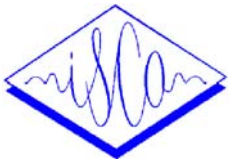

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

FUZZY SEGMENTATION OF LIP IMAGE USING CLUSTER ANALYSIS

Alan W.C. Liew, K.L. Sum, S.H. Leung and W.H. Lau
Department of Electronic Engineering,
City University of Hong Kong.
{eewcliew,eeeugshl,eewhlau}@cityu.edu.hk

ABSTRACT

A clustering-based lip segmentation algorithm is described here. The fuzzy clustering algorithm presented here is able to take into account the local smoothness property of image data. The objective functional of our algorithm utilizes a new distance metric that takes into account the influence of the neighboring pixels on the centre pixel in a 3 by 3 window. Computational steps involved in the segmentation of color lip image and segmentation result for a set of color lip images are given.

Keywords: lip segmentation, spatial fuzzy clusteing.

1. INTRODUCTION

Lip information can significantly improve the performance of automatic speech recognition especially in noisy environmnet[1]. Simple geometric information such as mouth width and height can be obtained from a segmented lip image. One method for lip segmentation is based on color thresholding as proposed in [2]. In this method, the lip region is assumed to be consisted of predominantly red hue. A hue filter is used to weight the red hue more heavily. The lip region is then extracted by simple thresholding. Besides the need to specify a hue value for the lip region, the method does not make use of spatial continuity feature typical of real image data.

In this paper, we propose to use a fuzzy c-mean clustering algorithm incorporating spatial context for lip segmentation. Cluster-based segmentation allows the segmentation to be based on color difference between lip and non-lip region. It does not assume a particular hue or pdf for the lip region. In contrast to hard clustering, fuzzy clustering allows overlap between classes and each data can be classified according to the degree of membership associated with each cluster[3]. It is thus better suited for classifying real world data that have inherent uncertainty. However, conventional fuzzy clustering algorithm does not take into account the spatial relationship between image data, i.e., image data are classified independently of their neighbor. For most image data, however, pixels having similar features usually aggregate together to form coherent patches. By making use of fuzzy clustering incorporating this spatial coherent property, the segmentation of an image would be more accurate.

2. SPATIAL FUZZY CLUSTERING

Consider the 3 by 3 window of image pixels as shown in Fig. 1. If the 3 by 3 patch is homogeneous, then we would like the centre pixel to be smoothed by its neighboring pixels so that eventually all 9 pixels in the window are assigned to the same cluster.

Fig. 1. A 3 by 3 pixel window.

Let us define the sigmoid function $\lambda(\partial)$ as

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

The parameter μ specifies the displacement of λ from ∂=0 and σ specifies the steepness of the sigmoid curve. For large ∂, $\lambda \to 1$, else, $\lambda \to 0$.

For every pixel in the image, we can compute the feature distance ∂ between itself and its neighbors. Thus, if $x_{r,c}$ and $x_{r-1,c-1}$ are close together in feature space, then $\partial_{\{(r,c),(r-1,c-1)\}}$ will be small.

Let $d_{i,j,k}$ be the distance between the pixel $x_{j,k}$ and the ith cluster centroid v_i. This distance measures the closeness of pixel $x_{j,k}$ to the cluster C_i. Consider the pixel $x_{r,c}$ and its neighbor $x_{r-1,c-1}$. If their distance $\partial_{\{(r,c),(r-1,c-1)\}}$ is small, we would like $d_{i,r,c}$ to be greatly influenced by $d_{i,r-1,c-1}$. On the other hand, if $\partial_{\{(r,c),(r-1,c-1)\}}$ is large, $d_{i,r,c}$ should be largely independent of $d_{i,r-1,c-1}$. Taking all the 8-neighborhood of $x_{r,c}$ into account and let $\lambda(\partial_{\{(r,c),(r-1,c-1)\}}) = \lambda^{r,c}_{r-1,c-1}$, we define a new distance $\hat{d}_{i,r,c}$ which measures the distance between $x_{r,c}$ and v_i as,

$$\hat{d}^2_{i,r,c} = \sum_{l_1=-1}^{1}\sum_{l_2=-1}^{1}\left[d^2_{i,r,c}\lambda^{r,c}_{r+l_1,c+l_2} + d^2_{i,r+l_1,c+l_2}(1-\lambda^{r,c}_{r+l_1,c+l_2})\right] \qquad (2)$$

where l_1 and l_2 cannot both be zero.

