Effective Lip Localization and Tracking for Achieving Multimodal Speech Recognition

Wei Chuan Ooi, Changwon Jeon, Kihyeon Kim, Hanseok Ko and David K. Han

Abstract Effective fusion of acoustic and visual modalities in speech recognition has been an important issue in Human Computer Interfaces, warranting further improvements in intelligibility and robustness. Speaker lip motion stands out as the most linguistically relevant visual feature for speech recognition. In this paper, we present a new hybrid approach to improve lip localization and tracking, aimed at improving speech recognition in noisy environments. This hybrid approach begins with a new color space transformation for enhancing lip segmentation. In the color space transformation, a PCA method is employed to derive a new one dimensional color space which maximizes discrimination between lip and non-lip colors. Intensity information is also incorporated in the process to improve contrast of upper and corner lip segments. In the subsequent step, a constrained deformable lip model with high flexibility is constructed to accurately capture and track lip shapes. The model requires only six degrees of freedom, yet provides a precise description of lip shapes using a simple least square fitting method. Experimental results indicate that the proposed hybrid approach delivers reliable and accurate localization and tracking of lip motions under various measurement conditions.

1 Introduction

A multimodal speech recognition system is typically based on a fusion of acoustic and visual modalities to improve its reliability and accuracy in noisy environments. Previously, it has been shown that speaker lip movement is a significant visual component that yields linguistically relevant information of spoken utterances. However, there have been very few lip feature extraction methods that work robustly under various conditions. The difficulty is caused by variation of speakers, visual capture devices, lighting conditions, and low discriminability in lip and skin color.

W.C. Ooi (✉)
School of Electrical Engineering, Korea University, Seoul, Korea
e-mail: wcooi@ispl.korea.ac.kr

S. Lee et al. (eds.), *Multisensor Fusion and Integration for Intelligent Systems*,
Lecture Notes in Electrical Engineering 35, DOI 10.1007/978-3-540-89859-7_3,

Historically, there have been two main approaches [1] in extracting lip features from image sequences. The first method is called the Image-based approach. In this approach, image pixels (e.g. intensity values) around the lip region are used as features for recognition. For instance, these approaches are based on a DCT or a PCA method. The projected low dimensional features are used for speech recognition. The extracted features not only consist of lip features but also of other facial features such as tongue and jaw movement depending on ROI size. The drawback is that it is sensitive to rotation, translation scaling, and illumination variation.

The second type is known as the model-based method. A lip model is described by a set of parameters (e.g. height and width of lips). These parameters are calculated from a cost function minimization process of fitting the model onto a captured image of the lip. The active contour model, the deformable geometry model, and the active shape model are examples of such methods widely used in lip tracking and feature extraction. The advantage of this approach is that lip shapes can be easily described by low order dimensions and it is invariant under rotation, translation, or scaling. However, this method requires an accurate model initialization to ensure that the model updating process converges.

In this paper, we propose a model-based method designed primarily to improve accuracy and to reduce the processing time. The model-based method requires good initialization to reduce the processing time. By integrating color and intensity information, our algorithm maximizes contrast between lip and non-lip regions, thus resulting accurate segmentation of lip. The segmented lip image provides initial position for our point based deformable lip model which has built in flexibility for precise description of symmetric and asymmetric lip shapes.

We describe in more detail of the PCA based color transformation method in Sect. 2. In Sect. 3, we present a new deformable model for Lip Contour Tracking. We also describe cost function formulation and model parameter optimization in the same section. Experimental Results and comparison with other color transformation are presented in Sect. 4. The conclusion is presented in Sect. 5.

2 Lip Color and Intensity Mapping

Many methods have been proposed for segmenting the lip region that based on image intensity or color. We propose a new color mapping of the lips by integrating color and intensity information. Among color based lip segmentation methods are Red Exclusion [2], Mouth-Map [3], R–G ratio [4], and Pseudo hue [5]. Theoretically, pseudo hue method gives better color contrast, but we found that it is only useful in performing coarse segmentation which is not adequate for our purpose. Thus, we perform a linear transformation of RGB components in order to gain maximum discrimination of lip and non-lip colors. We employ a PCA to estimate the optimum coefficients of transformation. From a set of training images, N pixels of lip and non-lip are sampled and its distribution shown in Fig. 1(a). Each pixel is regarded as three dimensional vector $\mathrm{x}_i = (R_i, G_i, B_i)$. The covariance matrix

