size, respectively, (d) is the segmentation result obtained by the proposed method.

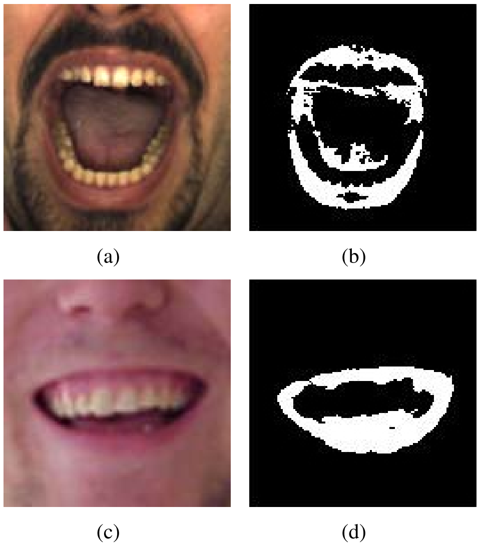


Fig. 12. Unsuccessful segmentation results obtained by the proposed method. (a) and (c) are the clips of source images $m-004-17$ in AR and $046-MVC-007F$ in CVL, (b) and (d) are the corresponding segmentation results obtained by the proposed method, in which some parts of teethridge and tongue are misclassified.

Liang06 utilizes the variance ratio (VR) criterion [49], which is also known as Fisher's criterion, to estimate the number of segments online. However, this method often leads to a poor estimate of the number of segments in the cases we have tried so far. In previous experimental comparative studies, the results of Liang06 were demonstrated with an appropriate pre-defined number of segments. In fact, the performance of Liang06 depends highly on the selection of the number of segments. A corresponding example is shown in Fig. 10. Obviously, the lip can be segmented accurately when the number of segments is select correctly as shown in Fig. 10 (b). Nevertheless, a bias of the estimation of the

TABLE IV
THE VARIATION OF OL AND SE CAUSED BY THE WEIGHTING CHANGE OF $U(t)$

(β, γ)	(0.2, 0.2)	(0.4, 0.4)	(0.6, 0.6)	(0.8, 0.8)	(1, 1)	(2, 2)	(4, 4)	(6, 6)	(8, 8)	(10, 10)
OL	85.6%	89.3%	89.5%	90.3%	91.4%	90.7%	89.0%	52.7%	41.1%	34.5%
SE	13.3%	12.4%	15.3%	14.5%	14.9%	11.5%	17.0%	63.1%	113.8%	137.5%

TABLE V
THE VARIATION OF OL AND SE CAUSED BY THE WEIGHTING CHANGE OF $U(f^{(t)})$

(β, γ)	(0.2, 2)	(0.4, 2)	(0.6, 2)	(0.8, 2)	(1, 2)	(2, 2)	(4, 2)	(6, 2)	(8, 2)	(10, 2)
OL	28.5%	32.2%	59.0%	73.5%	91.7%	90.7%	87.5%	86.0%	84.9%	81.7%
SE	291.0%	210.5%	103.4%	27.6%	13.1%	11.5%	21.8%	24.4%	33.5%	25.7%

TABLE VI
THE VARIATION OF OL AND SE CAUSED BY THE WEIGHTING CHANGE OF $U(x^{(t)} \mid f^{(t)})$

(β, γ)	(2, 10)	(2, 8)	(2, 6)	(2, 4)	(2, 2)	(2, 1)	(2, 0.8)	(2, 0.6)	(2, 0.4)	(2, 0.2)
OL	86.9%	87.5%	89.7%	90.1%	90.7%	91.1%	81.3%	51.1%	33.3%	28.6%
SE	25.3%	21.3%	19.8%	13.2%	11.5%	11.9%	22.3%	71.9%	120.1%	194.0%

number of segments will generally lead to under- or over-segmentation as shown in Fig. 10 (b) and (c). By contrast, the proposed method utilizing a rival-penalized mechanism features the automatic selection of the number of segments as long as $\hat{m}$ is no less than m^*.