This new distance reflects the spatial influence of the neighboring pixels on $x_{r,c}$. With $\hat{d}_{i,r,c}$, the spatial

context can be incorporated nicely into the fuzzy clustering algorithm. Notice that $\hat{d}_{i,r,c}$ can be viewed as a generalization of the usual distance $d_{i,r,c}$ as it takes into account both the distance to cluster centroid, $d_{i,r,c}$, and the distance between neighboring pixel, $\partial_{\{(r,c),(r-l,c-l)\}}$.

With $\hat{d}_{i,r,c}$ defined, the objective functional for our spatial fuzzy c-mean clustering algorithm is given by

$$J_m(u_{i,k_1,k_2}, v_i) = \sum_{k_1=1}^{n_1}\sum_{k_2=1}^{n_2}\sum_{i=1}^{C} u_{i,k_1,k_2}^m \hat{d}_{i,k_1,k_2} \tag{3}$$

subject to $\sum_{i=1}^{C} u_{i,k_1,k_2} = 1, \quad \forall(k_1,k_2)$.

The parameter $m \in (1,\infty)$ defines the fuzziness of the clustering results, n_1, n_2 are the x and y dimension of the image data respectively and C is the number of cluster in the data. The value u_{i,k_1,k_2} gives the membership of the pixel x_{k_1,k_2} to the cluster C_i.

The solution of $\min_{u_i,k_1,k_2,v_i}\{J_m(u_i,k_1,k_2,v_i)\}$ is the least-squared stationary point of J_m. Using the Lagrange multiplier method, we have

$$u_{i,k_1,k_2} = \sum_{j=1}^{C}\left[\frac{\hat{d}_{j,k_1,k_2}^2}{\hat{d}_{i,k_1,k_2}^2}\right]^{1/(m-1)} \tag{4}$$

$$v_i = \frac{\sum_{k_1=1}^{n_1}\sum_{k_2=1}^{n_2} u_{i,k_1,k_2}^m \hat{x}_{k_1,k_2}}{\sum_{k_1=1}^{n_1}\sum_{k_2=1}^{n_2} u_{i,k_1,k_2}^m} \tag{5}$$

$$\hat{x}_{k_1,k_2} = \frac{1}{8}\sum_{l_1=-1}^{1}\sum_{l_2=-1}^{1}\left[\lambda_{k_1+l_1,k_2+l_2}^{k_1,k_2} x_{k_1,k_2} + \left(1-\lambda_{k_1+l_1,k_2+l_2}^{k_1,k_2}\right)x_{k_1+l_1,k_2+l_2}\right]$$

with $(l_1, l_2) \neq (0,0)$. By iterating (4) and (5), we can reach a stationary point of J_m.

Assuming that the majority of pixels in an image will be in the homogeneous region, the parameter μ in (1) can be given by

$$\mu = \frac{1}{n_1 n_2}\sum_{k_1=1}^{n_1}\sum_{k_2=1}^{n_2}\frac{1}{8}\left(\sum_{l_1=-1}^{1}\sum_{l_2=-1}^{1}\partial_{\{(k_1,k_2),(k_1+l_1,k_2+l_2)\}}\right) \tag{6}$$

with $(l_1, l_2) \neq (0,0)$. The steepness parameter σ in (1) controls the influence of the neighboring pixels on the centre pixel. If the 3 by 3 window is in a non-homogeneous region, i.e., contain edges, then the average

$$\partial_{av} = \frac{1}{8}\left(\sum_{l_1=-1}^{1}\sum_{l_2=-1}^{1}\partial_{\{(k_1,k_2),(k_1+l_1,k_2+l_2)\}}\right) \tag{7}$$

with $(l_1, l_2) \neq (0,0)$, will be large. Thus, σ can be made adaptive to the image content by letting it be inversely proportional to ∂_{av} such that at non-homogeneous region, the influence of the neighboring pixels is small.

