Lip Segmentation Using Adaptive Color Space Training

Erol Ozgur, Berkay Yilmaz, Harun Karabalkan, Hakan Erdogan, Mustafa Unel

Faculty of Engineering and Natural Sciences,
Sabanci University, Istanbul, Turkey
{erol,berkayyilmaz,karabalkan}@su.sabanciuniv.edu,{haerdogan,munel}@sabanciuniv.edu

Abstract

In audio-visual speech recognition (AVSR), it is beneficial to use lip boundary information in addition to texture-dependent features. In this paper, we propose an automatic lip segmentation method that can be used in AVSR systems. The algorithm consists of the following steps: face detection, lip corners extraction, adaptive color space training for lip and non-lip regions using Gaussian mixture models (GMMs), and curve evolution using level-set formulation based on region and image gradients fields. Region-based fields are obtained using adapted GMM likelihoods. We have tested the proposed algorithm on a database (SU-TAV) of 100 facial images and obtained objective performance results by comparing automatic lip segmentations with hand-marked ground truth segmentations. Experimental results are promising and much work has to be done to improve the robustness of the proposed method.

Index Terms: Lip segmentation, color spaces, GMM, level-sets.

1. Introduction

Lip boundary extraction is an important problem that has been studied to some extent in the literature [1, 2, 3, 4]. Lip segmentation can be an important part of audio-visual speech recognition, lip-synching, modeling of talking avatars and facial feature tracking systems.

In audio-visual speech recognition, it has been shown that using lip texture information is more valuable than using the lip boundary information [5, 6]. However, this result may have been partly due to inaccurate boundary extraction as well, since lip segmentation performance was not independently evaluated in earlier studies. In addition, it is possible to use lip segmentation information complementary to the texture information. Lip boundary features can be utilized in addition to lip texture features in a multi-stream Hidden Markov model framework with an appropriate weighting scheme. Thus, we conjecture it is beneficial to use lip boundary information to improve accuracy in AVSR. Once the boundary of a lip is found, one may extract geometric or algebraic features from it. These features can be used in audio-visual speech recognition systems as complementary features to audio and other visual features.

In this paper, we use statistical color distributions represented by Gaussian mixture models (GMMs) to obtain region-based fields to be employed in a level-set curve-evolution framework. The GMMs are first trained in a speaker-independent way by taking lip and non-lip examples from multiple subjects. Then, for each test subject, we abtain initial lip and non-lip examples using good initial guesses and adapt lip and non-lip GMMs to the subject by using the maximum aposteriori probability (MAP) algorithm.

We compare the performance of different color spaces for lip segmentation. Our level-set formulation enables fusion of different regional fields. We assess different combinations of color spaces using our approach as well. For performance comparison, we employ hand-marked lip boundaries as ground truth. We report precision and recall rates for each method we tried.

The rest of the paper is organized as follows, Section 2 explains the main steps of the lip segmentation algorithm we used. The conducted experiments are presented in Section 3. Finally in Section 4 we conclude the paper and propose some future improvements.

2. Lip Segmentation

For lip segmentation, we employ a series of algorithms for face detection, lip corner detection and exact boundary extraction. We detail those steps in the following discussion.

2.1. Face Detection

Fast and accurate face detection based on Viola and Jones's method is performed on frontal face images to extract the face [7]. The algorithm is based on efficiently extracting Haar-like features and using those features in an Adaboost classification/feature selection framework. After the face is detected, it is resized to fixed dimensions to enable further invariant processing.

2.2. Lip Corners Extraction

2.2.1. PCA Training

After face detection, we need to extract lip corners to initialize the level set algorithm. Lip corners are extracted from the face image using PCA template matching [8]. First we train a PCA model for patches of fixed dimension centered around left and right lip corners, separately, using a large training database of face images. We experimented with different patch window sizes to get the best result. We downsized (by 2) extracted face image for faster processing. We obtained a PCA lip-corner-patch space of reduced dimension for each corner. Mean images and the primary PCA eigen-patch images are shown in Figure 1. Finally, a Gaussian mixture model (GMM) is built using the PCA coefficients and the coordinates of each lip corner. We expect that this model yields high likelihood for feature vectors extracted from correct lip corners.

We obtained the best result for corner detection with a 41×61 window. Mean error of detection as measured in Euclidean distance between the predicted lip corner and the ground truth is 4.4 and 5.2 pixels for left and right lip corners respectively.