In the comparative study, Wang07 worked quite well and even slightly better than the proposed method in the cases of AR and CVL in Table I. Nevertheless, such results were obtained by using the size of image clips as suggested in [12] (see Fig. 11 (b)) rather than the other sizes such as 128 × 128 or 64 × 64. In fact, our empirical studies have also found that the performance of Wang07 somewhat depends on the size of image clips. For example, if we utilized the 128 × 128 image from AR as the input, the segmentation results given by Wang07 deteriorate, as shown in Fig. 11 (c). In contrast, the proposed method is robust against the clip size and always gives the correct result, as shown in Fig. 11 (d).

Although the proposed method shows the promising results in most cases, the teethridge and tongue may be classified into the lip segment as shown in Fig. 12 because the visible tongue or teethridge has a similar value to the lip in the observation space. From the theoretical viewpoint, this kind of problem is hard to circumvent by a color analysis-based segmentation method. One feasible way to tackle this problem is to integrate prior shape modeling with the proposed method. We shall leave this for future work.

Another issue to be discussed here is the segmentation error caused by the selection of parameters. Four parameters are utilized in our method: β, γ, η_w and η_r. For simplicity, their values can be selected by a rule of thumb. As soon as they are fixed, there is no need to adjust them for different images. Although the proposed method is robust against these parameters to a certain level as long as they are not changed significantly, it is necessary to explain the influence of parameters upon segmentation accuracy. In Eq. (16), β and γ are weighting parameters to determine how much each component contributes to the whole energy function. With different values, segmentation results may fall into the following six cases.

TABLE VII
THE SEGMENTATION RESULT AND CONVERGENCE EPOCH OBTAINED BY DIFFERENT η_w (η_r IS FIXED TO 0.001)

η_w ($\eta_r = 0.001$)	OL	SE	epoch
0.004	91.2%	10.3%	10.5
0.008	91.3%	10.8%	9.8
0.01	91.3%	10.7%	9.5
0.012	91.3%	10.7%	9.4
0.014	91.3%	10.6%	9.4
0.02	89.9%	12.7%	8.7
0.04	75.5%	43.0%	6.1
0.1	12.9%	140.2%	4.8

1) $U(t)$ is over-weighted. During the segmentation procedure, local scales may be unified gradually because the smooth constraint of local scales is dominant in the whole energy function. Under extreme conditions, the proposed model may degenerate to a scale fixed one.
2) $U(t)$ is under-weighted. The scale configuration calculation is sensitive to noise (especially pepper and salt noise) because the priori information (e.g. spatial relationship) of local scales is ignored.
3) $U(f^{(t)})$ is over-weighted. Under-segmentation may occur because the smooth constraint of labels is dominant in the whole energy function. Under extreme conditions, all pixels may be classified into one segment.
4) $U(f^{(t)})$ is under-weighted. The segmentation result is sensitive to noise (especially pepper and salt noise) because the priori information (e.g. spatial relationship) of labels is ignored.
5) $U(x^{(t)} \mid f^{(t)})$ is over-weighted. The feature modeling component is dominant in the whole energy function. The proposed model may degenerate to the MAP estimation because all priori terms (e.g. spatial relationship) are ignored.
6) $U(x^{(t)} \mid f^{(t)})$ is under-weighted. The segmentation accuracy may degrade significantly because the