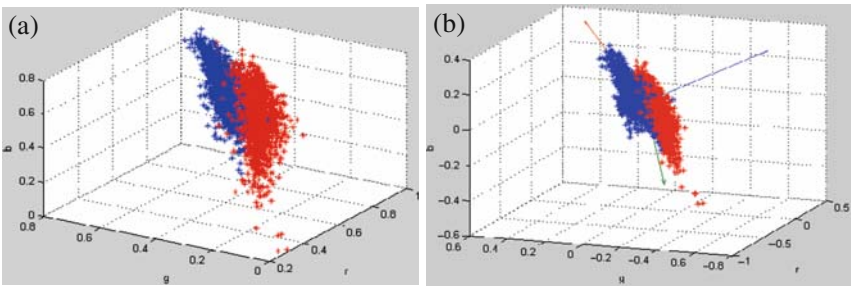

Fig. 1 (**a**) Distribution of lip and non-lip in RGB color space (60 people under different lighting conditions), (**b**) Eigenvectors of distribution in Fig. 1(a)

is obtained form the three dimensional vector and the associated eigenvectors and eigenvalues are determined from the covariance matrix.

$v_3 = (v_1, v_2, v_3)$ is an eigenvector corresponding to the third smallest eigenvalue where lip and non-lip pixels are the least overlapping as shown in Fig. 2(c). Experimentally, $v_1 = 0.2$, $v_2 = -0.6$, $v_3 = 0.3$ are obtained. Thus a new color space, C is defined as

$$C = 0.2 \times R - 0.6 \times G + 0.3 \times B \tag{1}$$

The new color space C is normalized as

$$C_{\text{norm}} = \frac{C - C_{\min}}{C_{\max} - C_{\min}} \tag{2}$$

Note that after normalization, the lip region shows higher value than the non-lip region. By squaring the C_{norm}, we can further increase the dissimilarity between these two clusters as shown in Fig. 3. A similar conversion of RGB values using the Linear Discriminant Analysis (LDA) was employed by Chan [6] to direct the evolution of snake. PCA based method is simpler compared to LDA especially in dealing with three dimensional RGB components.

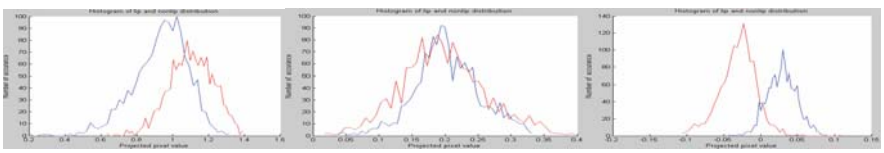

Fig. 2 (**a**) Histogram of projected pixels onto first principle component, (**b**) Histogram of projected pixels onto second principle component, (**c**) Histogram of projected pixels onto third principle component

Fig. 3 (**a**) Original image, (**b**) Transformed image, C, (**c**) C squared image

Fig. 4 (**a**) C_{map} image, (**b**) C_{map} negative image, (**c**) Gray-scale image

After the color transformation, the C squared image may still show low contrast in the upper lip region. This problem can be resolved by using the intensity information I. The upper lip region typically consists of lower intensity values. So by combining the C squared image (which is well separable in the lower lip) and intensity image (which has a stronger boundary in the upper lip), we can obtain an enhanced version of the lip color map C_{map} as follows.

$$C_{map} = \alpha C_{squar} + (1 - \alpha)\frac{1}{I} \tag{3}$$

Empirically, $\alpha = 0.75$ are derived. Higher weight is given to the C squared image since it captures most of the lip shape except the upper part and corners of lips.

2.1 Threshold Selection

In this paper, the global threshold is selected based on Otsu [7] method. The optimal threshold T_{opt} is chosen so that between classes variance σ_B^2 is maximized.

$$T_{opt} = \underset{0<T<1}{Arg\ \max}\ \sigma_B^2(T) \tag{4}$$

Fig. 5 Segmented images

3 Lip Contour Tracking

3.1 Lip Model

Most of the deformable geometric models are established using quadratic fittings (e.g. parabolic) with a prior assumption of lip shape being always symmetric about the center axis. Our lip model is an enhanced version of the proposed method in [8]. In [8] the writers integrated flexibility and constrained deformable template with point distribution model in order to reduce computations. However the geometric model in the paper is described by 15 parameters resulting in significant computation for the parameter updating process. Our proposed lip model is established by six parameters and is composed of three curves defined as follows:

- Lower lip, $\{0 < x < 1\}$

$$y_{low} = \alpha_{low} \cdot x \cdot (\log_2 x) + \beta_{low} \cdot (1 - x) \cdot (\log_2(1 - x)) + \gamma_{low} \left(\frac{(x - 0.5)^4}{0.5^4} - \frac{(x - 0.5)^2}{0.5^2} \right) \tag{5}$$

