Segmentation of Color Lip Images by Spatial Fuzzy Clustering

Alan Wee-Chung Liew, *Member, IEEE,* Shu Hung Leung, *Member, IEEE,* and Wing Hong Lau, *Member, IEEE*

***Abstract*—In this paper, we describe the application of a novel spatial fuzzy clustering algorithm to the lip segmentation problem. The proposed spatial fuzzy clustering algorithm is able to take into account both the distributions of data in feature space and the spatial interactions between neighboring pixels during clustering. By appropriate pre- and postprocessing utilizing the color and shape properties of the lip region, successful segmentation of most lip images is possible. Comparative study with some existing lip segmentation algorithms such as the hue filtering algorithm and the fuzzy entropy histogram thresholding algorithm has demonstrated the superior performance of our method.**

***Index Terms*—Color lip segmentation, local spatial interactions, spatial fuzzy clustering.**

I. Introduction

DYNAMIC visual information from the lip movement can significantly improve the accuracy and robustness of an automatic speech recognition system in a noisy environment [1], [2]. Useful geometric information about lip movement, such as the temporal variation of mouth width and height, can be obtained easily from a segmented lip. However, accurate lip segmentation has proved to be difficult due to the weak color contrast and the significant overlap in color features between the lip and the face regions.

Many image segmentation techniques have been proposed [3]–[7]. For color image segmentation, histogram-based and clustering-based methods have been widely used. In [6], [7] Cheng *et al.* proposed a histogram segmentation technique which involves performing a fuzzy partition on a two-dimensional (2-D) histogram based on the maximum fuzzy entropy principle. The 2-D histogram is constructed such that the horizontal axis represents one of the pixel color component and the vertical axis represents the local average of that color component, thus capturing the local spatial interactions between neighboring pixels. Color segmentation is obtained by performing the fuzzy partition on each of the color components and then combining the results. However, as the color components are treated independently, consistent segmentation of a region at every color components is impossible. In addition, the maximum fuzzy entropy criterion does not always correspond to good segmentation, as noted in [8].

Manuscript received March 4, 2002; revised August 12, 2002. This work was supported by CERG under Grant 9040272.

A. W.-C. Liew and W. H. Lau are with the Department of Computer Engineering and Information Technology, City University of Hong Kong, Kowloon Tong, Hong Kong.

S. H. Leung is with the Department of Electronic Engineering, City University of Hong Kong, Kowloon Tong, Hong Kong.

Digital Object Identifier 10.1109/TFUZZ.2003.814843

There are very few reports on color lip segmentation. In [8], the color lip region is segmented using a fuzzy thresholding algorithm with connectivity processing. The histogram of R/G averaged over several image frames and updated with time is used to determine an adaptive threshold T, obtained by minimizing a fuzzy entropy measure. Pixels above T are treated as definite lip pixels whereas pixels below T but above a heuristic point are considered definite lip pixels if they have eight-neighbor connectivity to a definite lip pixel. This algorithm, however, shares the same difficulties as in [6] and [7], namely, unable to handle multicomponent feature vector and the lack of clear relationship between the fuzzy entropy measure and good segmentation. In [9], a hue filter is used to weight the red hue, which is presumed to be the lip region, preferentially. The lip region is then segmented by thresholding. This method requires a hue distribution to be prespecified for the lip region. In practice, however, the lip hue varies between speakers, and is also affected by illumination and by the use of makeup.

In our approach, color lip segmentation is treated as a two-class clustering and segmentation problem. Clustering-based method allows the segmentation to be based on color difference between lip and nonlip regions, without assuming a particular hue distribution for the lip. A novel spatial fuzzy C-means (FCM) clustering algorithm [10], [11] is employed. With appropriate pre- and postprocessing, good lip segmentation results can be obtained.

II. Color Lip Segmentation Problem

The original lip images are in the RGB color format. It is desirable to work in a uniform color space, where the l_2 distance between two points in the color space is directly proportional to the perceived color difference. Two approximately uniform color spaces are the 1976 CIELAB color space (L^*, a^*, b^*) and the 1976 CIELUV color space (L^*, u^*, v^*) [12].

Although a number of researchers [13], [14] have suggested that the skin hue is fairly consistent across different people, small variation in hue can be significant in view of the weak color contrast between skin and lip regions. For many lip images, the colors of the lip and skin region usually overlap considerably. In addition, the lip color of a person can significantly overlap the skin color of another person.

The color lip segmentation problem thus involves the delineation of the lip region from the face region based on the color contrast between these two regions. The lip region is assumed to be a single connected component within the lip image, with a shape that is approximately elliptical, and with a detectable color contrast.

III. Fuzzy Clustering With Spatial Continuity

Fuzzy cluster analysis has been a powerful tool in the field of pattern recognition [15]–[17]. A conventional FCM algorithm classifies pixel data based solely on their feature space distribution without explicitly considering the spatial interactions between neighboring pixels. Although it is possible to include the coordinate information as features [16], [17], such approach will result in loose clusters even for perfect data since the coordinate information does not form a compact mass in feature space. Moreover, the clustering result will be influenced by the coordinate system chosen when the coordinates are used as features. For image data, pixels with similar features are usually found together forming homogenous patches and should therefore be assigned to the same cluster. However, image noise may alter the feature value of a pixel to the extent that it is misclassified. In addition, many pixels in real images are ambiguous. The incorporation of spatial information can resolve this ambiguity and yields better classification result.