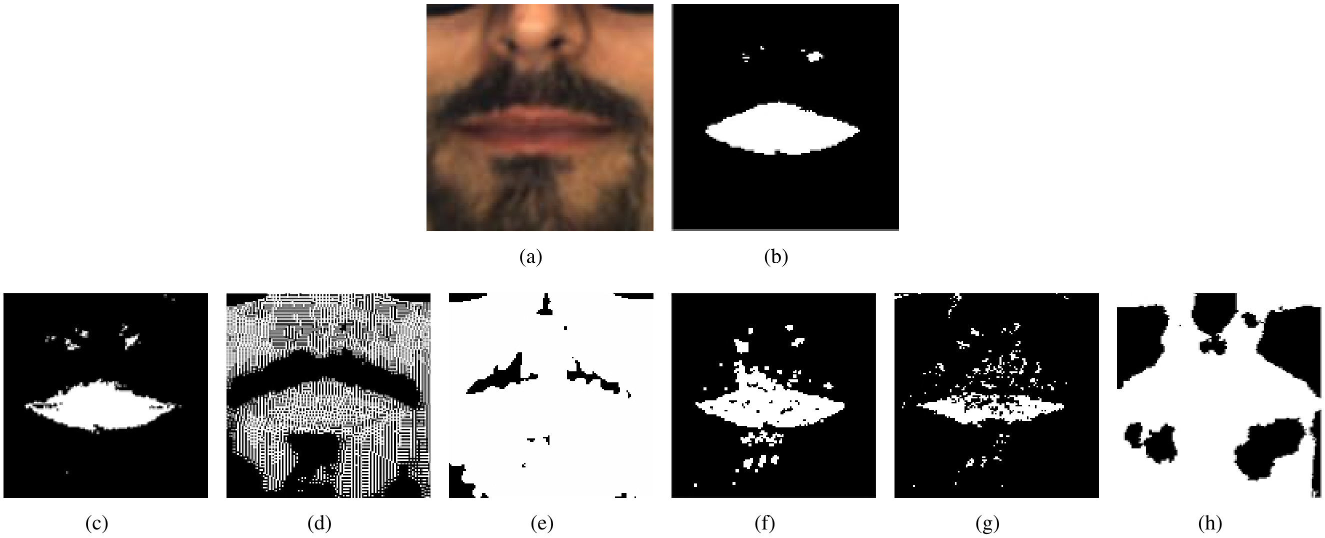


Fig. 13. Sample of the segmentation result with different weighting parameters β and γ: (a) the source image, (b) the segmentation with proper parameters ($\beta = \gamma = 2$), (c) the segmentation when $U(t)$ is over-weighted ($\beta = \gamma = 0.2$), (d) the segmentation when $U(t)$ is under-weighted ($\beta = \gamma = 10$), (e) the segmentation when $U(f^{(t)})$ is over-weighted ($\beta = 10$, $\gamma = 2$), (f) the segmentation when $U(f^{(t)})$ is under-weighted ($\beta = 0.2, \gamma = 2$), (g) the segmentation when $U(x^{(t)} \mid f^{(t)})$ is over-weighted ($\beta = 2$, $\gamma = 10$), and (g) the segmentation when $U(x^{(t)} \mid f^{(t)})$ is under-weighted ($\beta = 2$, $\gamma = 0.2$).

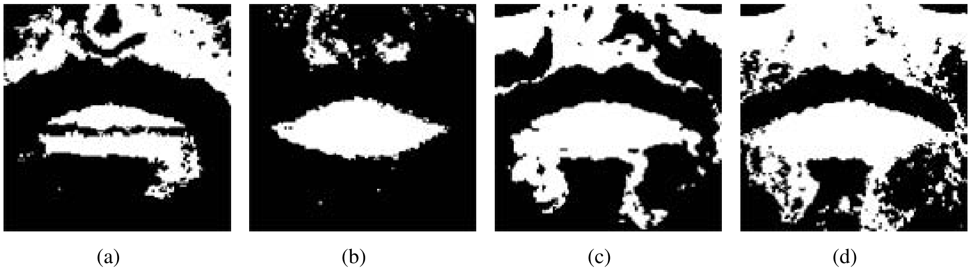


Fig. 14. A snapshot of segmentation result obtained by $\eta_w = 0.1$, where the four images (a)-(d) are the results after epoch 1, 2, 3, 4, respectively.

