Lip Feature Extraction Towards an Automatic Speechreading System

X. Zhang, R. M. Mersereau

Abstract

The use of color information can significantly improve efficiency and robustness of lip feature extraction capability over purely grayscale-based methods. Edge information provides another useful tool in characterizing lip boundaries. In this paper we present a method of integrating both types of information to address the problem of lip feature extraction for the purpose of speechreading. We first examine various color models and view hue as an effective descriptor to characterize the lips due to its invariance to luminance and human skin color, and its discriminative properties. We use prominent red hue as an indicator to locate the position of the lips. Based on the identified lip area, we further refine the interior and exterior lip boundary using both color and spatial edge information, where those two are combined within a Markov random field (MRF) framework. Experimental results are presented to show the effectiveness of this method.

1 Introduction

Various studies have demonstrated that automatic speech recognition systems can yield better recognition performance by adding visual information to the acoustic data, especially in environment corrupted by acoustic noise and multiple talkers. This motivated many research activities in the area of speechreading. It is generally agreed that most of the visual information is contained in the lips, therefore extraction of lip features is the first crucial step towards an automatic speechreading system.

Considerable research in automatic speechreading systems during the last sixteen years has been devoted to extracting lip contours from gray-scale image video ([1]). However most of these are relatively sophisticated dynamic contours/active shape/deformable templates methods, which prohibit real-time analysis. Other drawbacks include their sensitivity to lighting variations and appearance variations such as facial hair. In recent years, another approach using color information is gaining interest (see [2]-[4]). The main difficulty of lip feature extraction lies in the accuracy and reliability of the system. Recent research has shown that color is a powerful tool with regard to those two aspects. Unlike the gray-level approach, color image analysis increases the efficiency and robustness of locating the lips, and easily adapts to detect beards, teeth and tongue. However, certain restrictions and assumptions are required in those methods. [4] requires that the talker's head be fixed relative to the camera by using a micro-camera mounted on a light helmet. In [3], a 2D lookup table was manually determined from the sample images. In [2], individual chromaticity models were needed for each of the speaker. What we are looking for, in contrast, is an algorithm that allows a natural test environment with normal lighting conditions. Talkers can move around freely in front of the camera, and the models can be extended to new talkers.

This paper is organized as follows. Section 2 examines various color spaces and demonstrates results of finding mouth position in a video input. In Section 3, we refine the lip segmentation by using an MRF framework to combine both color and edge information. The experimental results and the summary are presented in Section 4 and 5, respectively.

2 Color analysis

We start by examining various color spaces. RGB is the most widely used among many existing color spaces. However the triple [R,G,B] represents not only color but also brightness, which hinders the effectiveness of color in detection. Several studies have shown that even though different people have different colors in appearance, the major difference lies in intensity rather than color itself. To separate the chromatic and luminance components, various transformed color spaces can be employed, such as the normalized RGB

space (we denote it as rgb in the following), YCbCr, and HSV. Many transforms from RGB to HSV are presented in the literature. Here the transformation is implemented after [5].

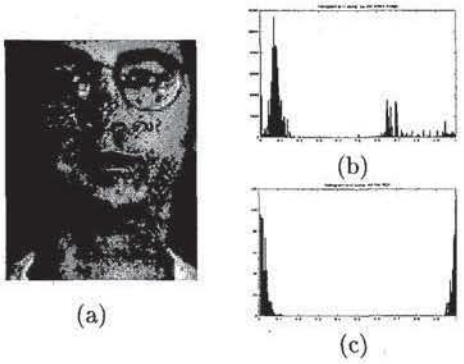

Figure 1: (a): Original image with a black contour highlighting the lip region, (b): Histogram of the hue component for the entire image, (c): Similar histogram for the lip region bounded within the black contour

To analyze the statistics of each color model, we build histograms of color components. We construct histograms for the entire image and for the extracted lip region bounded within the estimated boundary, as shown in Figure 1(a). From experiments on various video sequences taken under different test conditions and for different test subjects we have the following observations: i) Color components (r,g,b), (Cb,Cr) and (H) exhibit peaks in their histograms. This indicates that the feature distribution of the lip region is narrow and implies that the color for the lip region is fairly uniform. ii) The color histogram of (r,g,b) and (Cb, Cr) of the lip region more or less overlaps with that of the whole image, while the hue component has the least similarity between the entire image and lip region only (see (b) and (c) in Figure 1). This shows that hue has high discriminative power. iii) The distribution of (r,g,b) and (Cb,Cr) vary for different test subjects, while hue is relatively constant under varying conditions, such as lighting conditions, and for different talkers. We therefore conclude that hue is an appropriate model for our application.