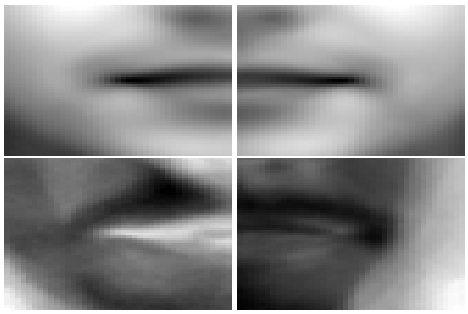
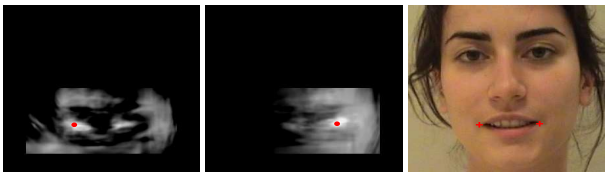


Figure 1: Mean lip-corner-patch (top row) and primary PCA eigen-patch (bottom row) images for left and right lip corners.

2.2.2. *PCA Template Matching*

Once we train a PCA lip corner model, we find the lip corners in a given image using PCA template matching. After resizing the test image, we look in a patch around each pixel and reduce dimension by PCA. We get a feature vector which consists of the PCA coefficients in that neighborhood concatenated by x and y coordinates of the pixel. GMM likelihood (or score) of each pixel is calculated using the model trained before, and the pixel with the greatest GMM score is decided to be the lip corner. Our testing data is distinct from the training data. In addition, the subjects in training and testing images are distinct as well. Sample GMM score images and the found corners are shown in Figure 2.



(a) (b) (c)

Figure 2: GMM score images for (a) left lip-corner, (b) right lip-corner. (c) Found lip corners shown superimposed on the face image.

2.3. Probabilistic Modeling

We would like to learn color distributions of the pixels in lip and non-lip regions by modeling them using Gaussian mixture models. The training data we used has the lip region hand-marked. We take a region of interest around the marked lip region and label the pixels as lip or non-lip within that region of interest. We randomly select a fixed number of pixels from each region in our training data. Next, we extract color-space features from each chosen pixel. We use these features to train a GMM for each region.

Let $\boldsymbol{x}$ denote x and y coordinates and $\boldsymbol{c} = \boldsymbol{c}(\boldsymbol{x})$ denote the color space feature(s) associated with a pixel. Then a GMM distribution for $\boldsymbol{c}$ is given as:

$$p(\boldsymbol{c}|R) = \sum_{k=1}^{K} w_k \mathcal{N}(\boldsymbol{c}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k). \quad (1)$$

Here K is the number of mixtures, w_k, $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ are the weight, mean vector and covariance matrix of the k^{th} mixture component. $\mathcal{N}$ indicates a Gaussian distribution with specified mean vector and covariance matrix. R is an indicator of the region (lip (L) or non-lip (N)). We used diagonal covariance matrices in this work.

We use the EM algorithm for training the GMMs and we initialize the iterations with the k-means algorithm. We call these distributions generic region models in the following discussion.

2.4. Adaptation and Testing

During testing, we first adapt the generic region models to the subject of interest by initially choosing conservative regions in the test image making use of the extracted (or assumed) lip corners. We find two concentric ellipses that pass through the lip corners as shown in Figure 4 part (a). Inside of the smaller ellipse is assumed to be a part of the lip region and outside of the outer ellipse is assumed to be a part of the non-lip region. We choose the ellipses such that for almost all subjects the assumptions are correct. We then randomly take a fixed number of samples from two assumed regions to adapt the generic models to the subject. This adaptation step yields models that are well-suited to the subject of interest. We use MAP adaptation as described in [10] with a relevance parameter ρ. After adaptation of GMMs, we obtain adapted regional models for the subject of interest.

For testing, we first find the region of interest using the lip corner points. For each pixel within the region of interest, we calculate a detection score based on the likelihood ratio as follows:

$$\hat{S}(\boldsymbol{x}) = \log p(\boldsymbol{c}(\boldsymbol{x})|N) - \log p(\boldsymbol{c}(\boldsymbol{x})|L) + \log \frac{P(N)}{P(L)}. \quad (2)$$

Here $P(N)$ and $P(L)$ denote the probability that a pixel belongs to non-lip and lip regions, respectively. This score is precisely the (natural) logarithm of the likelihood ratio plus a prior imbalance term. The range of the score function is $(-\infty, \infty)$. We can assume $P(N)/P(L)$ to be in the range 5-10 since there are more non-lips than lips in a typical region of interest. In order to remove regional discontinuities, we median filter this score field using a 9×9 window.

