Lip Segmentation Using Level Set Method: Fusing Landmark Edge Distance and Image Information

Seyed Reza Banimahd
Electrical Engineering Department
Sahand University of Technology
Sahand New Town, Tabriz, Iran
r_banimahd@sut.ac.ir

Hossein Ebrahimnezhad
Electrical Engineering Department
Sahand University of Technology
Sahand New Town, Tabriz, Iran
ebrahimnezhad@sut.ac.ir

Abstract—Lip segmentation is an essential step in audio-visual processing systems. In this paper, we incorporate the color and edge information in level set formulation, for extraction of lip contour. We build two initiative auxiliary images by mixing of different color spaces, to extract the landmark edges for upper and lower part of lip. The performance of this approach on VidTIMIT databases is tested and accuracy of 91.2% is reached.

Keywords—Lip contour extraction; level sets.

I. INTRODUCTION

Audio-visual speech processing is an important field of computer vision science. Fusion of audio and visual information has many applications in human-computer interaction such as audio-visual speech recognition [1]. It is shown in [1], that the use of visual cues, significantly, outperforms the recognition performance in acoustic noisy environment. In fact, in such application the visual cues of lip movement contain important information. Thus, lip segmentation in an image is a critical and challenging task in audio-visual speech processing system. The higher accuracy of the segmentation results in an increase in the performance of the global system.

Level sets provide an energy minimization framework that extensively used in modeling and boundary detection in computer vision field. Level set methods consider general concepts like boundary smoothness and can incorporate additional information defined as energy function into curve evolution. In [2] and [3], lip contour extraction was performed based on level set curve evolution. A non-linear learning method in the form of an SVM classifier is trained to recognize lip color over a variety of faces [2]. There is an overlap of color distribution between lip and skin pixels. The classifier error raises from this overlap forms the segmentation error. In [3] a novel lip contour extraction method based on level set curve evolution using shape constraint has been proposed. In this method, the curve is evolved by minimizing an energy function that incorporates shape constraint function using parametric lip contour model, as an internal energy. The shape constraint function prevents the curve to converge to arbitrary shapes, occurred due to weak color contrast between lip and skin regions. The main drawback of this approach is that the lip contour model is cubic polynomials and hence a constraint for degree of freedom of extracted contour is imposed.

We propose in this paper, a new level set formulation for lip contour extraction. Contour evolution in level set formulation results from minimization process of an energy function. The energy function that is used in this paper contains two terms. The first term is a modified version of Chan-Vese [4] formulation in which, for segmentation, local statics information is used [5]. The second term is related to the distance of evolving curve from a set of extracted edge points that this landmark edges guide the curve to lip boundary in critical and troublesome locations. In next subsections, we describe each energy function in more details.

In addition, we define two auxiliary images for extraction of landmark edges in up and bottom of lip. These images are obtained by initiative mixing of different color spaces. Some post-processing operations are performed on extracted edge of these images, to produce the landmark edges.

The remainder of the paper is organized as follows. In section 2, we will present the pre-processing procedure, which contains face detection and mouth localization. In section 3, the construction of auxiliary images is introduced. In section 4, the level set segmentation formulation and details of the proposed energy functions in segmentation, is presented. Experimental results and conclusion will be given in section 4 and 5, respectively.

II. LIP LOCALIZATION AND EXTRACTION IN FACE IMAGES

In audio-visual research domain, the problem of face and its component detection is somewhat connived. In this paper, a simple algorithm for face and facial feature extraction is used. At first, image is converted from RGB color space to HSV system. In this space, hue component for skin is different from background. Therefore, on resulting image, face is extracted using a threshold. In the second step, for determining mouth region, a method similar to one proposed in [6] is performed. In [6] an Elliptical Gabor Filter (EGF) is used in a coarse-to-fine process for detection of eye location.
The shape of Elliptical Gabor Filter is similar to an ellipse and is defined as follows [6]:

\[
G(x, y) = g(x', y') \exp \left(2\pi jF \sqrt{x^2 + y^2}\right)
\]  \hspace{1cm} (1)

where, \((x', y') = (x \cos \theta + y \sin \theta, -x \sin \theta + y \cos \theta)\) and \(F\) is the spatial central frequency of the filter in the frequency domain, \(G(x, y)\) is 2-D Gaussian envelope as follows:

\[
g(x, y) = \frac{1}{2\pi\sigma_x\sigma_y} \exp \left(-\frac{1}{2} \left(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2}\right)\right)
\]  \hspace{1cm} (2)

where, \(\sigma_x\) and \(\sigma_y\) are variances (scale factors) along \(x\) and \(y\) axes respectively, \((x, y)\) is the center of the receptive field in the spatial domain and \(\theta\) is the rotation angle of the filter. Because of the elliptical shape of lips, in appropriate frequency and scale factors, the EGF emphasizes the edge of lips. Therefore, these parameters are adjusted in a way that the lip and mouth region are apparently extracted. After performing EGF filtering on face image, binarization of image carried out. Then on black and white image, post-processing morphological operations is used to smooth the extracted binary mask and eliminate small erroneous blobs and holes. By determining mouth region, a rectangular region of interest (ROI) is selected. The pre-processing procedure is shown in Fig.1.