- Upper right lip, $\{0.5 \leq x < 1\}$

$$y_{up\_r} = -3.148 \cdot \alpha_{up\_r} \cdot (x - 0.4)^{\frac{1}{2}} \cdot (\log_2 x) + \gamma_{up\_r} \left(\frac{(x - 0.5)^4}{0.5^4} - \frac{(x - 0.5)^2}{0.5^2} \right) \tag{6}$$

- Upper left lip, $\{0 < x \leq 0.5\}$

$$y_{up\_l} = -3.148 \cdot \alpha_{up\_l} \cdot (0.6 - x)^{\frac{1}{2}} \cdot (\log_2(1 - x)) + \gamma_{up\_l} \left(\frac{(x - 0.5)^4}{0.5^4} - \frac{(x - 0.5)^2}{0.5^2} \right) \tag{7}$$

3.2 Parameters Description

There are six parameters to fully model the lower and upper lips. Parameters α_{low} and β_{low} control vertical height and skewness of lower lip as shown in Fig. 6.

- If $\alpha_{low} = \beta_{low}$, lip (lower) is symmetric with respect to center point, then center height=$\alpha_{low} = \beta_{low}$
- If $\alpha_{low} > \beta_{low}$, lip shape slides to the left
- If $\alpha_{low} < \beta_{low}$, lip shape slides to the right

The parameter γ_{low} controls curvature of lower lip shape with values between 0.15 and 0.15 as shown in Fig. 7.

Compared to the lower lip, upper lip shape remains relatively symmetric. This is due to the fact that the lower lip motion is a result of the mandible movement. The articulation available of the jaw joint allows the mandible movement of left or

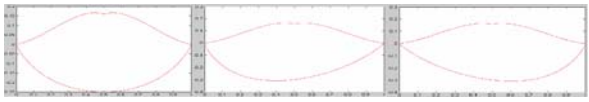

Fig. 6 Lip model for above cases

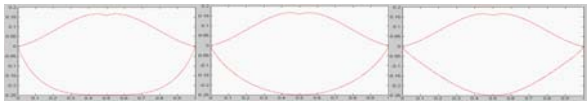

Fig. 7 (**a**) $\gamma_{low} = -1.5$, (b) $\gamma_{low} = 0$, (c) $\gamma_{low} = 1.5$

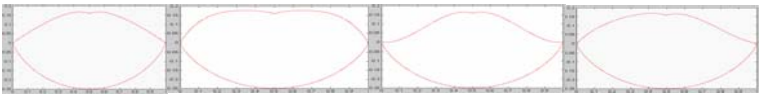

Fig. 8 (**a**) $\gamma_{up\_r} = \gamma_{up\_l} = 0$ (**b**)−0.3,−0.3 (**c**) 0.2, 0.2 (**d**)−0.2, 0.1

Fig. 9 Lip contour points converge from initial position to optimum position

right of the centerline of the face resulting in asymmetric movement of the lower lip. Thus, we assume the upper lip always remains symmetric

- $\alpha_{up\_l} = \alpha_{up\_r}$ = center height of upper lip
- γ_{up_r} and γ_{up_l} control curvature of upper lip (with value between 0.2 and −0.3)

Lip shapes according to the variations of the upper lip parameters are shown in Fig. 8.

3.3 Model Initialization and Normalization

Our model initialization is based on a segmented lip shape image. In order to reduce the computation time, we limited our model with just 16 points of p = $\{p_1, p_2, ..., p_{16}\}$ where $p_i = (x_i, y_i)$. The lip points are labeled in anti-clockwise direction starting from the left corner. These points are divided into three groups where $p_1, ..., p_9$ describe lower lip, $p_9, ..., p_{13}$ describe upper right points, and $p_{13}, ..., p_{16}$ describe upper left. The contour point normalization process is applied to reduce processing time and to simplify the curve fitting process. The left corner point is fixed as the origin. Rotation and scaling transformations are employed to normalized all points so that p_1 is at (0, 0) and p_9 is at (1, 0). Reverse normalization is applied after curve fitting for obtaining the original coordinates.

3.4 Model Optimization

The optimization procedure is an iterative process and the lip points are adjusted in order to reduce the cost function in each iteration process.

Our cost function F is defined in (8)

$$F = \arg\min_{p} \sum_{i=1}^{16} aE_{\text{int}}(p_i) + bE_{ext}(p_i) + cE_{bal}(p_i) \tag{8}$$

$E_{\text{int}}(p_i)$ is an energy function dependent on the shape of the contour points and it is the continuity energy that enforces the shape of the contour. $E_{ext}(p_i)$ is an energy function based on image properties (we use gradient in this paper). $E_{bal}(p_i)$ is a balloon force that causes the contour to expand (or shrink). In most cases, our binary image provides a good model initialization. Hence, the model usually takes only 5–8 iterations to converge to lip shape in a given image.