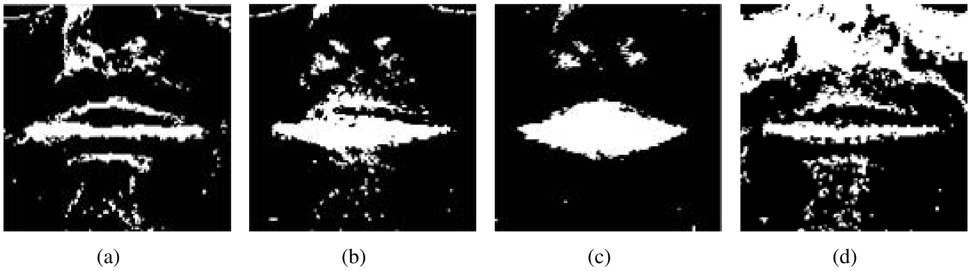


Fig. 15. A snapshot of segmentation result obtained by $\eta_r = 0.02$, where the four images (a)-(d) are the results after epoch 1, 2, 3, 4, respectively.

observation values that play the foundation role in segmentation task are ignored.

To show the segmentation error caused by weighting parameter selection quantitatively, we conduct an experiment with different β and γ based on the six cases above. The input of this experiment is 30 images randomly selected from the four databases, i.e. AR, CVL, GTAV and VidTIMIT. Experimental results in terms of average OL and SE are shown in Table IV - Table VI. Each table represents the variation of segmentation result caused by changing of one energy function component weighting (i.e. from over- to under-weighted). Fig. 13 illustrates the samples of segmentation results obtained when each energy function component is over- and under-weighted.

As for η_w and η_r, η_w controls the learning speed of the "true" segment centroid, and η_r controls the speed of driving the redundant segment centroid far from the observation data in the feature space. If the values selected are too small, the convergence time may be very long. However, if the values are too large, the segmentation result may not be reached smoothly. From the practical viewpoint, the learning rate and penalty rate can generally be assigned to a small positive constant. Table VII and Table VIII show the segmentation accuracy, and the number of epochs needed for learning convergence (simply called *convergence epoch*) obtained with different η_w and η_r. The input of this experiment is ten images randomly selected from the four databases: AR, CVL, GTAV

TABLE VIII

THE SEGMENTATION RESULT AND CONVERGENCE EPOCH OBTAINED BY DIFFERENT η_r (η_w IS FIXED TO 0.01)

$\boldsymbol{\eta_r}$ ($\eta_r = 0.01$)	**OL**	**SE**	**epoch**
0.0004	91.3%	10.1%	10.5
0.0008	91.2%	10.8%	9.9
0.001	91.3%	10.7%	9.5
0.002	91.2%	10.7%	9.5
0.004	91.2%	10.5%	9.3
0.01	34.3%	121.6%	5.1
0.02	24.2%	230.5%	4.7

and VidTIMIT. The samples of over-learning caused by η_w and η_r are illustrated in Fig. 14 and Fig. 15, respectively.

VI. CONCLUSION

We have proposed a local-scale dependent segmentation model within the MAP-MRF framework. In this model, the classical energy function in MAP-MRF-based segmentation has been extended by adding an observation scale variable for each site. Then, an automatic lip segmentation method has been presented, featuring automatic selection of the appropriate scale configuration and the number of segments simultaneously. Experimental results have shown the efficacy of the proposed method in comparison to the existing counterparts.

APPENDIX I

Based on the Bayesian rule, Eq. (14) can be further expressed as

$$\{f^*, t^*\} = \mathop{\arg\max}_{\substack{f \in \Omega^L \\ t \in \Omega^D}} P(t) \cdot P(f \mid t) \cdot P(x \mid f, t). \tag{44}$$

Under the proposed hierarchical model, we have $x_i = x_i^{(t_i)}$ and $f_i = f_i^{(t_i)}$ when $t = \{t_i \mid i \in S, t_i \in D\}$ is determined. The conditional terms can be rewritten as:

$$P(f \mid t) = P(\mathcal{F}_1 = f_1^{(t_1)}, \mathcal{F}_2 = f_2^{(t_2)}, \ldots, \mathcal{F}_s = f_s^{(t_s)}), \tag{45}$$

and

$$\begin{aligned} P(x \mid f, t) = P(&\mathcal{X}_1 = x_1^{(t_1)}, \mathcal{X}_2 = x_2^{(t_2)}, \ldots, \mathcal{X}_s \\ &= x_s^{(t_s)} \mid \mathcal{F}_1 = f_1^{(t_1)}, \mathcal{F}_2 = f_2^{(t_2)}, \ldots, \mathcal{F}_s = f_s^{(t_s)}). \end{aligned} \tag{46}$$