We consider a 3×3 image window. If the 3×3 patch belongs to the same class, then the center pixel should be smoothed by its neighboring pixels so that eventually all pixels in the window have high and similar membership values in one of the clusters. Now, consider the feature vector $\boldsymbol{x}_{r,s}$ and its topological neighbor $\boldsymbol{x}_{r-1,s-1}$. Let ∂ be the l_2 distance between them, i.e., $\partial_{\{(r,s),(r-1,s-1)\}} = \|\boldsymbol{x}_{r,s} - \boldsymbol{x}_{r-1,s-1}\|$. Let $d_{i,r,s}$ be the l_2 distance between $\boldsymbol{x}_{r,s}$ and the cluster centroid $\boldsymbol{v}_i$. If $\partial_{\{(r,s),(r-1,s-1)\}}$ is small (i.e., similar in feature), we would like $d_{i,r,s}$ to be greatly influence by $d_{i,r-1,s-1}$. Otherwise, $d_{i,r,s}$ should be largely independent of $d_{i,r-1,s-1}$. Taking the eight-neighborhoods into account, we define a dissimilarity index $D_{i,r,s}$ which measure the dissimilarity between $x_{r,s}$ and $\boldsymbol{v}_i$

$$D_{i,r,s} = \frac{1}{8}\sum_{l_1=-1}^{1}\sum_{l_2=-1}^{1}\left[d_{i,r,s}^2\lambda_{l_1,l_2}^{r,s} + d_{i,r+l_1,s+l_2}^2(1-\lambda_{l_1,l_2}^{r,s})\right], \qquad (l_1, l_2) \neq (0,0) \tag{1}$$

where $\lambda(\partial_{\{(r,s),(r+i,s+j)\}}) = \lambda_{i,j}^{r,s}$ is the weighting factor controlling the degree of influence of the neighboring pixels $(r+i, s+j)$ on the center pixel (r, s)

$$\lambda(\partial) = \frac{1}{1+e^{-(\partial-\mu)/\sigma}} \tag{2}$$

and μ, σ specifies the displacement of λ from 0, and the steepness of the sigmoid curve, respectively.

Note that $D_{i,r,s}$, in effect, smoothes the cluster assignment (via the distance to centroid term) of the center pixel by the cluster assignment of the neighboring pixels. This is different from a noise filtering perspective in [18]. When the center pixel is along edges, its feature value will be very different from that of its neighbors, reflecting the unlikeliness that they belong to the same class. Hence, ∂ will be large and $\lambda \to 1$ for all its neighbors. In this case, $D_{i,r,s} \approx d_{i,r,s}^2$, i.e., neighboring influence is turned off. When the window is on a step boundary, the center pixel is only affected by the neighboring pixels in the same class (i.e., on the same step level) as the center pixel. When the center pixel is on a smooth region and is affected by all its neighbors, the degree of influence of each neighbor on the center pixel is determined by the similarity of the neighbor's feature value with the center pixel. Hence, $D_{i,r,s}$ enables local spatial interactions between neighboring pixels that is adaptive to image content. $D_{i,r,s}$ can be easily modified to allow larger region of influence by using a larger window. Weighting can also be applied to the neighboring pixels such that more distance pixels become less relevant.

The parameter μ in (2) can be viewed as the average "randomness" of the homogeneous region. It takes into account the noise present in homogeneous region. When the difference in feature value between the center pixel and its neighbor is larger than the average "randomness," i.e., $\partial > \mu$, the center pixel and the neighboring pixel is less likely to belong to the same class and the influence on the center pixel is suppressed in $D_{i,r,s}$. Let us denote

$$\partial_{av}(r, s) = \frac{1}{8}\sum_{l_1=-1}^{1}\sum_{l_2=-1}^{1}\partial_{\{(r,s),(r+l_1,s+l_2)\}} \qquad (l_1, l_2) \neq (0, 0) \tag{3}$$

as the average ∂ for a 3×3 window centered at (r, s). Assuming that in real images, most 3×3 windows fall on homogeneous region, then μ can be set to be the average of $\partial_{av}(r, s)$ over all (r, s). Since σ controls the slope of the sigmoid curve, it can be made adaptive to the image content and noise property as well. Clearly, σshould be chosen such that the clustering results of important image structures are not smoothed out, i.e., $\lambda(\partial) \approx 1$ when ∂ is due to genuine structures, such as region borders or edges, in the image. We determine σ as follows. From the $\partial_{av}(r, s)$ computed over the image data, we take ∂_t equal to the 95th percentile of $\partial_{av}(r, s)$. Then, we let $\lambda(\partial_t) = 0.8$ and solve for σ using (2).

With $D_{i,r,s}$, both the feature space information and the local spatial interactions between neighboring pixels (spatial space information) can be incorporated into the fuzzy clustering. Due to the formulation of $D_{i,r,s}$, no explicit weighting of the contributions of the feature space information and the spatial space information is needed, thus avoiding the difficulty of choosing an appropriate weight. For an image data I of dimension n_1 by n_2, and the number of cluster c, the objective functional of the spatial FCM (SFCM) clustering algorithm is given by

$$J_m(U, v) = \sum_{r=1}^{n_1}\sum_{s=1}^{n_2}\sum_{i=1}^{c} u_{i,r,s}^m D_{i,r,s}$$

$$\text{subjected to} \quad \sum_{i=1}^{c} u_{i,r,s} = 1 \qquad \forall (r, s) \in I. \tag{4}$$

The $c \times n_1 n_2$ matrix U is a fuzzy c-partition of the set of image feature vector X from I, v is the set of fuzzy cluster centroids, $m \in (1, \infty)$ defines the fuzziness of the clustering results and $u_{i,r,s}$ gives the membership of pixel (r, s) in fuzzy cluster C_i.