3. LIP SEGMENTATION

3.1 Color Feature Vector

Since intensity contrast between lip and non-lip region is generally very poor, our segmentation is based solely on color information. The lip image is originally in the RGB format. One drawback of RGB color space is that the color features are heavily correlated and the metrics does not represent color differences on a uniform scale, making it impossible to evaluate the similarity of two colors from their distance in the RGB color space. Instead, we transformed the RGB image into the CIELAB color space and the CIELUV color space. These color space are chosen because they have approximately uniform chromaticity diagrams, i.e., any two colors having a perceptual color difference of the same magnitude will be represented by lines of equal length on the chromaticity diagram. In both color space, the luminance information (L^* component) is separated from the chrominance information (a^*, b^* or u^*, v^* component). Intensity variation due to shadow and lighting will have minimal effect on the chrominance components. An RGB color image is transformed into the CIELAB color space and the CIELUV color space as follows[4]. First, we convert the RGB into the XYZ tristimulus values

$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} = \begin{pmatrix} 0.490 & 0.310 & 0.200 \\ 0.177 & 0.813 & 0.011 \\ 0.000 & 0.010 & 0.990 \end{pmatrix}\begin{pmatrix} R \\ G \\ B \end{pmatrix} \tag{8}$$

Let $X' = X/X_n$, $Y' = Y/Y_n$ and $Z' = Z/Z_n$ where X_n, Y_n and Z_n are the X,Y,Z values for the reference white. We take the reference white in RGB color space as $R_n = G_n = B_n = 1$. The L^*, a^*, b^* of CIELAB color space is then given by

$$L^* = \begin{cases} 116(Y')^{1/3} - 16 & if\ Y' > 0.008856 \\ 903.3Y' & otherwise \end{cases}$$
$$a^* = 500\left[K_1^{1/3} - K_2^{1/3}\right] \tag{9}$$
$$b^* = 200\left[K_2^{1/3} - K_3^{1/3}\right]$$

where

$$K_i = \begin{cases} \Phi_i & if\ \Phi_i > 0.008856 \\ 7.787\Phi_i + 16/116 & otherwise \end{cases}$$

for $i = 1,2,3$ and $\Phi_1 = X'$, $\Phi_2 = Y'$, $\Phi_3 = Z'$.
The L^* of CIELUV color space is the same as that of CIELAB. The u^*, v^* of CIELUV color space is given by

$$u^* = 13L^*(u' - u'_n)$$
$$v^* = 13L^*(v' - v'_n) \tag{10}$$

where u'_n and v'_n are the u' and v' values for reference white and

$$u' = 4X/(X+15Y+3Z)$$
$$v' = 9Y/(X+15Y+3Z) \quad (11)$$

By examining the histograms of various color components of the lip image, we decided to use the set of color features {a^*, b^*, u^*, v^*, hue_{ab}, hue_{uv}, $chroma_{uv}$} where $hue_{ab} = arctan(b^*/a^*)$, $hue_{uv} = arctan(v^*/u^*)$ and $chroma_{uv} = \sqrt{(a^{*2}+b^{*2})}$, to form the feature vector for each pixel. These color features are chosen because their histograms show discernable separation between lip and non-lip region. The luminance component L^* is not used as a feature since we do not want the clustering result to be affected by intensity variation caused by lighting and shadow.

3.2 Unwanted Region Masking

The present of outliers in the image data will affect the clustering process and the final segmentation. By masking out unreliable pixels or unwanted objects in the image, the clustering and segmentation results can be significantly improved. In the clustering process, we preset the number of cluster to be two. As teeth might be present in some lip images, we decided to mask out possible teeth pixels so that they do not enter into the clustering process. The chrominance value for teeth region is fairly consistent across speakers. A hue threshold for teeth region can be estimated from the sample data. This threshold is then used to mask out possible teeth region. We also find that the chrominance information for pixels with low luminance value are less reliable, i.e., the color difference between lip and non-lip region tends to decrease and fluctuates at low luminance level. Including such pixels into the clustering process will bias the cluster centroids and spread out the clusters. Low luminance region is therefore masked out before clustering.