Thus, Eq. (44) can be described as:

$$\begin{aligned} \{f^*, t^*\} = &\mathop{\arg\max}_{\substack{\{f_i^{(t_i)} \mid i \in S\} \in \Omega^L \\ \{t_i \mid i \in S\} \in \Omega^D}} P(\mathcal{T}_1 = t_1, \mathcal{T}_2 = t_2, \ldots, \mathcal{T}_s = t_s) \cdot \\ &P(\mathcal{F}_1 = f_1^{(t_1)}, \mathcal{F}_2 = f_2^{(t_2)}, \ldots, \mathcal{F}_s = f_s^{(t_s)}) \cdot \\ &P(\mathcal{X}_1 = x_1^{(t_1)}, \mathcal{X}_2 = x_2^{(t_2)}, \ldots, \mathcal{X}_s = x_s^{(t_s)} \mid \mathcal{F}_1 = f_1^{(t_1)}, \\ &\mathcal{F}_2 = f_2^{(t_2)}, \ldots, \mathcal{F}_s = f_s^{(t_s)}). \end{aligned} \tag{47}$$

According to the Hammersley-Clifford theorem, Eq. (47) can be written in the form of energy function $U(\cdot)$:

$$\begin{aligned} \{f^*, t^*\} = &\mathop{\arg\min}_{\substack{\{f_i^{(t_i)} \mid i \in S\} \in \Omega^L \\ \{t_i \mid i \in S\} \in \Omega^D}} U(\mathcal{T}_1 = t_1, \mathcal{T}_2 = t_2, \ldots, \mathcal{T}_s = t_s) \\ &+ \beta U(\mathcal{F}_1 = f_1^{(t_1)}, \mathcal{F}_2 = f_2^{(t_2)}, \ldots, \mathcal{F}_s = f_s^{(t_s)}) \\ &+ \gamma\, U(\mathcal{X}_1^{(t_1)} = x_1^{(t_1)}, \mathcal{X}_2^{(t_2)} = x_2^{(t_2)}, \ldots, \mathcal{X}_s^{(t_s)} \\ &= x_s^{(t_s)} \mid \mathcal{F}_1 = f_1^{(t_1)}, \mathcal{F}_2 = f_2^{(t_2)}, \ldots, \mathcal{F}_s = f_s^{(t_s)}) \end{aligned} \tag{48}$$

where β and γ are the weighting parameters, and the equation

$$\begin{aligned} \mathcal{E}(f, t) = &U(\mathcal{T}_1 = t_1, \mathcal{T}_2 = t_2, \ldots, \mathcal{T}_s = t_s) \\ &+ \beta U(\mathcal{F}_1 = f_1^{(t_1)}, \mathcal{F}_2 = f_2^{(t_2)}, \ldots, \mathcal{F}_s = f_s^{(t_s)}) \\ &+ \gamma\, U(\mathcal{X}_1^{(t_1)} = x_1^{(t_1)}, \mathcal{X}_2^{(t_2)} = x_2^{(t_2)} \ldots, \mathcal{X}_s^{(t_s)} \\ &= x_s^{(t_s)} \mid \mathcal{F}_1 = f_1^{(t_1)}, \mathcal{F}_2 = f_2^{(t_2)}, \ldots, \mathcal{F}_s = f_s^{(t_s)}) \end{aligned} \tag{49}$$

can be denoted by

$$\mathcal{E}(f, t) = U(t) + \beta U(f^{(t)}) + \gamma\, U(x^{(t)} \mid f^{(t)}). \tag{50}$$

When x is regarded as a parameter, we can obtain Eq. (15).

REFERENCES

[1] D. Stork and M. Hennecke, *Speechreading by Humans and Machines*. New York, NY, USA: Springer-Verlag, 1996.

[2] I. Matthews, T. F. Cootes, J. A. Bangham, S. Cox, and R. Harvey, “Extraction of visual features for lipreading,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 2, pp. 198–213, Feb. 2002.