It can be shown [11] that a local minimum of J_m can be reached by performing Picard iteration through (5) and (7), when $D_{i,r,s}$ is nonzero for $i = 1$ to c

$$v_i = \frac{\sum_{r=1}^{n_1}\sum_{s=1}^{n_2} u_{i,r,s}^m \hat{x}_{r,s}}{\sum_{r=1}^{n_1}\sum_{s=1}^{n_2} u_{i,r,s}^m} \tag{5}$$

$$\hat{x}_{r,s} = \frac{1}{8} \sum_{l_1=-1}^{1} \sum_{l_2=-1}^{1} \left[\lambda_{l_1,l_2}^{r,s} x_{r,s} + (1 - \lambda_{l_1,l_2}^{r,s}) x_{r+l_1,s+l_2} \right] \quad (l_1, l_2) \neq (0, 0) \tag{6}$$

$$u_{i,r,s} = \left[\sum_{j=1}^{c} \left(\frac{D_{i,r,s}}{D_{j,r,s}} \right)^{1/(m-1)} \right]^{-1}. \tag{7}$$

In (5) and (6), $\boldsymbol{v}_i$ and $\hat{\boldsymbol{x}}_{r,s}$ denote the ith fuzzy cluster centroid and the locally smoothed feature vector at pixel (r, s), respectively. When at least one $D_{i,r,s}$ is zero, i.e., at a singularity, then $u_{j,r,s}$ is zero when $D_{j,r,s}$ is nonzero, and $u_{i,r,s}$ is set to be equal to one over the number of zero D terms. In practice, the Picard iteration is terminated when the l_∞ difference between two consecutive iterations of the fuzzy c-partition matrix U falls below a small threshold.

Although in theory, the Picard iteration only ensures convergence to a local minimum of J_m, the incorporation of spatial interactions between neighboring pixels in the clustering seems to make the convergence more consistent in the sense that the same solution is observed for different starting points. This is in contrast to the FCM algorithm where the final result is very dependent on the starting points. Details on the SFCM algorithm and its image segmentation performance compare to the FCM algorithm can be found in [11].

IV. Lip Image Preprocessing

The lip image is first transformed into the CIELAB, CIELUV color spaces. A feature vector consisted of $\{L^*, a^*, b^*, u^*, v^*\}$ is constructed for each pixel in the image. Then, the intensity nonuniformity (i.e., smooth variation in intensity value) in the L^* component due to uneven illumination is reduced by the following procedure.

1) Estimate the intensity nonuniformity along the column direction on the upper and lower border of the image using a small window and denote them by $u(j)$ and $l(j)$, respectively.
2) Compute the mean value m by averaging $u(j)$ and $l(j)$.
3) Modify the luminance value $L^*(i, j)$ for each pixel along each column j to

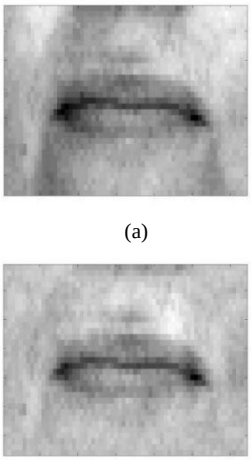
$$\hat{L}^*(i, j) = L^*(i, j) + \frac{l(j) - u(j)}{row - 1}(i - 1) + m - l(j). \tag{8}$$

Fig. 1(a) and (b) shows the luminance image of a mouth region before and after equalization, respectively. The obvious intensity nonuniformity below the two corners of the mouth has been reduced.

The teeth region, the low luminance region (where the chrominance information is noisy), and the high luminance pixels caused by highlight can adversely affect the clustering result. The teeth region has been observed to have a lower a^* and u^* value than the rest of the lip image and has a fairly consistent hue. From experimentation with different lip images, the teeth thresholds t_a and t_u was found to be given by

$$t_a = \begin{cases} \mu_a - \sigma_a, & \text{if } (\mu_a - \sigma_a) < 9 \\ 9, & \text{otherwise} \end{cases} \tag{9}$$

$$t_u = \begin{cases} \mu_u - \sigma_u, & \text{if } (\mu_u - \sigma_u) < 29 \\ 29, & \text{otherwise} \end{cases} \tag{10}$$



(a) (b)

Fig. 1. (a) Original luminance image. (b) Equalized luminance image.

where μ_a, σ_a and μ_u, σ_u are the mean and standard deviation of a^* and u^*, respectively. Possible teeth pixels, i.e., $a^* \leq t_a$ or $u^* \leq t_u$, or pixels with $L^* \leq 35\%$ or $L^* \geq 95\%$ of the reference white, are masked out from subsequent clustering stage. Finally, the chrominance components are normalized by restricting their range to within ± 2 standard deviations around the mean.

V. Lip Membership Computation and Postprocessing

A. Lip Membership Computation

The settings for the spatial fuzzy clustering algorithm are $m = 2$ and $c = 2$. Mean feature vectors for the skin class and the lip class from hand labeled training images are used as initial cluster centroids. After the final cluster centroids for the lip class and the skin class are obtained, the lip membership map is computed. In order for the teeth region to be included in the lip region, the lip membership values of teeth region are set to 0.75.