3.3 Segmentation Experiments

We run the clustering and segmentation algorithm over a set of color lip images collected from different speakers, uttering different words and under slightly different lighting condition. The initial cluster centroids are initialized using the average lip and non-lip feature vectors obtained from a set of training sample data. Although the clustering algorithm is not sensitive to the initial cluster centroid value, such initialization enables convergence in less number of iterations. Fig. 2 shows the luminance component of a lip image. Clearly, the intensity contrast does not allow an edge-based segmentation algorithm to perform satisfactorily.

The output of the clustering algorithm consists of the cluster centroids and a membership image for each cluster. Fig. 3 shows the membership image of the lip cluster. The higher the membership value, the closer the pixel belongs to the lip class. From the membership image, we can see that the lip region stands out clearly from the non-lip region. The histogam of the membership image has two distinct peaks. A threshold can be chosen by detecting the location of the minimum point m in the valley between the two peaks. However, the exact position of m could be hard to determine if the valley is broad. To determine the appropriate threshold, we fit a two gaussian mixture curve onto the histogram and estimate the value as follows[5].

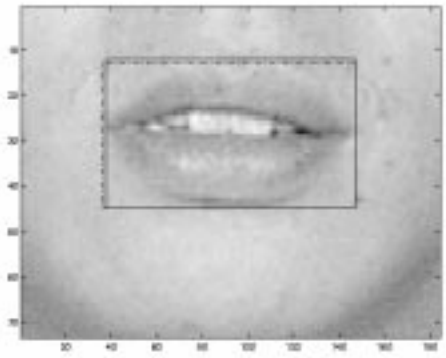

Fig.2. The luminance component of a color lip image.

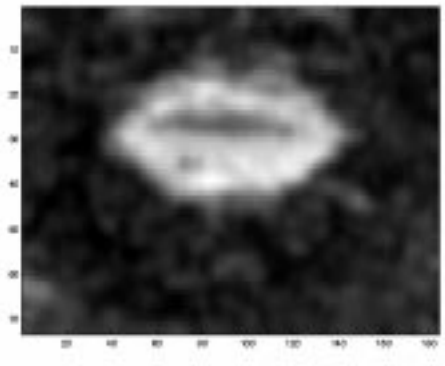

Fig. 3. The membership image of the lip cluster.

Let $p(z)$ be the mixture pdf function of two gaussian densities,

$$p(z) = \rho \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left(-\frac{(z-m_1)^2}{2\sigma_1^2}\right) + (1-\rho)\frac{1}{\sqrt{2\pi}\sigma_2}\exp\left(-\frac{(z-m_2)^2}{2\sigma_2^2}\right) \quad (12)$$

Let $h(z)$ be the normalized histogram of the lip membership image. Using steepest descent, the optimum values of ρ, m_1, m_2, σ_1, σ_2, can be obtained by minimizing

$$E(\rho, m_1, m_2, \sigma_1, \sigma_2) = \frac{1}{2}\sum_{z=0}^{L-1}\left|p(z)-h(z)\right|^2 \quad (13)$$

where L is the number of bins in the histogram function. The optimal threshold T can then be determined by solving for

$$AT^2 + BT + C = 0 \quad (14)$$

where

$$A = \sigma_1^2 - \sigma_2^2$$

$$B = 2(m_1\sigma_2^2 - m_2\sigma_1^2)$$

$$C = \sigma_1^2 m_2^2 - \sigma_2^2 m_1^2 + 2\sigma_1^2\sigma_2^2 \ln\left(\frac{\sigma_2\rho}{\sigma_1(1-\rho)}\right)$$

Fig. 4 shows the segmentation result obtained after thresholding the lip membership image. It can be seen that the segmentation is of good quality, the general shape of the lip is clearly visible.