[3] S. L. Wang, A. Liew, W. H. Lau, and S. H. Leung, “An automatic lipreading system for spoken digits with limited training data,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 2, pp. 1760–1765, Dec. 2008.

[4] H. E. Cetingul, Y. Yemez, E. Erzin, and A. M. Tekalp, “Discriminative analysis of lip motion features for speaker identification and speech-reading,” *IEEE Trans. Image Process.*, vol. 15, no. 10, pp. 2879–2891, Oct. 2006.

[5] G. N. Kodandaramaiah, M. B. Manjunatha, S. A. K. Jilani, M. N. G. Prasad, R. B. Kulkarni, and M. M. Rao, “Use of lip synchronization by hearing impaired using digital image processing for enhanced perception of speech,” in *Proc. 2nd Int. Conf. Comput., Control, Commun.*, New Delhi, India, Feb. 2009, pp. 1–7.

[6] G. I. Chiou and J.-N. Hwang, “Lipreading from color video,” *IEEE Trans. Image Process.*, vol. 6, no. 8, pp. 1192–1195, Aug. 1997.

[7] A. C. Hulbert and T. A. Poggio, “Synthesizing a color algorithm from examples,” *Science*, vol. 239, no. 4839, pp. 482–485, 1998.

[8] M. Liévin and F. Luthon, “Nonlinear color space and spatiotemporal MRF for hierarchical segmentation of face features in video,” *IEEE Trans. Image Process.*, vol. 13, no. 1, pp. 63–71, Jan. 2004.

[9] G. Potamianos and C. Neti, “Audio-visual speech recognition in challenging environments,” in *Proc. Eur. Conf. Speech Commun. Technol.*, Geneva, Switzerland, 2003, pp. 1293–1296.

[10] H. Mirzaalian and G. Hamarneh, “Vessel scale-selection using MRF optimization,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, San Francisco, CA, USA, Jun. 2010, pp. 3273–3279.

[11] Y.-M. Cheung and M. Li, “MAP-MRF based LIP segmentation without true segment number,” in *Proc. 18th IEEE Int. Conf. Image Process.*, Brussels, Belgium, Sep. 2011, pp. 796–772.

[12] S. Wang, W. Lau, W. Liew, and S. Leung, “Robust lip region segmentation for lip images with complex background,” *Pattern Recognit.*, vol. 40, no. 12, pp. 3481–3491, Dec. 2007.

[13] T. Wark, S. Sridharan, and V. Chandran, "An approach to statistical lip modelling for speaker identification via chromatic feature extraction," in *Proc. 14th Int. Conf. Pattern Recognit.*, Brisbane, Australia, Aug. 1998, pp. 123–125.

[14] X. Zhang and R. M. Mersereau, "Lip feature extraction towards an automatic speechreading system," in *Proc. Int. Conf. Image Process.*, Vancouver, BC, Canada, 2000, pp. 226–229.

[15] M. Pardàs and E. Sayrol, "Motion estimation based tracking of active contours," *Pattern Recognit. Lett.*, no. 22, no. 13, pp. 1447–1456, 2001.

[16] P. Delmas, N. Eveno, and M. Liévin, "Towards robust lip tracking," in *Proc. 16th Int. Conf. Pattern Recognit.*, Quebec, QC, Canada, 2002, pp. 528–531.

[17] N. Eveno, A. Caplier, and P.-Y. Coulon, "Accurate and quasi-automatic lip tracking," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 5, pp. 706–715, May 2004.

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

[19] 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, ON, Canada, May 2006, pp. 2021–2024.

[20] 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, no. 9, pp. 2215–2223, May 2008.

[21] Q. Nguyen and M. Milgram, "Online active feature model for lip tracking," in *Proc. 12th Int. Conf. Comput. Vis. Workshops*, Kyoto, Japan, Sep./Oct. 2009, pp. 368–375.

[22] B. Beaumesnil and F. Luthon, "Real time tracking for 3D realistic lip animation," in *Proc. 18th Int. Conf. Pattern Recognit.*, Hong Kong, China, 2006, pp. 219–222.

