Real-time Outer Lip Contour Tracking for HCI Applications

Sabri Gurbuz

ATR Human Information Science Laboratories, Kyoto, Japan
sabrig@atr.jp

Abstract

A new method for tracking outer lip contours of individuals in real world conditions is presented. For an arbitrary speaker, lip color properties are learned for the Bayes decision from the current image frame using the nose tip location as a reference point. Estimated outer contour data is fit to an ellipsoid for further eliminating the effect of outliers in the contour. The algorithm, which is posed as a real-time solution to lip contour tracking in real world conditions, is made efficient by the use of an online learning method. Demonstrations of lip contour tracking and its application to a mouth movement imitation are presented.

1. Introduction

Robust real-time lip tracking is one of the important research topics in multimodal Human-Computer Interface applications. For example, addition of visual speech information to speech recognizer clearly meets at least two practicable criteria: It mimics human visual perception of speech recognition, and it may contain information that is not always present in the acoustic domain [1]. Another application example is enhancing the social interaction between humans and humanoid agents with their human-like mouth movements during speech [2]. The motivation of this research is to develop an algorithm to track the mouth and its outer contour in real world conditions without using prior training data.

This work is organized as follows. In section 2, related earlier works are described. Section 3 discusses face and mouth region of interest (ROI) localization, and lip contour tracking in real-time. Experimental results and discussions are presented in Section 5. Finally, Conclusion is given in Section 6.

2. Related Work

Several researchers have utilized the frame-work of active contour models (snakes) [3] for outer lip contour tracking. Active contour models were first introduced in 1987 by Kass et. al. The contour model represents an object boundary or some other image features as a parametric curve. A user or a higher level process initializes the curve close to the object boundary. Iteratively at the end the curve is expected to shrink-wrap around the lip boundaries. In general, the contour constraints are defined by internal (elasticity and bending) and external (boundary) energies. Due to the complex nature of the human mouth surroundings such as wrinkles and mustache, defining a good external energy depends on the parameter selection. Therefore, in the case of cluttered lip ROI, the snakes algorithm may fail to converge the correct boundary.

Xu and Prince [4] defined a new external energy for snakes called gradient vector flow (GVF) for object boundary tracking. The GVF field is defined to be a vector field to improve the attraction of the snakes to object boundaries. Though in general there is a significant improvement over the traditional snakes algorithm, the GVF snake algorithm is also sensitive to parameter selection and computationally more expensive.

Chan et. al. [5] employed an iterative B-spline approach using a reference shape model for lip contour tracking. Their algorithm as most others requires prespecified knowledge about the user to track the user’s mouth.

Although most approaches reported success, their computational complexity and requirements of prior knowledge about the user’s skin color or mouth shape creates disadvantage for many practical usages. Our work extends these efforts to a ubiquitous lip tracking system by adding an online learning algorithm which doesn’t require prior knowledge about the user.

3. Face and Mouth ROI Localization

In general, face tracking approaches are either image based or direct feature search based methods. The image based (top-down) approaches utilize statistical models of skin color pixels to find the face region first, accordingly preprocessed face templates or feature search algorithms are used to match the candidate face regions [6]. Feature based approaches use specialized filters directly such as templates or Gabor filter of different frequencies and orientations to locate the facial features. Our work falls into the latter category. That is, first we track the eyes employing the method described in [7]. Then, we utilize the eye locations for face ROI selection (typically width is 1.2 and height is 1.6 times the distance between the eyes) to trace the nose bridge and nose tip. The nose tip is then utilized as a reference for the selection of mouth ROI.

3.1. Eye Tracking

The patterns of between the eyes are detected and tracked with updated pattern matching. To cope with scales of faces, various scale down images are considered for the detection, and an appropriate scale is selected according to the distance between the eyes. The algorithm calculates the intermediate representation of the input image called “Integral image” described in [8]. Then, a six-segmented rectangular (SSR) filter is used for fast filtering of bright-dark relations of the eye region in the image. The details of the eye tracking algorithm can be found in [7]. Resulting face candidates around the eyes are further verified by a support vector machine (SVM) algorithm. The eye locations are then utilized for the face ROI selection for locating the nose bridge and nose tip.