B. Morphological Filtering

Grayscale morphological closing and opening [19] with an eight-neighborhood structuring element is used to smooth the membership map and eliminate small erroneous blobs and holes. The morphological closing is realized by performing a maximum operation, followed by a minimum operation. Since the object of interest has higher grey value than the background, this operation will fill up small holes and gaps. The morphological opening is realized by reversing the maximum and minimum operations and has the effect of opening up small-connected regions and protrusions.

C. Symmetry Processing

Large spurious protrusions, occasionally found around the upper or lower lip boundary, cannot be eliminated by the morphological filtering operation. Taking advantage of the symmetry of the left and right side of the lip with frontal view can eliminate these protrusions. The x-coordinate of the left lip corner, x_l, is found by scanning from the left and detecting the first occurrence of a group of 5 pixels with membership value > 0.5, arranged column-wise, and located approximately in the

center row. The x-coordinate of the right lip corner, x_r, is detected likewise and the x-coordinate of the lip center, x_c, can then be found. Next, the row-wise integral projection from the left lip corner to the lip center given in (11) is computed for every row y

$$\zeta_l(y) = \sum_{x=x_l}^{x_c} z(x, y) \tag{11}$$

where

$$z(x, y) = \begin{cases} 1, & \text{if } m(x, y) > 0.5 \\ 0, & \text{otherwise} \end{cases}$$

and $m(x, y)$ denotes the lip membership. The row-wise integral projections of the right, $\zeta_r(y)$, is obtained similarly. By scanning downward from the top, the upper y-coordinate of the lip region is determined if the two following conditions are satisfied:

$$\begin{aligned} &(1) \quad \zeta_l(y) > 0 \quad \cap \quad \zeta_r(y) > 0 \\ &(2) \quad (\zeta_l(y) < 3 * \zeta_r(y)) \quad \cap \quad (\zeta_r(y) < 3 * \zeta_l(y)) . \end{aligned} \tag{12}$$

The lower y-coordinate of the lip region is detected likewise. Lip pixels that are above the upper y-coordinate or below the lower y-coordinate are set to be nonlip.

D. Luminance Processing

Since lip pixels have lower luminance value than skin pixels, the lip membership map can be enhanced by up-weighting the membership value of pixels which are of low luminance value. We first estimate the statistics of the skin pixels by computing the mean, μ_{skin}, and standard deviation, σ_{skin}, of the pixels in a strip of region around the image border. Next, for pixel having a luminance value $v < t_{\text{skin}}$, where $t_{\text{skin}} = \mu_{\text{skin}} - 3.5\sigma_{\text{skin}}$, the difference, $d = t_{\text{skin}} - v$, is computed. For pixel with $v > t_{\text{skin}}$, its d is set to zero. Then, set $d_{\max} = \min(\sigma_{\text{skin}}, \max(d))$. Finally, for any pixel with a membership value $u > 0.45$ or $d > d_{\max}$, u is updated by adding to it a value equals to $d/(2d_{\max})$ if $d \leq d_{\max}$, or 0.5 if $d > d_{\max}$. The modified membership value is then clipped at the maximum membership value in the unmodified membership map. Note that this procedure will have no effect on lips with makeup such that the lip pixels have luminance value larger than the skin pixels.

E. Shape Processing

Prior knowledge about the mouth shape can be used to further reduce inaccurate classification. A best-fit ellipse can be fitted onto the lip membership map to suppress any remaining spurious protrusions. The parameters of the best fit ellipse, i.e., the center of mass, (x_m, y_m), the inclination θ about the center of mass, the semimajor axis x_a, and the semiminor axis y_a, are computed form the lip membership $m(x, y)$ by [20]

$$x_m = \sum_{x=1}^{M}\sum_{y=1}^{N} x * m(x, y) \Big/ \sum_{x=1}^{M}\sum_{y=1}^{N} m(x, y) \tag{13}$$

$$y_m = \sum_{x=1}^{M}\sum_{y=1}^{N} y * m(x, y) \Big/ \sum_{x=1}^{M}\sum_{y=1}^{N} m(x, y) \tag{14}$$

$$\theta = \frac{1}{2}\tan^{-1}\left\{\frac{2\mu_{11}}{\mu_{20} - \mu_{02}}\right\} \tag{15}$$

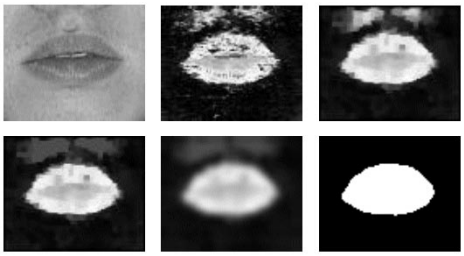


Fig. 2. (Top left) RGB lip image (shown in gray). (Top middle) Lip membership map. (Top right) After morphological filtering. (Bottom left) After symmetry, luminance, and shape processing. (Bottom middle) After Gaussian smoothing. (BOttom right) Final segmentation.

$$x_a = \left(\frac{4}{\pi}\right)^{1/4}\left[\frac{(I_y)^3}{I_x}\right]^{1/8} \tag{16}$$

$$y_a = \left(\frac{4}{\pi}\right)^{1/4}\left[\frac{(I_x)^3}{I_y}\right]^{1/8} \tag{17}$$

with

$$\mu_{pq} = \sum_{x=1}^{M}\sum_{y=1}^{N}(x - x_m)^p(y - y_m)^q m(x, y) \tag{18}$$

$$I_x = \sum_{x=1}^{M}\sum_{y=1}^{N}((y - y_m)\cos\theta - (x - x_m)\sin\theta)^2 m(x, y) \tag{19}$$

$$I_y = \sum_{x=1}^{M}\sum_{y=1}^{N}((y - y_m)\sin\theta + (x - x_m)\cos\theta)^2 m(x, y) \tag{20}$$

where M, N are the column and row dimensions, respectively. Only pixels having membership ≥ 0.5 (i.e., potential lip pixels) enter into the ellipse computation. After obtaining the best-fit ellipse, any potential lip pixels outside of the best-fit ellipse are flagged to nonlip. Finally, the lip membership map is smoothed with a Gaussian filter before thresholding at a threshold of 0.45 and retaining the single largest connected component as the lip region.