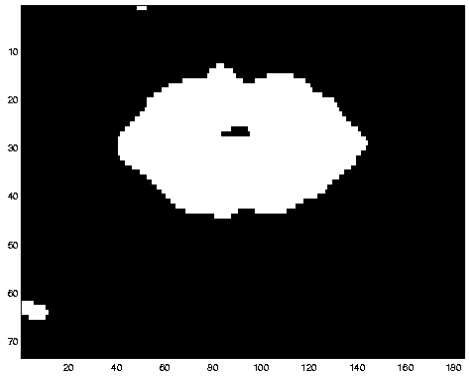

Fig.4. The segmented lip image.

To assess the quality of the segmentation, we compute the percentage of overlap of the bounding rectangle obtained from the segmented lip image and the bounding rectangle of the actual lip. The percentage overlap is given by

$$p = \frac{2(A_1 \cap A_2)}{A_1 + A_2} \times 100\% \quad (15)$$

where A_1, A_2 are the bounding rectangles of the segmented lip and the actual lip respectively. Using this measure, perfect segmentation will have an overlap of 100%. Fig.5 plots the number of lip images at any particular p for a total of 186 lip images. As can be seen, most lip images can be segmented satisfactorily. An example of the bounding rectangles of the segmented lip (solid box) and the actual lip (dashed box) overlaid on the lip image is shown in Fig. 2. The percentage overlap of the bounding rectangles for this image is 98%.

From the bounding rectangle, the height and width of the mouth can be measured. These features can be used directly for lipreading application. The membership image returned by the clustering algorithm contains much more information than just for segmentation, i.e., it can be used for instance, to guide a deformable model for lip contour extraction.

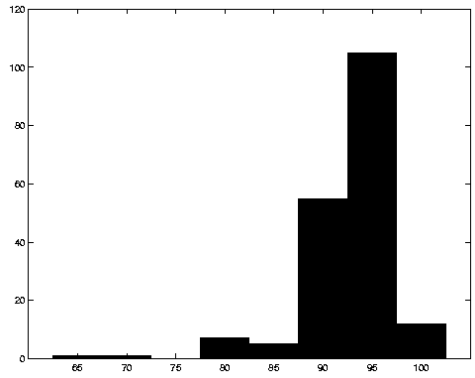

Fig.5. Number of lip images vs percentage overlap.

4. CONCLUSIONS

We present a lip segmentation algorithm based on fuzzy cluster analysis. The clustering algorithm is able to take into account the local spatial smoothness inherent in image data. This is due to a new distance metric that takes into account the influence of the neighboring pixels on the centre pixel in a 3 by 3 window. A color lip image is first transformed into the uniform CIELAB and CIELUV color space. A color feature vector for each pixel is derived from the chrominance components in the color spaces. Clustering is then carried out on the set of feature vectors. To minimize the effect of outliers on the clustering result, we masked out pixels having low luminance value and pixels that correspond to the teeth region. Finally, the segmentation is done by appropriately thresholding the lip membership image returned by the clustering algorithm. Our experiment indicates that most lips can be segmented satisfactorily.

5. REFERENCES

[1] E.D. Petajan, "Automatic Lipreading to Enhance Speech Recognition", Proc. Of IEEE Global Telecommunications Conference, Atlanta, Georgia, 1984, pp.265-272.

[2] Tarcisio Coianiz, Lorenzo Torresani, and Bruno Caprile, "2D Deformable Models for Visual Speech Analysis", in Speechreading by Humans and Machines, D.G.Stork & M.E.Hennecke Eds., Springer, NY, 1996.

[3] J.C. Bezdek, "Pattern Recognition with Fuzzy Objective Function Algorithms", Plenum Press, New York, 1981.

[4] R.W.G. Hunt, "Measuring Colour" 2nd Edition, Ellis Horwood Series in Applied Science and Industrial Technology, Ellis Horwood Limited, 1991.

[5] Z.Chi, H.Yan and T.Pham, "Fuzzy Algorithms:With Applications to Image Processing and Pattern Recognition", World Scientific Publishing Co. Pte. Ltd,1996.