[23] R. Rohani, S. Alizadeh, F. Sobhanmanesh, and R. Boostani, "Lip segmentation in color images," in *Proc. Int. Conf. Innov. Inform. Technol.*, Al Ain, UAE, Dec. 2008, pp. 747–750.

[24] E. Skodras and N. Fakotakis, "An unconstrained method for lip detection in color images," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Prague, Czech Republic, May 2011, pp. 1013–1016.

[25] P. Gacon, P.-Y. Coulon, and G. Bailly, "Non-linear active model for mouth inner and outer contours detection," in *Proc. 13th Eur. Signal Process. Conf.*, Antalya, Turkey, 2005, pp. 473–476.

[26] C. Bouvier, P.-Y. Coulon, and X. Maldague, "Unsupervised lips segmentation based on ROI optimisation and parametric model," in *Proc. IEEE Int. Conf. Image Process.*, San Antonio, TX, USA, Sep./Oct. 2007, pp. IV-301–IV-304.

[27] M. Li and Y.-M. Cheung, "Automatic segmentation of color lip images based on morphological filter," in *Proc. 20th Int. Conf. Artif. Neural Netw.*, Thessaloniki, Greece, 2010, pp. 384–387.

[28] S. Wang, A. Liew, W. H. Lau, and S. Leung, "Lip region segmentation with complex background," in *Visual Speech Recognition: Lip Segmentation and Mapping*, A. W. C. Liew and S. Wang, Eds. Hershey, PA, USA: IGI Global, 2009.

[29] 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, no. 4, pp. 542–549, Aug. 2003.

[30] A. W.-C. Liew and H. Yan, "An adaptive spatial fuzzy clustering algorithm for 3-D MR image segmentation," *IEEE Trans. Med. Imag.*, vol. 22, no. 9, pp. 1063–1075, Sep. 2003.

[31] S.-H. Leung, S.-L. Wang, and W.-H. Lau, "Lip image segmentation using fuzzy clustering incorporating an elliptic shape function," *IEEE Trans. Image Process.*, vol. 13, no. 1, pp. 51–62, Jan. 2004.

[32] M. Liévin, P. Delmas, P. Y. Coulon, F. Luthon, and V. Fristol, "Automatic lip tracking: Bayesian segmentation and active contours in a cooperative scheme," in *Proc. IEEE Int. Conf. Multimedia Comput. Syst.*, Barcelona, Spain, Jul. 1999, pp. 691–696.

[33] M. Liévin and F. Luthon, "Unsupervised lip segmentation under natural conditions," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Phoenix, AZ, USA, Mar. 1999, pp. 3065–3068.

[34] X. Zhang, R. M. Mersereau, M. Clements, and C. C. Broun, "Visual speech feature extraction for improved speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Orlando, FL, USA, May 2002, pp. II-1993–II-1996.

[35] I. Maglogiannis, D. Vouyioukas, and C. Aggelopoulos, "Face detection and recognition of natural human emotion using Markov random fields," *Pers. Ubiquitous Comput.*, vol. 13, no. 1, pp. 95–101, Jan. 2009.

[36] S. Bakshi, R. Raman, and P. K. Sa, "Lip pattern recognition based on local feature extraction," in *Proc. Annu. IEEE India Conf.*, Dec. 2011, pp. 1–4.

[37] J. Besag, "Spatial interaction and the statistical analysis of lattice systems," *J. Roy. Statist. Soc. Ser. B*, vol. 36, no. 2, pp. 192–236, 1974.

[38] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the bayesian restoration of images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-6, no. 6, pp. 721–741, Nov. 1984.

[39] S. Li, *Markov Random Field Modeling in Image Analysis*, 3rd ed. New York, NY, USA: Springer-Verlag, 2009.

[40] T. Lindeberg, "Principles for automatic scale selection," Dept. Numer. Anal. Comput. Sci., KTH (Roy. Inst. Technol.), Stockholm, Sweden, Tech. Rep. ISRN KTH NA/P-98/14-SE, 1999.