3.5 Model Fitting to Contour Points

We used least square approach to fit the model onto optimum contour points. By fitting the model on contour points, we can constraint the deformation of contour points and preserve a legal shape of lip shape. Furthermore, from the fitted model parameters, lip features can be extracted and used in visual speech recognition. The least square fitting is performed for three parts of outer lip separately when optimums lip contour points are obtained from optimization process. For example, in lower lip model that employs three parameters, $\theta = \{\alpha_{low}, \beta_{low}, \gamma_{low}\}$ show a process of least square method to fit model on contour points and parameter values are obtained. Note that p_1 and p_9 are not included since these two points are fixed in the normalization process.

$$
\begin{aligned}
H &= \begin{bmatrix} x_2 \cdot (\log_2 x_2)\ (1-x_2) \cdot (\log_2(1-x_2)) \left(\frac{(x_2-0.5)^4}{0.5^4} - \frac{(x_2-0.5)^2}{0.5^2}\right) \\ \vdots \qquad\qquad \vdots \qquad\qquad \vdots \\ x_8 \cdot (\log_2 x_8)\ (1-x_8) \cdot (\log_2(1-x_8)) \left(\frac{(x_8-0.5)^4}{0.5^4} - \frac{(x_8-0.5)^2}{0.5^2}\right) \end{bmatrix} \\
\theta &= \begin{bmatrix} \alpha_{low} \\ \beta_{low} \\ \gamma_{low} \end{bmatrix}, Y = \begin{bmatrix} y_{low2} \\ \vdots \\ y_{low8} \end{bmatrix} \\
&\because H\theta = Y \\
&\Rightarrow \theta = (H^T H)^{-1} H^T Y
\end{aligned}
\tag{9}
$$

After least square fitting, new contour points can be found by equally deriving from fitted curves. These contour points are de-normalized by employing reverse scaling and rotation. Before new iteration is processed, sum of distance of new contour points is computed and compare to sum of distance of previous contour points. If the result is less than threshold the iteration process will be terminated.

4 Experiment Result

4.1 Lip Contour Extraction Result

In order to test the performance of our proposed hybrid procedure, we use 2,000 lip images with different sizes over 50 people (not including images that were used in color space training). In the testing images, we also use some images which consist of complex background like mustache and beard. For evaluating the flexibility of model asymmetrical lip shape images are incorporated.

Overall, 97% of the lip contours are accurately extracted. Figures 10 and 11 show examples of such images. We also show that our proposed lip color and intensity mapping have successfully improved the lip contour extraction performance under different lightning conditions. With our algorithm implemented in Matlab, the average computation time for 85×100 size images was approximately 0.9 s.

From experimental results, it had shown that accuracy of lip contour tracking not merely depend upon flexibility of built model but also contingent on preprocessing part. For instance, our proposed lip color and intensity mapping efficiently maximize dissimilarity of lip and non-lip region. This mapping result will be used to localized lip model and also provide clear edge information to derive contour points moving toward lip boundary.

Fig. 10 Lip contour extraction results of female and also male lips

Fig. 11 Lip contour extraction results under different lightning conditions

4.2 Comparative Studies of Lip Mapping

We apply a quantitative technique to evaluate the performance of our color space transformation algorithm. Since no ground truth is available, we manually draw the boundaries of 25 lip images. The first measurement method is the degree of overlap (DOL) between the lip and the non-lip histograms. DOL [9] is used to measure discriminability of the transformed color spaces for differentiating the lip and the non-lip colors. A lower percentage of DOL means a higher contrast between the lip and the non-lip regions.

$$\begin{aligned} DOL &= \sum_{i=0}^{1} \min(P_{lip}(i), P_{nonlip}(i)) \\ \text{where} \quad P_{lip}(i) &= Num_{lip}(i)/Total_{lipPixel} \\ P_{nonlip}(i) &= Num_{nonlip}(i)/Total_{nonlipPixel}, \{0 \le i \le 1\} \end{aligned} \tag{10}$$

The second method is Classification Error (CE) which is the average of the False Positive (FP) rate and the False Negative (FN) rate FP is error rate of classifying a non-lip as a lip pixel. FN is the error rate in classifying the lip as non-lip pixel.

$$\begin{aligned} CE &= (FN + FP)/2 \\ \text{where } FN &= False_{nonlip}/(False_{nonlip} + True_{lip}) \\ FP &= False_{lip}/(False_{lip} + True_{nonlip}) \end{aligned} \tag{11}$$

From the results, we can see that our proposed lip color transformation method gives the lowest DOL and CE.