VI. Comparative Studies and Experimental Results

Fig. 2 shows a lip segmentation example. The lip region can be identified clearly from the lip membership map produced by the SFCM algorithm. However, there are two spurious blobs above the mouth. After morphological filtering [Fig. 2(c)], the membership map is smoothed and small holes are filled up. Fig. 2(d) shows the membership map after symmetry, luminance and shape processing. The two spurious blobs have been eliminated. Fig. 2(e) shows the membership map after Gaussian smoothing and Fig. 2(f) shows the final segmentation result.

For comparison, the hue filtering segmentation algorithm (HFS) of [9] and the fuzzy entropy histogram thresholding algorithm (FEHT) of [6] are implemented. As each color component is treated independently in the FEHT algorithm, bilevel segmentation of color images cannot be performed directly. Instead, we use the FEHT algorithm to automatically segment the hue-filtered image (a scalar value image). The

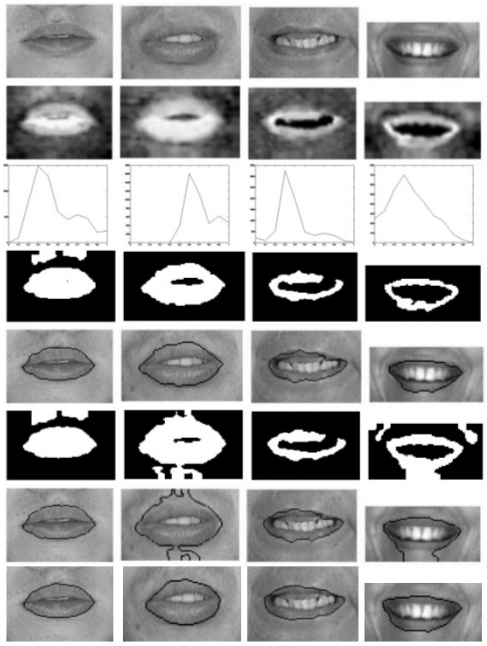


Fig. 3. Lip segmentation using the hue filtering (HFS) algorithm [11] and the fuzzy entropy histogram thresholding (FEHT) [7]. Row 1: RGB lip images (in gray). Row 2: Post-processed hue filtered images. Row 3: Histograms of hue-filtered images. Rows 4 and 5: HFS segmented lip regions with threshold of 0.6, 0.8, 0.6, and 0.65, respectively. Rows 6 and 7: FEHT segmented lip regions. Row 8: Segmented lip regions using our algorithm.

FEHT algorithm performs an exhaustive search to maximize the total fuzzy entropy for a bilevel segmentation of the hue-filtered image. The hue image is computed from CIELAB color space by $hue_{ab} = \tan^{-1}(b^*/a^*)$. In the HFS algorithm, the lip is assumed to be prevalently red. To determine the hue filter, the hue values for pure red, pure green and pure blue are calculated, and are found to be 0.0788 rad, 2.2999 rad and -0.7071 rad, respectively. Then, the two end points where the filter response falls to zero are taken to be midway between pure red and pure green, i.e., 1.189 rad, and midway between pure green and pure blue, i.e., -0.314 rad, respectively. The hue image is first smoothed and then hue-filtered as follows:

$$f(\mathrm{hue}_{ab}) = \begin{cases} 1 - \dfrac{(\mathrm{hue}_{ab} - 0.0788)^2}{1.1102^2} & 0.0788 < \mathrm{hue}_{ab} \le 1.189 \\ 1 - \dfrac{(\mathrm{hue}_{ab} - 0.0788)^2}{0.3928^2} & -0.314 \le \mathrm{hue}_{ab} \le 0.0788 \\ 0, & \text{otherwise.} \end{cases} \tag{21}$$

The hue-filtered image is subject to morphological filtering, luminance processing, and Gaussian smoothing to improve the final segmentation. The symmetry and shape processing are not applied because the threshold for the lip region is not known *a priori*. The final segmentation is obtained by thresholding the hue-filtered image with an optimum threshold, chosen manually to be the valley of the histogram of the hue-filtered image.

Fig. 3 shows some examples of lip segmentation using HFS, FEHT, and our method. The first row shows the original RGB lip images. The second row shows the post-processed hue-filtered images. The third row shows the histograms of the hue-filtered images. For the first example, there are two valleys, at 0.6 and 0.9, and the optimum threshold was found to be 0.6. The optimum thresholds for the second and third examples are 0.8 and 0.6, respectively. No discernable valley exists for the fourth example and the optimum threshold of 0.65 was obtained by trial. The fourth row shows the binary images obtained by thresholding the hue-filtered images using the aforementioned thresholds. By taking into account the teeth region which is detected independently, and retaining the single largest connected region, the final segmented lip regions are shown in row 5. Although the lip segmentation results are sufficiently good, the major difficulty with the HFS algorithm is the determination of the optimum threshold. It can be seen that the optimum threshold varies considerably for different lip images. Row 6 shows the binary images obtained from the thresholds (0.571, 0.714, 0.565, and 0.433, respectively) found automatically using the FEHT algorithm. The fuzzy entropy criterion obviously does not match human perception of good lip segmentation in these examples. Row 7 shows the segmented lip regions using the FEHT algorithm. For comparison, the segmentations obtained using our method are shown in the last row. The quality of segmentation using our method is clearly superior to both the HFS method and the FEHT method.