Figure 1(c) shows the histogram of hue for the lip region. We observe that the red hue mainly falls into two separate subsets at the low and high ends of the whole range. Due to the wrap-around nature of hue, low values of hue lie close to high ones. For easy use of hue component, we shift hue by a 1/8 of the total length to the left. This results in a connected range for the red hue, which is close to 1 (if hue is defined in $[0,1]$). Figure 2(b) shows the hue-color image. Since the modified red hue value is at high end, the lips appear to be the brightest region. There is considerable noise in the hue image though. This is mainly related to the unfortunate singularity property of RGB to HSV conversion, which occurs when R=G=B (saturation=0)([6]). In order to use hue, we require that S must exceed a certain preset value. For segmenting the lip, we use the following H and S constraints:

$$BW(x,y) = \begin{cases} 1 & H(x,y) > H_0, S(x,y) > S_0 \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

where $H_0 = 0.8, S_0 = 0.25$ for $H/S \in [0,1]$. The accuracy of those two values are not very critical, and they proved to generalize well for other talkers. The resulting binary image is shown in Figure 2(c).

Figure 2: (a): Original image, (b): Hue image, (c): Resulting binary image after thresholding

From the binary image, our algorithm extracts the lip region from its surrounds. It works as follows: we sum up all horizontal pixels and relate the largest value to the lip line. Then we extract the two corners by detecting the two intensity extrema from the horizontal lip line. Based on the two corners, vertical middle line across the lip can be derived and the upper and lower lip are detected from the intensity. Once we detect the lip region, a generous big area around the lip is extracted for further processing. The detected lip area is shown as a white bounding box in Figure 2(c).

Note that the color space conversion, binary image thresholding and lip region extraction described above need to be done only once for the first image. For the proceeding frames in the sequence, we can estimate the lip region from the previous frame based on the assumption that the mouth doesn't move abruptly in subsequent frames.

In the following, only the sub-image of the lip region is considered.

3 Combining edge and color information

Edge characterizes object boundaries and provides additional useful information in lip feature extraction. Since hue in [5] is defined on the ring rather on the interval $\mathbf{R}$, standard edge detection doesn't work well with it. In [8] another hue definition was suggested, $H = \frac{R}{R+G}$. It is defined on $\mathbf{R}$, and achieves nearly as good a reduction of intensity dependency as the normal hue definition. The image based on this hue provides fairly good boundary information. We perform edge detection on the hue image using a Canny detector. In the Canny edge detection algorithm [7], the input image is convolved with the first derivative of a 1-dimensional Gaussian function in x- and y-direction separately, and it results in I'_x, I'_y. The magnitude of the result is then computed at each pixel (x, y) as

$$\sqrt{c_1 I'_x(x,y)^2 + c_2 I'_y(x,y)^2}, \tag{2}$$

where c_1 and c_2 are normally set to 1. Based on this magnitude, a non-maxima suppression and double thresholding algorithm are performed and the edge map is derived. Since the lip contains mainly horizontal edges, we assign $c_2 = 10$ to accentuate the importance of horizontal edges. This modification results in an improved edge map for lip images.

To combine edge and hue color information, we have chosen to use the machinery of the Markov random field (MRF). The reason is twofold. First, extraction of lip features recovers the true image from the noisy observed image. It is, therefore, an inverse problem with many possible solutions and is ill-posed [9]. This problem can be solved by the use of regularization methods employed in the MRF framework. Second, the MRF formulation allows us to embed many features of interest by simply adding appropriate terms in the energy function, therefore it provides an easy tool for fusing multiple low-level vision modules.

Our problem can be formulated as a 'site' labeling problem - to assign each site a label x_i from the set {lip,non-lip}, and b_i from {edge, non-edge}. The maximum *a posterior* (MAP) criterion is used to formulate what the best labeling should be. Bayes' Rule and the Hammersley-Clifford Theorem allows to reduce the estimation problem to the minimization of a Gibbs distribution energy function consisting of two parts: the prior energy U_i and the energy data term U_d.

U_i describes the interaction potential between neighbors, and regularizes the solution. It is expressed as

$$U_i(x,b) = \lambda_1 \sum_{c \in C} V_c(x) + \lambda_2 V_e(x,b). \tag{3}$$

C is the set of all cliques; here we use the 1^{st} order neighborhood system. The first term in (3) is responsible for piecewise smoothing and is given by

$$V_c(x) = \begin{cases} -1 & \text{if all } x_{ij} \text{ in } c \text{ are equal} \\ +1 & \text{otherwise.} \end{cases} \tag{4}$$

The second term in (3) can be written as

$$V_e(x,b) = \sum_{(i,j)} \Psi(|x_i - x_j|)(1 - b_{(i,j)}) + V_{eo}(b), \tag{5}$$

where the first term is for smoothing with $\Psi(0) = -1$, and $\Psi(\Delta) = 1$ for $\Delta \neq 0$, and the second one for boundary organization for the edges. We assign to each local edge configuration a potential based on the heuristics that edges are likely to be linked horizontally and that close parallel edges and isolated edge elements are improbable. The former are given a small potential as an encouragement, while the latter are given a large potential as a penalty. $b_{(i,j)}$ indicates the edge between site i and j. It is 1 if there is an edge, and 0 if otherwise. Since the edge map is defined on each pixel, we imaginarily shift the edge map by $\frac{1}{2}$ pixel downwards against the original image, and we have $b_{(i,j)} = e_i$ if the site j is one pixel below the site i. For simplicity, we only consider horizontal edges in this work.