Since the logarithm of a small likelihood value tends towards $-\infty$, it may be beneficial to limit the dynamic range of the score function. To achieve this, we obtain a clipped score $\tilde{S}(\boldsymbol{x})$ which is obtained by limiting the absolute value of the score by S_M which is the maximum absolute score allowed.

We expect pixel $\boldsymbol{x}$ to belong to the lip region if the score is less than a chosen threshold and vice versa. This threshold can be varied to adjust precision-recall trade-off (or ROC curve). However, we would like to choose a single optimal threshold value in this work, for which we make use of two ellipses that go through lip and non-lip regions as shown in Figure 4 part (b).

We calculate the means μ of the clipped score field $\tilde{S}(\boldsymbol{x})$ values on the boundary pixels of each ellipse. We then choose the single best threshold value to be in between these two means, given by $t_{opt} = k\mu_{lip} + (1-k)\mu_{nonlip}$. We experimentally found that using a k value larger than 0.5 worked better.

Our level-set formulation requires a score field which is between -1 and +1 and we would like to have a negative value within the lip region and a positive value within the non-lip region ideally. So, we need to map the score value to the range $(-1, 1)$. For this purpose, we *linearly* map the shifted score value to the desired range by dividing the score by the maximum absolute score Z:

$$S(\boldsymbol{x}) = (\tilde{S}(\boldsymbol{x}) - t_{opt})/Z. \quad (3)$$

We also experimented with some more sophisticated score mapping techniques with no improvement in results.

Figure 3 shows an example potential field surface $S(\boldsymbol{x})$ and the final lip contour which is governed by it.

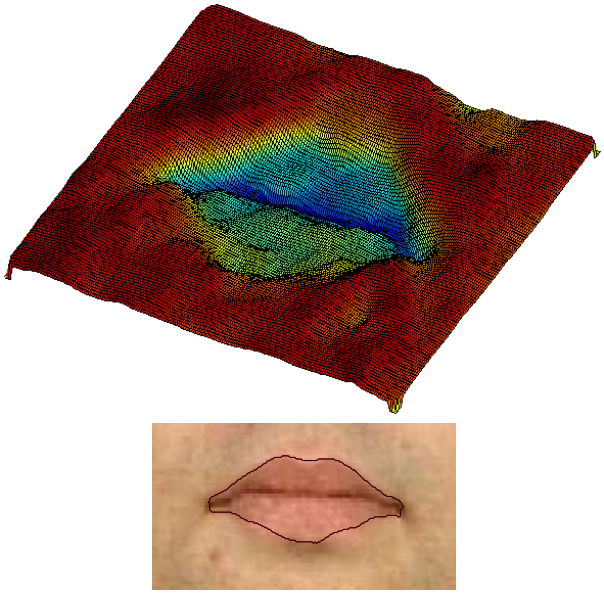
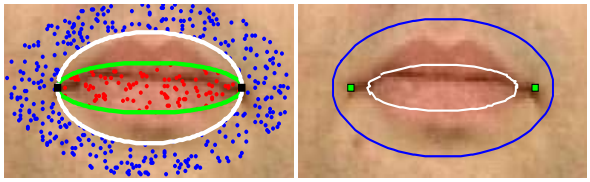


Figure 3: Potential field surface and segmented lip boundary



(a) (b)

Figure 4: (a) Sampling for adaptation (b) Ellipses used for score mapping

2.5. Curve Evolution Using Level Sets

We consider a level set formulation [9] based on both region and image gradients information. For this purpose, we construct color based potential fields to be used in level set equation to drive the interface to the boundaries of a lip and stop there by the help of a stopping function which uses image gradients. The initial curve is chosen to be an ellipse which passes through the extracted lip corners around the mouth. The evolution of the level set function is given by:

$$\Phi_t + F|\nabla\Phi| = 0, \tag{4}$$

where the speed function F on a pixel $\mathbf{x}$ is designed as:

$$F(\boldsymbol{x}) = -g(\boldsymbol{x})(\epsilon\kappa + \sum_{i=1}^{n} \alpha_i S_i(\mathbf{x})), \tag{5}$$

where g, κ and S denote stopping function, curvature and the potential field constructed from a color space, respectively. The coefficients ϵ and α_i are positive scalars. The number of color spaces used is given by n. We have used up to three color spaces together in this paper. The potential field $S_i(\boldsymbol{x})$ is computed using equation (3) for the i^{th} color feature.