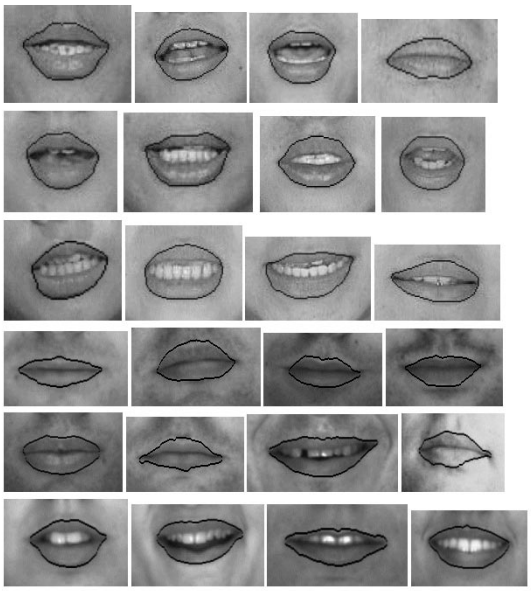


Fig. 4. Some color lip segmentation results obtained using the proposed algorithm.

To evaluate the performance of our algorithm, we carried out lip segmentation experiment with a homebrew database (about 2000 color lip images from more than 20 speakers) and some color lip images taken from the XM2VTS [21] and the AR Face databases [22]. The segmentation results are evaluated based on subjective judgment since no ground truth is available. The algorithm has been observed to perform well and some of the results are shown in Fig. 4, where the first three rows are from our own database, and the last three rows are images from the XM2VTS and AR Face databases. Rows 3 and 6 are female speakers, whereas the rest are male speakers.

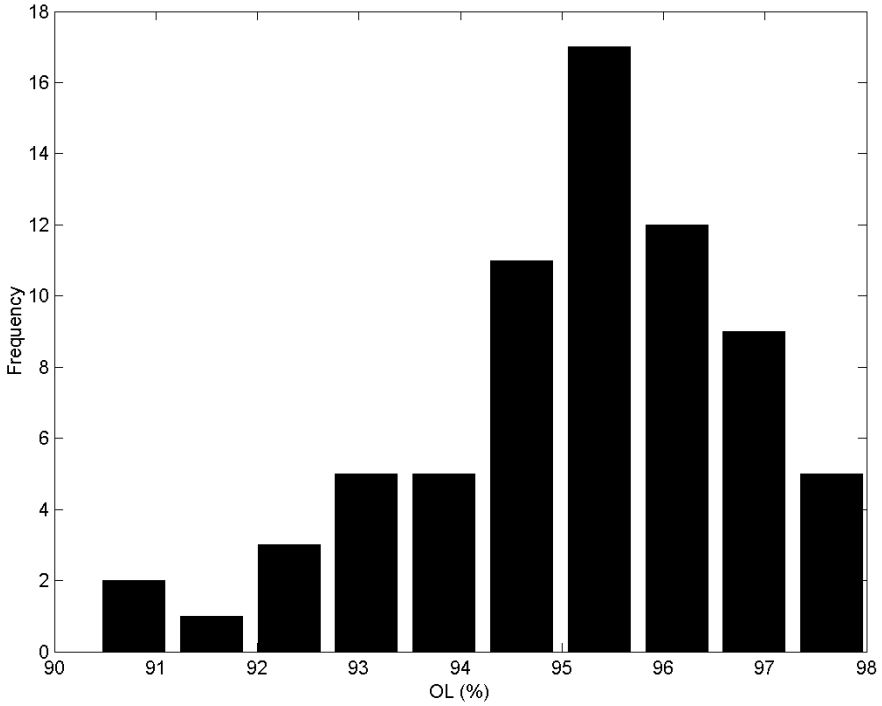


Fig. 5. Percentage overlap of segmented lips with ground truth.