3.2. Nose Bridge and Nose Tip Tracking

The human nose has a convex curvature shape and the ridge of the nose from the eye level to the tip of the nose lies on a line as depicted in Figure 1. Our system utilizes the information in the integral intensity profile of convex curvature shape. The peak of the profile of a segment that satisfies Equation 1 is the convex hull point. A convolution filter with three segments traces the ridge with the center segment greater or equal to the side seg-
ments, and the sum of the intensities in all three segments gives a maximum value on the convex hull point. Figure 2 shows the proposed filter with three segments that traces the maximum intensity (convex hull) pattern starting from the eye line. The criteria for finding the convex hull point on a integral intensity profile of a row segment is as follows,

\[ S_1 \leq S_2 \geq S_3 \quad \text{and} \quad \arg\left\{ \max_j (S_1 + 3S_2 + S_3) \right\}, \]

where \( S_j \) denotes the integral value of the intensities of a segment in the maximum filter shown in Figure 2, and \( j \) is the center location of the filter in the current integral intensity profile. The filter is convolved with the integral intensity profile of every row segment. A row segment typically extends over 5 to 10 rows of the face ROI image, and a face ROI image typically contains 20 row segments. Integral intensity profiles of row segments are processed to find their hull points (see Figure 1) using Equation 1 until either the end of the face ROI is reached or until Equation 1 is no longer satisfied. That is, some of the nose bridge candidate points may not be on the nose bridge and may be below the nose tip. Thus, for the refinement process, we found that the first derivative of the horizontal integral intensity profile (see Figure 3 for an example of a horizontal integral intensity profile) at the nose tip is a maximum, and the second derivative is zero at the nostril level [9].

4. Lip Tracking

The nose tip location is utilized for the initial mouth ROI selection. Then, the maximum-likelihood estimation of class conditional densities for subsets of lip (\( w_1 \)) and non-lip (\( w_2 \)) classes are formed in real-time for Bayes decision rule. That is, multivariate class conditional Gaussian density parameters are estimated for every frame using an unsupervised maximum-likelihood estimation method.

4.1. Online Learning and Extraction of Lip and Non-lip Data Samples

In order to alleviate the influence of ambient lighting on the sample class data, chromatic color transformation is adopted for color representation [6, 10]. Yang et al. in [10] pointed out that human skin colors are less variant in the chromatic color space than the RGB color space. Although in general the skin-color distribution of each individual may be modeled by a multivariate normal distribution, the parameters of the distribution for different people and different lighting conditions are different. Therefore, Online learning and sample data extraction are important keys for handling different skin-tone colors and lighting changes. To solve these two issues, In [10], authors proposed an adaptation approach to transform the previous developed color model into the new environment by combination of the known parameters from the previous frames. This approach has two drawbacks in general. First, it requires an initial model to start, and second, it may fail in the case of a different user with completely different skin-tone color starts using the system.

We propose an online learning approach to extract sample data for lip and non-lip classes to estimate their distribution in real time. Chiang et al. [6] in their work provides hints for this approach. They pointed out that lip colors are distributed at the lower range of green channel in the (r,g) plane. Figure 4 shows an example distribution of lip and non-lip colors in the normalized (r,g) space.

Utilizing the nose tip, time dependent (r,g) spaces for lip and non-lip are estimated for every frame by allowing the points stay within the lip (r,g) space as shown (typical 10%) of the points stay within the lip (r,g) space as shown in Figure 4. Then, using the obtained (r,g) space information in the initial classification, the pixels below the nostril line that falls within the lip space are considered as lip pixels, and the other pixels are considered as non-lip pixels in the sample data set extraction process, and RGB color values of pixels are stored as class attributes, respectively.