Table 1 Comparison of DOL and CE based on five mapping methods for below three images

	Image 1 (Fig. 12)		Image 2 (Fig. 13)		Image 3 (Fig. 14)	
	DOL	CE	DOL	CE	DOL	CE
MM	0.274	0.195	0.202	0.237	0.232	0.229
Our	0.182	0.150	0.201	0.143	0.173	0.144
RE	0.372	0.466	0.423	0.319	0.291	0.5
PH	0.223	0.243	0.294	0.188	0.210	0.162
RG	0.977	0.496	0.318	0.500	0.993	0.499

Table 2 Comparison for average DOL and CE for below three images and additional 22 testing images

	MM	Our	RE	PH	RG
Average DOL (%)	23.8	16.4	36.7	22.9	55.7
Average CE (%)	21.4	12.5	35.8	17.2	36.4

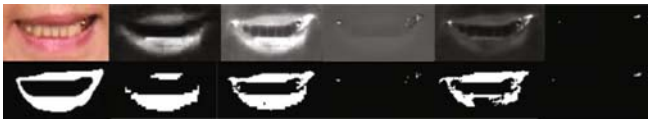

Fig. 12

Fig. 13

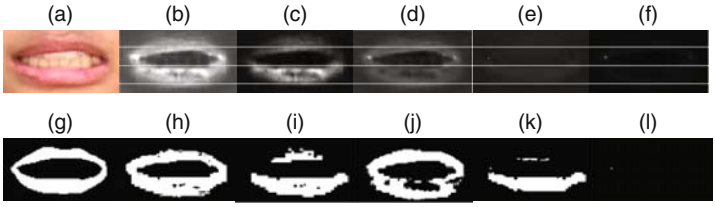

Fig. 14 (**a**) Test image, (**b**) Mouth-map method (MM), (**c**) Our method, (**d**) Red Exclusion method (RE), (**e**) Pseudo Hue method (PH), (**f**) R–G ratio method (RG), (**g**) Ground truth image, (**i**) is segmented image of (c) based on Otsu thresholding, (**h**), (**j**), (**k**), (**l**) are segmented images of (b), (d), (e), (f) with the threshold values proposed by corresponding previous methods

5 Conclusion

In this paper, we describe a new hybrid approach to improve lip localization and tracking. The first part of our proposed algorithm is lip mapping based on color and intensity information. From experimental results, our proposed mapping method successfully enhances the contrast between lip and non-lip regions. Results from the contrast enhancement process allowed more accurate lip region segmentation. In the second part, a new flexible while constrained deformable geometric model is established to accurately locate and track lip shape. Overall, our implemented hybrid approach has shown high reliability and is able to perform robustly under various conditions.

Acknowledgments This research was supported by MKE (Ministry of Knowledge Economy), Korea, under the ITFSIP (IT Foreign Specialist Inviting Program) supervised by the IITA.

References

1. G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. W. Senior, Recent advances in the automatic recognition of audio-visual speech, Invited, IEEE Proc., 91, 1306–1326, 2003.

2. T. W. Lewis.and D. M. Powers, Lip feature extraction using Red Exclusion, Proc. Selected papers from Pan-Sydney Workshop on Visual Information Processing, pp. 61–67, 2000.
3. R. L. Hsu, M. Abdel, A. K. Jain, Face detection in color images, IEEE Trans. Pattern Anal. Mach. Intelli., 2002.
4. S. Igawa, A. Ogihara, A. Shintani, and S. Takamatsu, Speech recognition based on fusion of visual and auditory information using full-frame color image, ZEZCE Trans. Fundam., 1996.
5. A. Hulbert and T. Poggio, Synthesizing a color algorithm from examples, Science, 239, 482–485, 1998.
6. M. T. Chan, Automatic lip model extraction for constrained contour-based tracking, ICIP, 848–851 1999.
7. N. Otsu, A threshold selection method from gray-level histograms, IEEE Trans. Syst. Man Cyber., 62–66, 1979.
8. S. L. Wang, W. H. Lau, and S. H. Leung, A new real-time lip contour extraction algorithm, ICASSP, 217–220, 2003.
9. T. C. Terrillon, M. N. Shirazi, and H. Fukamachi, Comparative performance of different chrominance skin chrominance models and chrominance spaces for the automatic detection of human faces in color images, Proc. IEEE Int. Conf. Autom. Face Gesture Recogn., 54–61, 2000.