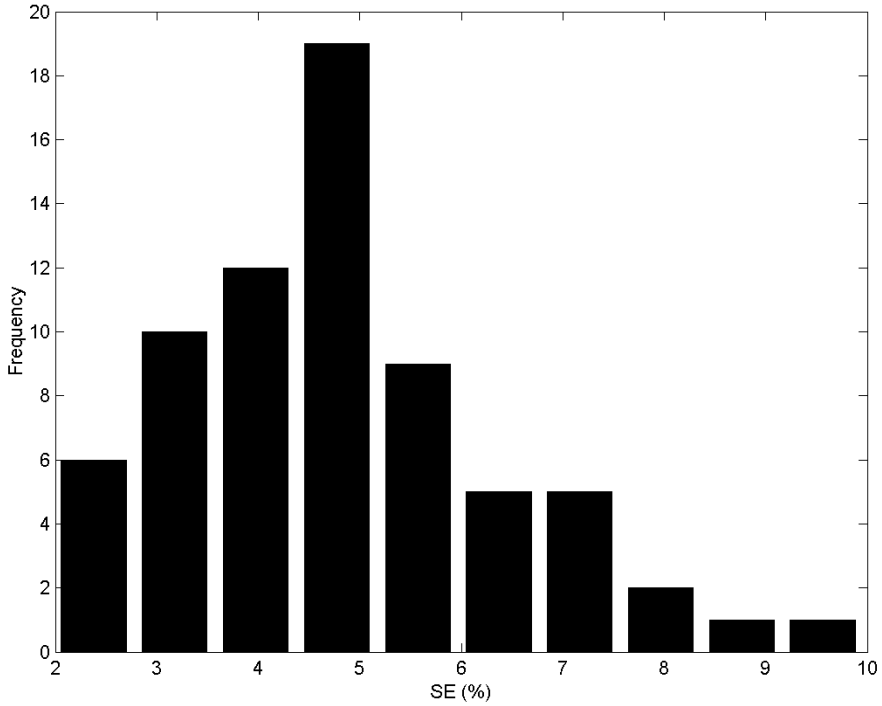


Fig. 6. Percentage segmentation error of segmented lip.

We have also attempted to perform some quantitative evaluations about the performance of our algorithm. We randomly selected 70 images from our database and hand-segmented the lip by manually fitting a parametric lip model [24]. These serve as ground truth. The use of a parametric lip model allows relatively fast manual segmentation, but compromises on the accuracy of some of the extracted lip contours since the model may not fit the actual lip exactly. Two measures are used to evaluate the accuracy of our algorithm. The first measure determines the percentage of overlap (OL) between the segmented lip region A_1 and the ground truth A_2. It is defined as

$$\mathrm{OL} = \frac{2(A_1 \cap A_2)}{A_1 + A_2} * 100\%. \tag{22}$$

Using this measure, total agreement will have an overlap of 100%. The second measure is the segmentation error (SE) defined as

$$\mathrm{SE} = \frac{\mathrm{OLE} + \mathrm{ILE}}{2 * \mathrm{TL}} * 100\%. \tag{23}$$

OLE is the number of nonlip pixels being classified as lip pixels (outer lip error) and ILE is the number of lip-pixels classified as nonlip ones (inner lip error). TL denotes the number of lip-pixels in the ground truth. Total agreement will have an SE of 0%. The segmentation results are presented as bar charts in Fig. 5 (OL) and Fig. 6 (SE). The results indicated that the proposed algorithm was able to give good segmentation, with an OL of $\sim$95% or an SE of $\sim$5%. We remark that some of the error is due to the misfit between the parametric lip model and the actual lip. Based on our experience, this type of error accounts for about 2%–5% in both OL and SE, depending on the actual lip shape. Currently our segmentation algorithm does not consider the presence of facial hair such as moustaches and beards. A preprocessing stage would be needed to mask out those regions.

VII. Conclusion

Accurate lip segmentation is a difficult problem due to the weak color contrast between the lip and the face region. In this paper, the color lip segmentation problem is formulated as a two-class clustering and segmentation problem. The proposed spatial fuzzy clustering algorithm is able to exploit the spatial interactions between neighboring pixels through the use of a novel dissimilarity index. Comparative study with the hue-filtering segmentation algorithm [9] and the fuzzy entropy histogram thresholding algorithm [6] has indicated the superior performance of our algorithm. In particular, our algorithm produces better segmentation, and without the need to determine an optimum threshold for each lip image. Finally, we remark that the lip membership map obtained from the proposed method can be viewed as a lip probability map. Such a probability map can serve as a robust image feature in a deformable-model based contour extraction algorithm employing a probabilistic region-based cost function [23], [24].

References

[1] E. D. Petajan, "Automatic lipreading to enhance speech recognition," *Proc. IEEE Conf. Computer Vision Pattern Recognition*, pp. 40–47, 1985.

[2] C. Bregler, H. Hild, S. Manke, and A. Waibel, "Improving connected letter recognition by lipreading," *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, pp. 557–560, 1993.

[3] N. R. Pal and S. K. Pal, "A review on image segmentation techniques," *Pattern Recogn.*, vol. 26, no. 9, pp. 1277–1294, 1993.

[4] C. L. Huang, T. Y. Cheng, and C. C. Chen, "Color image segmentation using scale space filter and Markov random field," *Pattern Recogn.*, vol. 25, no. 10, pp. 1217–1229, 1992.

[5] S. H. Park, I. D. Yun, and S. U. Lee, "Color image segmentation based on 3-D clustering: Morphological approach," *Pattern Recogn.*, vol. 31, no. 8, pp. 1061–1076, 1998.

[6] H. D. Cheng, Y. H. Chen, and X. H. Jiang, "Thresholding using two-dimensional histogram and fuzzy entropy principle," *IEEE Trans. Image Processing*, vol. 9, pp. 732–735, Apr. 2000.

[7] H. D. Cheng, J. R. Chen, and J. Li, "Threshold selection based on fuzzy C-partition entropy approach," *Pattern Recogn.*, vol. 31, no. 7, pp. 857–870, 1998.

[8] S. Lucey, S. Sridharan, and V. Chandran, "Chromatic lip tracking using a connectivity based fuzzy thresholding technique," presented at the 5th Int. Symp. Signal Processing Applications ISSPA'99, Brisbane, Australia, Aug. 22–25, 1999.

[9] T. Coianiz, L. Torresani, and B. Caprile, "2D deformable models for visual speech analysis," in *Speechreading by Humans and Machines*, D. G. Stork and M. E. Hennecke, Eds. New York: Springer-Verlag, 1996.