III. AUXILIARY IMAGES

In this paper, for reduction of shadow effect and extraction of a proper model for lip, we use a mixture of different color spaces to construct a new color space. By edge detection algorithms in critical regions, a series of landmark edges points are extracted. In level set curve evolution, for lip segmentation, these points are embedded as an energy function, and direct the contour to correct lip boundary. We propose two empirical formulations to define the new color spaces for extraction of landmark edges in upper and lower lip.

To extract the upper lip edge, we use the HSV color space. In fact we use an empirical formulation to make an image that distinctly extracts the upper lip edge:

\[
I_{UE} = 1 - (H \times \exp(V/S))^*
\]  \hspace{1cm} (3)

where, \(H\), \(S\) and \(V\) are hue, saturation and value component of the HSV color space. The \((d)^*\) phrase means that the min-max normalization performs on data \(d\).

We propose a combination form of three color space (i.e. HSV, Lab, and RGB) to extract the boundary of lower part of lip. The proposed formulation is as follows:

\[
I_{LE} = h' \times \exp\left(-\left(1 - a' \times h_y \times h'\right) / \sigma\right)^*
\]  \hspace{1cm} (4)

in which, \(a'\) is 2nd component in Lab color space in normalized form. \(h_y\) is normalized pseudo hue defined by [7]. \(h'\) is a modified version of hue component that obtained by:

\[
h' = 1 - (h)^*
\]  \hspace{1cm} (5)

The use of edge detection algorithms as demonstrated in Fig.2 is not efficient. Some extra edges in mouth region are...
produced. This landmark edges are not necessarily determines the exact location of lip boundary. But, they provide an attraction force to guide the contour to the lip boundary in defined energy function of level set formulation.

Extraction of the landmark edges in auxiliary image is performed on edge images. Upper landmark edges are selected in up to down scan of upper edge auxiliary image. In the same way, lower landmark edges are selected in down to up scan of lower edge auxiliary image. In this way, the extra edges in the interior of mouth region are eliminated. In Fig.2, the extracted upper and lower landmark edges for up and down region of lip are shown.

IV. LIP CONTOUR EXTRACTION USING LEVEL SET

Level Set method, is a common way for description of an active contour in image segmentation problem. Here, a variation of the level set method is proposed and used for extraction of lip contour from ROI region. As an introductory explanation of level set theory [8], let \( \omega \) be a bounded open subset of \( \mathbb{R}^2 \), with \( \partial \omega \) its boundary, and \( I: \Omega \to \mathbb{R} \) denote a given image defined on the domain \( \Omega \subset \mathbb{R}^2 \). In level set framework, an evolving curve \( C \) (boundary of an open set \( \omega \) in \( \Omega \), i.e. \( C = \partial \omega \)) is represented implicitly as zero level of an embedding function \( \phi: \Omega \to \mathbb{R} \) (level set function that usually a signed distance function to evolving curve is used) [8]:

\[
C = \{ x \in \Omega \mid \phi(x) = 0 \} \tag{6}
\]

Thus, given the initial state of level set function \( \phi \), the evolution of curve \( C \) based on a speed function \( F \) can be written as (level set equation) [8]:

\[
\frac{\partial \phi}{\partial t} + F \nabla \phi = 0 \tag{7}
\]

where, \( \nabla \phi \) denotes the normalized gradients of the level set function. This equation also is called gradient flow that minimizes an energy function \( E \) such that \( F = \frac{\partial E}{\partial \phi} \). For image segmentation problems, the energy function \( E \) depends on the image data and the level set function \( \phi \).

Thus, the equation (7) can be reformulated as [8]:

\[
\frac{\partial \phi}{\partial t} = \nabla \phi E_{\text{int}} + \lambda \delta(\phi) \text{div} \left( \frac{\nabla \phi}{|\nabla \phi|} \right) \tag{8}
\]

The energy function \( E \) that has been used in our proposed method consists of various parts. In fact, various constraints in the form of some energy functions try to perform correct segmentation.

A. Image-Based Energy Term

The mouth region is heterogeneous. In the presence of the mustaches and beard in around of the mouth (especially in open mouth), appearance of teeth and tongue prevents the correct modeling of lip region using region averages. In [5], a localize version of Chan-Vese energy function [4] has been discussed that uses local mean of interior and exterior regions instead of global ones that can be used for extraction of lip contour. By eliminating the length term in Chan-Vese energy function, the corresponding localized energy function is formed by replacing global means by their local equivalents form as follows [5]:

\[
E_{\text{cv-l}} = \int_{\Omega} H(\phi(y))(I(y) - u_s)^2 + (1 - H(\phi(y)))(I(y) - v_s)^2 \, dy \tag{9}
\]

The flow equation for the evolution of \( \phi \) is derived by taking the first variation of this energy function with respect to \( \phi \) [5]:

\[
\frac{\partial \phi}{\partial t} = \delta(\phi(x)) \int_{\Omega} B(x,y) \delta(\phi(y)) \times \left( (I(y) - u_s)^2 - (I(y) - v_s)^2 \right) \, dy \tag{10}
\]

where, \( \delta \) is the derivate of \( H \), a smoothed version of the Dirac delta. In the localized version, the minimum is obtained when each point on the curve has moved such that the local interior and exterior around every point along the curve is best approximated by local means \( u_s \) and \( v_s \).