The stopping function is designed in terms of the gradient of the Gaussian smoothed image as follows,

$$g(\boldsymbol{x}) = \frac{1}{1 + |\nabla((G * I)(\boldsymbol{x}))|^p}\ ,\ p \geq 1. \tag{6}$$

3. Experiments

We performed experiments of the proposed method using images from the Sabanci University Turkish audio-visual (SUTAV) database which has been collected at Sabanci University, Istanbul, Turkey. We obtained 100 training images and 100 test images from the database. In addition, we used 220 additional images (a total of 320 images for training) from the IMM database [11] to train the model for the lip-corner extraction module. The training part of SUTAV contains 60 female and 40 male subjects. The test data consists of 56 females and 44 males. The 16 of males have facial hairs such as moustaches and beards. The database is challenging due to poor lightening conditions.

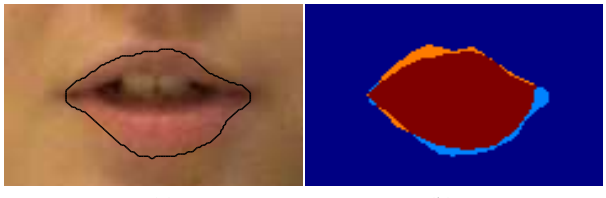
After face detection, we resized face images to 300×320 dimensions. We first detect lip corners using the method described. We used a window size of 41×61 and a PCA dimension of 50. We used 3 mixtures in GMMs to represent lip-corner-patches. For lip boundary extraction, we used 100 training images from SUTAV for GMM color space training with 5 mixtures. We adapted the GMMs using 300 and 900 random samples from lip and non-lip regions obtained using ellipses with a relevance factor of $\rho = 50$. Our score limit was $S_M = 4$. We found the threshold t_{opt} using 100 and 150 lip and non-lip samples and $k = 0.6$. In the level-set formulation, we used numeric values of $\epsilon = 7$ and $\alpha_i = 12$, and $p = 1.6$ which were determined experimentally.

3.1. Performance Metric

In order to assess the segmentation performance we used the following *precision* (p) and *recall* (r) metrics,

$$p = \frac{t_p}{t_p + f_p}\ ,\ r = \frac{t_p}{t_p + f_n}, \tag{7}$$

where t_p , f_p and f_n denote the true positives, the false positives and the false negatives with respect to ground truth binary image of the lip, respectively. The p and r are closer to 1, the better the segmentation. Segmented lip region is equal to $t_p + f_p$. Figure 5 shows a segmented lip and its t_p in brown, f_p in light blue and f_n in orange colors.



(a) (b)

Figure 5: (a) Segmented boundary (b) t_p in brown, f_p in light blue and f_n in orange. ($p = 0.9178$ $r = 0.9212$)

3.2. Color Spaces

In our experiments, we used 4 types of color space: $\{RGB\}$, $\{\frac{R}{R+G}\}$, $\{Hue\}$ and $\{rg\}$, to train GMMs and to construct potential fields. We define $r = R/(R+G+B)$ and $g = G/(R+G+B)$ as red and green ratios independent of illumination, and $\{rg\}$ denotes this normalized chromatic space. The $\{\frac{R}{R+G}, Hue^*\}$ and $\{\frac{R}{R+G}, Hue, rg\}$ combined color spaces are also employed. In the first combined color space, $\{Hue^*\}$ represents hue image itself, used as potential field, after mapping to the range (-1,1).

3.3. Results

The final segmented binary image is also post-processed to get rid of spurious blobs and pixels (using connected component analysis and choosing the biggest blob) to increase the accuracy of results. Figure 6 shows some of the visually good segmentations. In failure cases of our method we observed that there are leakages through a bowl between upper lip and nose and through a saddle point located between the lower lip and the chin. Figure 7 depicts such poor segmentation results. Table 1 tabulates the average metric values of different color spaces and their combinations for the performance evaluation of boundary segmentation process. We obtained the best results with $\{\frac{R}{R+G}\}$ and $\{\frac{R}{R+G}, Hue, rg\}$. The results show that it is possible to obtain about 85% average precision and recall performance in lip detection accuracy using the proposed technique.