[10] A. W. C. Liew, K. L. Sum, S. H. Leung, and W. H. Lau, "Fuzzy segmentation of lip image using cluster analysis," presented at the Eurospeech'99, Budapest, Hungary, Sept. 5–9, 1999.

[11] A. W. C. Liew, S. H. Leung, and W. H. Lau, "Fuzzy image clustering incorporating spatial continuity," *Proc. Inst. Elect. Eng.—Vision, Image, Signal Processing*, vol. 147, no. 2, pp. 185–192, Apr. 2000.

[12] R. W. G. Hunt, *Measuring Color*, 2nd ed. New York, Ellis Horwood Series in Applied Science and Industrial Technology: Ellis Horwood, 1991.

[13] K. Sobottka and I. Pitas, "A novel method for automatic face segmentation, facial feature extraction and tracking," *Signal Processing: Image Commun.*, vol. 12, pp. 263–281, 1998.

[14] M. Ulises, R. Sanchez, J. Matas, and J. Kittler, "Statistical chromaticity-based lip tracking with B-splines," *Proc. IEEE Int. Conf. Acoustic, Speech, Signal Processing*, vol. 4, pp. 2973–2976, 1997.

[15] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*. New York: Plenum, 1981.

[16] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*. Upper Saddle River, NJ: Prentice-Hall, 1988.

[17] R. Krishnapuram and C. P. Freg, "Fitting an unknown number of lines and planes to image data through compatible cluster merging," *Pattern Recogn.*, vol. 25, no. 4, pp. 385–400, Apr. 1992.

[18] Y. Choi and R. Krishnapuram, "A robust approach to image enhancement based on fuzzy logic," *IEEE Trans. Image Processing*, vol. 6, pp. 808–825, June 1997.

[19] R. Klette and P. Zamperoni, *Handbook of Image Processing Operators*. New York: Wiley, 1996.

[20] A. K. Jain, *Fundamentals of Digital Image Processing*. Upper Saddle River, NJ: Prentice-Hall, 1989.

[21] The XM2VTS face database. [Online]http://www.ee.surrey.ac.uk/Research/VSSP/xm2vtsdb/

[22] A. M. Martinez and R. Benavente. (1998, June) The AR face database. [Online]CVC Technical Report #24

[23] A. W. C. Liew, S. H. Leung, and W. H. Lau, "Region-based approach to robust lip contour extraction," *Electron. Lett.*, vol. 36, no. 15, pp. 1272–1274, July 2000.

[24] A. W. C. Liew, S. H. Leung, and W. H. Lau, "Lip contour extraction from color images using a deformable contour," Pattern Recogn., vol. 35, no. 12, pp. 2949–2962, to be published.

Alan Wee-Chung Liew (M'02) received the B.E. degree (first class honors) in electrical and electronic engineering from the University of Auckland, Auckland, New Zealand, and the Ph.D. degree in electronic engineering from the University of Tasmania, Tasmania, Australia, in 1993 and 1997, respectively.

He is currently a Senior Research Fellow in the Department of Computer Engineering and Information Technology, City University of Hong Kong, Hong Kong. His current research interests include image processing, pattern recognition, and bioinformatics.

Shu Hung Leung (M'02) received the B.Sc. degree (first class honors) in electronics from the Chinese University of Hong Kong, Hong Kong, and the M.Sc. and Ph.D. degrees, both in electrical engineering, from the University of California at Irvine, in 1978, 1979, and 1982, respectively.

From 1982 to 1987, he was an Assistant Professor with the University of Colorado, Boulder. Since 1987, he has been with the Department of Electronic Engineering at the City University of Hong Kong, Hong Kong, where he is currently an Associate Professor. He is the leader of the Digital and Mobile Communication Team in the Department of Electronic Engineering. His current research interest is in digital communications, speech and image processing, intelligent signal processing, and adaptive signal processing.

Dr. Leung has served as the Program Chairman of the Signal Processing Chapter of the IEEE Hong Kong Section since 1992; he is now the Vice Chairman of this Chapter. He serves as a Technical Reviewer for a number of international conferences, the IEEE TRANSACTIONS ON SIGNAL PROCESSING, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS, the IEEE TRANSACTIONS ON COMMUNICATIONS, *Proceedings of Institution of Electrical Engineers*, and *Electronics Letters*. He is listed in the *Marquis Who's Who in Science and Engineering* and *Marquis Who's Who in the World*.

Wing-Hong Lau (M'88) received the B.Sc. and Ph.D. degrees in electrical and electronic engineering from University of Portsmouth, Portsmouth, U.K., in 1985 and 1989, respectively.

In 1985, he joined the Microwave Telecommunications and Signal Processing Research Unit of the University of Portsmouth as a Research Assistant. In 1990, he joined the City University of Hong Kong, where he is currently an Associate Professor in the Department of Computer Engineering and Information Technology. His current research interests are in the areas of digital signal processing, digital audio engineering, and visual speech signal processing.

Dr. Lau is currently the Treasurer of the IEEE Hong Kong Section, Committee Member of the IEEE Hong Kong Signal Processing Chapter. He is the Registration Co-Chair of the ICASSP 2003 and also a Member of the International Steering Committee for APCCAS. He was the Chairman of the IEEE Hong Kong Joint Chapter on CAS/COM for 1997 and 1998, and the Registration Co-Chairman of ISCAS'97.