The parameters in (3) are the weighting coefficients of the energy terms. λ_1 controls the degree of smoothing and λ_2 weights the importance of the presence of an edge. In our experiments, we set the ratio $\frac{\lambda_1}{\lambda_2}$ empirically to $\frac{1}{3}$.

The energy data term U_d binds the solution to the data and is defined as

$$U_d = \sum_i (y_i - \mu_{x_i})^2 / 2\sigma_{x_i}^2 \tag{6}$$

where y_i is the observed image data, μ_{x_i} and σ_{x_i} are the mean and variance of all pixels in the image with the region label x_i.

We first perform initial segmentation by thresholding the hue image, where the threshold is obtained by using Otsu's methods [10] based on histogram. Both μ and σ in equation (6) are also obtained from this method. In the second step, the segmentation labels are updated by utilizing the iterative deterministic algorithm proposed by Chou *et. al.* [11], known as the highest confidence first (HCF). This procedure converges to a local minimum of the Gibbs potential.

4 Results

We conducted tests on various sequences. Test persons have various skin complexions with no particular lip-stick. Results with different persons and different lip opening situations are demonstrated in Figure 3. We observe that the highlighted pixels can fairly well match the true lip area. The running curve of the labeling on the boundary of the lip is not very smooth in some images. But based on the obtained segmentation we can detect the key points on the lip and derive geometric features such as the width and height of the inner and outer lip fairly accurately, as seen in Figure 4. These features and their corresponding dynamic features are used for the speech recognition.

Besides the geometric dimensions of the lip, the visibility of the tongue and teeth also contributes to a better recognition. Note that it is trivial to detect the presence of these two features for a color image. For the former, we detect the "lip" labels within the inner lip region; and for detecting the teeth, we look for the pixels with the property: $|H - 0.5| < \epsilon$, where we use $\epsilon = 0.01$. Our experiments show that we can easily detect tongue and teeth.

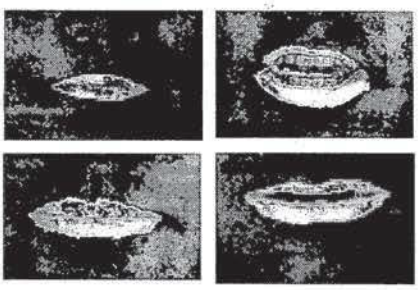

Figure 3: Segmented lip overlayed on original image

Figure 4: Detected key points on the lip

5 Conclusion

In this paper we have presented methods of extracting lip features for the purpose of automatic speechreading. Our algorithm first performs color space conversion from RGB to HSV on the whole image and locates the mouth region by means of thresholding. It subsequently segments the lip by using color and edge information, where those two are combined within a MRF framework. Relevant lip features can be successfully extracted from the segmented image.

Acknowledgments

We would like to thank the Advanced Multimedia Processing Lab at Carnegie Mellon University for providing us the database used in this work.

References

[1] D.G. Stock and M.E. Hennecke, "Speechreading by Humans and Machines", *NATO ASI Series F*, Vol. 150, Springer Verlag, 1996.

[2] M.U. Ramos Sanchez, J. Matas, and J. Kittler, "Statistical chromaticity models for lip tracking with B-Splines", *Proc. of the first international conference on Audio-and Video-based Biometric Person Authentication*, Lectures Notes in Computer Science, pp. 69-76, Springer Verlag, 1997.

[3] M. Vogt, "Interpreted Multi-State Lip Models for Audio-Visual Speech Recognition", *Proc. of the AVSP 97 workshop*, Sept. 1997.

[4] M. Lievin and F. Luthon, "Unsupervised Lip Segmentation under Natural Conditions", *ICASSP* 99.

[5] Keith Jack, "Video Demystified - A handbook for the Digital Engineer", 1996.

[6] J.R. Kender, "Instabilities in Color Transformations", *PRIP-77*, IEEE Computer Society, Troy NY, pp. 266-274, June 1977.

[7] J.F. Canny, "A Computational Approach to Edge Detection", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 8, pp. 679-698,1986.

[8] A. Hurlbert, and T. Poggio, "Synthesizing a Color Algorithm From Examples", *Science*, vol.239, pp. 482-485,1988.

[9] T. Poggio, V. Torre, C. Koch, "Computational vision and regularization theory" *Nature*, vol. 317 (26), pp. 314-319, Sept. 1985.

[10] Otsu, "Threshold Selection Method from Gray-level Histograms", *IEEE Transactoin Syst. Man Cybern.* vol.9 no. 1, Jan. 79.

[11] P. Chou, C. Brown, and R. Raman, "A Confidence-Based Approach to the Labeling Problem", Proc. IEEE Workshop on Computer Vision, pp. 51-56, Miami Beach, Florida, 1987.