B. Point-Based Energy Term

To alleviate the previously mentioned problems, some landmark edges in upper and lower lip boundary are extracted using the presented method in previous section. In the evolution of level set curve, point-based energy function as an external force ensures the crossing of these key-points by the curve. In [9], Cohen and Cohen propose the use of an attraction potential to landmark edges extracted by a local edge detector.

In this paper, we propose a new energy function to incorporate the landmark edges into the level set framework. The energy function must have the property of attraction of the whole contour to the landmark edges. The nearest contour points to the edges are more influenced. An energy function like this one can be used to incorporate the edges attraction energy into the curve evolution of level set:

\[
E_p = \int_{\Omega} \delta(\phi) \text{exp} \left( -\frac{||s - p||}{\eta} \right) \, d\Omega \tag{11}
\]

where, \( ||s|| \) is the Euclidian distance between \( i \)th point of extracted edge \( p_i \) and a point in evolving contour \( x \). The term \( \delta(\phi) \), guarantees that the energy function is only computed on the boundary curve. \( \eta \) determines the influence weight of the landmark edges on adjacent curve. So, for each point on the curve, the energy of landmark edges can be considered as the contribution of all landmark edges, i.e.:

\[
E_p = \sum_{i} E_{p_i} = \sum_{i} \delta(\phi) \text{exp} \left( -\frac{||s - p_i||}{\eta} \right) \, dx \tag{12}
\]

Force of the landmark edges that influences the contour points is considered as the first variation of energy function in normal direction and can be written as:

\[
F = \sum_{i} -\frac{1}{\eta ||s - p_i||} \text{exp} \left( -\frac{||s - p_i||}{\eta} \right) \frac{x}{||x||} \tag{13}
\]
So, the flow equation of level set function will be:
\[
\frac{\partial \phi}{\partial t}(x) = \delta(\phi(x)) \times \int_{\Omega} \frac{1}{\eta} \exp \left( -\frac{\|x - \mu\|}{\eta} \right) \times \phi \, dy
\]
(14)

V. EXPERIMENTAL RESULTS

In order to evaluate the performance of the proposed method, we performed experiments on an audio-visual database. Unfortunately, there is not a benchmark dataset in audio-visual speech processing domain. VidTIMIT database [10], consists of audio recordings and video sequences of 43 subjects (19 female and 24 male), with the resolution of 384×512 pixels, reciting short sentences in three sessions with an average delay of a week between sessions, allowing for appearance and mood changes. Since, our algorithm employs edge information as a part of energy function, we have selected 22 number of whitey-skin people images, with no mustaches and bread. In fact, the presence of such image information can make outlier edges. As, no ground truth is available for this database, the lip boundaries must be drawn manually. This can be a time consuming and prosaic task. Therefore, we chose only 70 frame of each person sequence with different shape of mouth.

To illustrate the performance of the proposed energy function, a quantitative measure must be used. We use the precision and recall criteria by the following formulation:
\[
p = \frac{t_p}{t_p + f_p}, \quad r = \frac{t_p}{t_p + f_n}
\]
(15)
where, \( t_p \), \( f_p \) and \( f_n \) denote the true positives, the false positives and the false negatives with respect to the ground truth binary image of lip, respectively. The closer values of \( p \) and \( r \) to 1 define the better segmentation. In Table 1, the mean of recall and precision metric for different methods is shown. In Fig.3 the segmentation results for some selected images are demonstrated. In first set, segmentation is done by Chan-Vese energy function. Second set shows the localized version of this energy function segmentation result and in third set, segmentation is performed by incorporating the extracted landmark edges.

![Segmentation results](image1)

VI. CONCLUSION

In this paper, we proposed the modified level set formulation for lip contour extraction task. New energy function that provides absorption force for contour to some landmark edge extracted from the auxiliary image was introduced. Imposing extra constraint can increase the segmentation accuracy.

VII. REFERENCES


![Segmentation results](image2)

![Segmentation results](image3)

<table>
<thead>
<tr>
<th>Energy Fun.</th>
<th>CV</th>
<th>L-CV</th>
<th>L-CV&amp;LE(proposed)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>0.832</td>
<td>0.866</td>
<td>0.912</td>
</tr>
<tr>
<td>Recall</td>
<td>0.933</td>
<td>0.871</td>
<td>0.864</td>
</tr>
</tbody>
</table>

TABLE I. SEGMENTATION ERROR BY CHAN-VES (CV), LOCALIZED CHAN-VES (L-CV) AND INCORPORATING LANDMARK EDGES (LE) IN LOCALIZED CHAN-VES.