[41] P. Soille, *Morphological Image Analysis: Principles and Applications*. New York, NY, USA: Springer-Verlag, 1999.

[42] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst., Man, Cybern.*, vol. 9, no. 1, pp. 62–66, Jan. 1979.

[43] A. M. Martinez and R. Benavente, "The AR face database," Centre Visió Computador, Universitat Autònoma Barcelona, Bellaterra, Spain, CVC Tech. Rep. 24, Jun. 1998.

[44] F. Solina, P. Peer, B. Batagelj, S. Juvan, and J. Kovač, "Color-based face detection in the '15 seconds of fame' art installation," in *Proc. Conf. Comput. Vis./Comput. Graph. Collaboration Model-Based Imag., Rendering, Image Anal. Graph. Special Effects*, Versailles, France, Mar. 2003, pp. 38–47.

[45] F. Tarrs and A. Rama. (2013, Mar.). *GTAV face database* [Online]. Available: http://gps-tsc.upc.es/gtav/researchareas/upcfacedatabase/gtavfacedatabase.htm

[46] C. Sanderson and K. K. Paliwal, "Polynomial features for robust face authentication," in *Proc. Int. Conf. Image Process.*, New York, NY, USA, Jun. 2002, pp. 997–1000.

[47] K.-H. Liang and T. Tjahjadi, "Adaptive scale fixing for multiscale texture segmentation," *IEEE Trans. Image Process.*, vol. 15, no. 1, pp. 249–256, Jan. 2006.

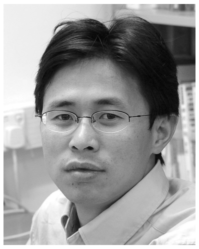
[48] M. K. Agoston, *Computer Graphics and Geometric Modeling: Implementation and Algorithms*. New York, NY, USA: Springer-Verlag, 2005.

[49] T. Caliński and J. Harabatz, "A dendrite method for cluster analysis," *Commun. Statist.*, vol. 3, no. 1, pp. 1–26, 1974.

Yiu-Ming Cheung (SM'06) received the Ph.D. degree from the Department of Computer Science and Engineering, Chinese University of Hong Kong, Hong Kong, in 2000. He is currently a Professor with the Department of Computer Science, Hong Kong Baptist University, Hong Kong. His research interests include machine learning, information security, signal processing, pattern recognition, and data mining. He is the Founding Chair of the Computational Intelligence Chapter of the IEEE Hong Kong Section. He is also a Senior Member of the Association for Computing Machinery.

Meng Li is currently pursuing the Ph.D. degree with the Department of Computer Science, Hong Kong Baptist University, Hong Kong. He received the B.E. degree from the Department of Automatic Test and Control, Harbin Institute of Technology, Harbin, China, in 2004, and the M.E. degree from the Department of General and Fundamental Mechanics, Harbin Institute of Technology, in 2007. His research interests include human lip segmentation and Markov random field-based image processing.

Xiaochun Cao received the B.E. and M.E. degrees in computer science from Beihang University, Beijing, China, and the Ph.D. degree in computer science from the University of Central Florida, Orlando, FL, USA, with his dissertation nominated for the University Level Outstanding Dissertation Award. He spent around three years with ObjectVideo Inc., Reston, VA, USA, as a Research Scientist. From 2008 to 2012, he was a Professor with Tianjin University, Tianjin, China. Since 2012, he has been a Professor with the Institute of Information Engineering, Chinese Academy of Sciences, Beijing. He has authored and co-authored over 100 journal and conference papers. He was a recipient of the Piero Zamperoni Best Student Paper Award at the International Conference on Pattern Recognition in 2004 and 2010.



Xinge You received the B.S. and M.S. degrees in mathematics from Hubei University, Wuhan, China, in 1990 and 2000, respectively, and the Ph.D. degree from the Department of Computer Science, Hong Kong Baptist University, Hong Kong, in 2004. He is currently a Professor with the Department of Electronics and Information Engineering, Huazhong University of Science and Technology, Wuhan. His current research interests include wavelets and its application, signal and image processing, pattern recognition, machine learning, and computer vision.