In most cases, sample data contains high variance and it is preferable to separate into subsets according to its time dependent intensity average. Let \( \text{avg}_L \) and \( \text{D}_k \) be the intensity average and \( k^{th} \) subset of the lip class, respectively. The subsets of the lip class are separated according to lip class’ intensity average as

\[
\begin{align*}
\text{assign to } D_1 & \quad \text{if } x_{\text{intensity}} < \text{avg}_L, \\
\text{assign to } D_2 & \quad \text{if } \text{avg}_L / 2 < x_{\text{intensity}} < 3 \text{avg}_L / 2, \\
\text{assign to } D_3 & \quad \text{if } x_{\text{intensity}} > 3 \text{avg}_L.
\end{align*}
\]

Using the same concept in Equation 2, we also separate the
where the mean value of the probabilities may be written as follows:

\[ p(x) = \frac{p(x|w_1)p(w_1)}{p(w_1)} \tag{6} \]

where \( p(x) \) is the density function and is positive constant for all classes. Then, re-arranging both sides, we obtain

\[ L(x) = \frac{p(x|w_1)}{p(x|w_2)} \tag{7} \]

where \( L(x) \) is called the likelihood ratio, and \( p(w_2)/p(w_1) \) is called the threshold value of the likelihood ratio for the decision. Because of the exponential form of the densities involved in Equation 7, it is preferable to work with the monotonic discriminant functions obtained by taking the logarithm as follows.

\[ q_{w_i}(x) = \ln(p(x|w_i)p(w_i)) \tag{8} \]

thus,

\[ q_{w_i}(x) = -\frac{1}{2}(x - \mu_i)(x - \mu_i)^T + c_i \tag{9} \]

where \( c_i = \ln(p(w_i)) - (1/2)\ln(2\pi - (1/2)||\Sigma_i||) \) is a constant for this image frame. In general, Equation 9 has only nonlinear quadratic form and a summation, and using this equation, the Bayes rule can be implemented for real-time lip tracking as follows.

\[ q_{w_1}(x) \geq q_{w_2}(x) \tag{10} \]

where \( q(i)(x) = \max\{q_{w_1}(x), q_{w_2}(x), q_{w_3}(x)\} \) for \( i = w_1, w_2 \) and referring to Figure 5. Threshold value of the likelihood ratio as shown in Equation 7 is based on a priori class probabilities. In our implementation, equally likely class probabilities are assumed.

4.3. Bayes Decision Rule

Let \( x \) be an observation vector formed from RGB attributes of a pixel location in an image frame. Our goal is to design a Bayes classifier to determine whether \( x \) belongs to \( w_1 \) or \( w_2 \) in two class classification problem. The Bayes test using a posteriori probabilities may be written as follows:

\[ p(w_1|x) \geq p(w_2|x) \tag{5} \]

where \( p(w_1|x) \) is the a posteriori probability of \( w_i \) given \( x \). Equation 5 shows that if the probability of \( w_1 \) given \( x \) is larger than the probability of \( w_2 \), then \( x \) is declared belonging to \( w_1 \), and vice versa. Since direct calculation of \( p(w_i|x) \) is not practical, we can re-write the a posteriori probability of \( w_i \) using Bayes’ Theorem in terms of a priori probability and the conditional density function \( p(x|w_i) \), as

\[ p(w_i|x) = \frac{p(x|w_i)p(w_i)}{p(x)} \tag{6} \]

where \( p(x) \) is the density function and is positive constant for all classes. Then, re-arranging both sides, we obtain

\[ L(x) = \frac{p(x|w_1)}{p(x|w_2)} \tag{7} \]

where \( L(x) \) is called the likelihood ratio, and \( p(w_2)/p(w_1) \) is called the threshold value of the likelihood ratio for the decision. Because of the exponential form of the densities involved in Equation 7, it is preferable to work with the monotonic discriminant functions obtained by taking the logarithm as follows.