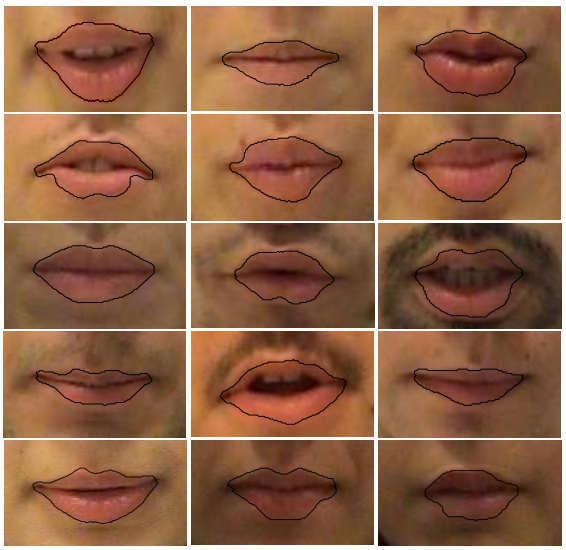


Figure 6: Examples for good segmented lip boundaries.

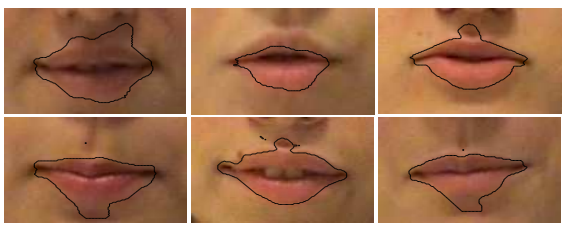


Figure 7: Examples for bad segmented lip boundaries.

Table 1: *Precision and recall values.*

Color Spaces	p	r
RGB	0.6252	0.9175
$R/(R+G)$	0.7981	0.9393
Hue	0.7387	0.9680
rg	0.7503	0.9736
$R/(R+G), Hue^*$	0.8810	0.7398
$R/(R+G), Hue, rg$	0.8125	0.9135

4. Conclusions and Future Work

In this work, we have introduced a method for lip segmentation based on GMMs for color space modeling and curve evolution using level set formulation. We have demonstrated performances of adaptively trained color spaces and their combinations. Conducted experiments show that the results are promising but we still need improvements to increase the robustness of the proposed method. Different color spaces which are not mentioned in this work and constraints on lip shape can be imposed to decrease failure cases.

5. Acknowledgements

This research has been funded by TUBITAK (Scientific and Technical Research Council of Turkey), research support program (program code 1001), project number 107E015.

6. References

[1] Paul Kuo, Peter Hillman and John Hannah, “Improved Lip Fitting and Tracking For Model-Based Multimedia and Coding”, International Conference on Visual Information Engineering Conference, Glasgow, UK, pp. 251-258, 2005.

[2] Mohammad Sadeghi, Josef Kittler and Kieron Messer, “Segmentation of Lip Pixels For Lip Tracker Initialization”, International Conference on Image Processing, ICIP, IEEE, Greece, 2001.

[3] Rainer Stiefelhagen, Jie Yang, Alex Waibel, “A Model based Gaze Tracking System”, Proc. of IEEE International Joint Symposia on Intelligence and Systems, pp. 304-310, Rockville Maryland, 1996.

[4] Nicolas Eveno, Alice Caplier, Pierre-Yves Coulon, “Accurate and quasi-automatic lip tracking”, IEEE Trans. Circuits Syst. Video Technology, vol. 14, no. 5, pp. 706-715, 2004.

[5] C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, and D. Vergyri, Large-vocabulary audio-visual speech recognition: A summary of the Johns Hopkins Summer 2000 Workshop, Proc. Works. Multimedia Signal Process. (MMSP), pp. 619-624, Cannes, France, 2001.

[6] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A.W. Senior, Recent advances in the automatic recognition of audio-visual speech, Invited, Proceedings of the IEEE, vol. 91, no. 9, pp. 1306-1326, 2003.

[7] Viola, P., and Jones, M., “Rapid object detection using a boosted cascade of simple features”, In IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2001.

[8] P. M. Hillman, J. M. Hannah, P. M. Grant, “Global Fitting of a Facial Model to Facial Features for Model-Based Video Coding”, Proc. of the 3rd International Symposium on Image and Signal Processing and Analysis, pp. 359-364, 2003.

[9] J.A. Sethian, “Level Set Methods: Evolving Interfaces in Geometry, Fluid Mechanics, Computer Vision and Material Sciences”, Cambridge University Press, 1996.

[10] Douglas A. Reynolds, Thomas F. Quatieri and Robert B. Dunn, “Speaker Verification Using Adapted Gaussian Mixture Models”, Digital Signal Processing, Vol: 10, Issue: 1-3, Jan 2000.

[11] http://www2.imm.dtu.dk/ aam/datasets/datasets.html