\[ q_{w_i}(x) = \ln(p(x|w_i)p(w_i)) \tag{8} \]

thus,

\[ q_{w_i}(x) = -\frac{1}{2}(x - \mu_i)(x - \mu_i)^T + c_i \tag{9} \]

where \( c_i = \ln(p(w_i)) - (1/2)\ln(2\pi - (1/2)||\Sigma_i||) \) is a constant for this image frame. In general, Equation 9 has only nonlinear quadratic form and a summation, and using this equation, the Bayes rule can be implemented for real-time lip tracking as follows.

\[ q_{w_1}(x) \geq q_{w_2}(x) \tag{10} \]

where \( q(i)(x) = \max\{q_{w_1}(x), q_{w_2}(x), q_{w_3}(x)\} \) for \( i = w_1, w_2 \) and referring to Figure 5. Threshold value of the likelihood ratio as shown in Equation 7 is based on a priori class probabilities. In our implementation, equally likely a priori class probabilities are assumed.

4.4. Mouth Shape Parameterization

After mouth tracking algorithm locates the mouth region, outer lip contours of the speaker’s lips are detected (see the center image in Figure 3). Then, the outer contour as a whole...
is parameterized by a generalized ellipse shape which is obtained using the estimated contour data. A parametric contour is found that corresponds to the general quadratic equation $a_1 x^2 + a_2 xy + a_3 y^2 + a_4 x + a_5 y + a_6 = 0$, where $a_i$s are constants, and $a_1$ and $a_3$ are non-zero. Let us denote the 2D positions over the traced outer lip contour as

\[
\begin{bmatrix}
  x_1 & x_2 & x_3 & \ldots & x_N \\
  y_1 & y_2 & y_3 & \ldots & y_N
\end{bmatrix}.
\]

The basic form used in the elliptical parameter estimation in matrix notation is $M \theta = 0$ where $\theta = (a_1 \ a_2 \ a_3 \ a_4 \ a_5 \ a_6)^T$. The dimensionality of $M$ is the number of points, $N$, in the segment multiplied by 6 ($N \times 6$). Each row of $M$ corresponds to one point in the segment. The parameters of each contour are then solved using the least-squares method to find $a_i$s, where $i = 1, 2, \ldots, 6$.

Using the estimated parameters, parametric lip contour data can be re-generated for each image frame. Five points are sufficient to represent a general elliptical shape, leading to a significant data reduction and representation of lip shape for a lip structure design and for a lip-reading system.

5. Experimental Results and Discussion

In this paper, our work specifically focused on a novel unsupervised online parameter learning algorithm for real-time lip contour tracking approach. RGB color attributes of lip pixels are used for the lip tracking, but only the intensity information is used for the eye and nose bridge tracking algorithms which makes the tracking robust.

The proposed lip contour algorithm has been tested on live data from various users without using any special markers or paintings. Figure 6 shows tracked lip contour results for various users under various lighting conditions. The facial feature tracking algorithm which uses Videre CCD camera works at 30 frames per second on a 2 GHz notebook PC with Windows platform. Because of the high volume of data to be examined, evaluation of the system is quite difficult, but we have included real-time tracking videos, and mouth movement imitation demonstrations as an application example under “http://www.his.atr.jp/~sabrig/liptracking.htm”.

6. Conclusions

A new method for tracking outer lip contours of individuals in real world conditions is described. Specifically, a novel unsupervised online parameter estimation algorithm for the Bayesian rule is proposed. That is, a speaker’s lip colors are learned from the current image frame using the nose tip as a reference point. Vertical and horizontal integral projections are utilized to guide the algorithm in picking out the correct lip contour. In the final stage, estimated outer contour data for every image frame is parameterized as a generalized ellipse.

7. Acknowledgements

This research was conducted as part of ‘Research on Human Communication’ with funding from the National Institute of Information and Communications Technology. Thanks are due to Shinjiro Kawato for the original eye tracking extended in this paper, and Marcia Riley for the graphical talking head system.

8. References