ROBUST LIP LOCALIZATION ON MULTI-VIEW FACES IN VIDEO

Yi Wu[1], Rui Ma[2], Wei Hu[3], Tao Wang[3], Yimin Zhang[3], Jian Cheng[1], Hanqing Lu[1]

[1]National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Science
{ywu, jcheng, luhq}@nlpr.ia.ac.cn
[2]State Key Laboratory of Intelligent Technology and Systems, Tsinghua University, Beijing, P.R. China
mr02@mails.tsinghua.edu.cn
[3]Intel China Research Center, Beijing, P.R. China
{wei.hu, tao.wang, yiming.zhang}@intel.com

ABSTRACT

In this paper, a fast and robust multi-view lip localization algorithm in video is proposed. We consider lip localization as a binary classification problem, where a classifier is learned to distinguish between the lip and the region surrounding it. The classifier we use here is a histogram-based one which exploits the anthropometrical properties of the human face with the help of face scale normalization. Due to the perceptual uniformity and robustness for lip/skin color variations across different people, we adopt CIELUV color model to represent the color of lip. After classification, we propose a novel projection-cut algorithm by Spatial Deviation Analysis (SDA) to locate the lip, which is effective to deal with the background clutters. Experimental results on teleplay videos demonstrate that the proposed approach is efficient and robust for lip localization.

***Index Terms*—** Lip localization, video analysis, CIELUV, projection-cut, binary classification

1. INTRODUCTION

Lip localization is a crucial step in many video analysis problems such as talking face detection, audio-visual speech recognition, speaker recognition, etc. However, lip localization in video, especially in teleplay and movie, is a very challenging problem due to the following reasons: 1) unknown face poses and deformability of the lip, 2) low image qualities with changing illumination condition and shadow, 3) background clutters, 4) unpredicted lip movement together with head movement and 5) other factors, such as beard surrounding the lip.

In recent years, many techniques have been proposed for lip segmentation. Zhang [1] uses hue and edge features to achieve mouth localization and segmentation. They view hue as an effective descriptor to characterize the lip. However, the hue difference between lip and skin is too trivial to discriminate lip from skin. In [2] Eveno et al. detect characteristic points in the mouth region and then use a parametric model to fit the lip. They propose to convert the image from RGB color space to pseudo hue plane, where lip can be separated from skin better than the conventional hue plane. In the presence of beard or shadow, however, it's still difficult to distinguish lip from them. Liew [3] utilizes a uniform color space to describe the lip and use spatial fuzzy clustering algorithm to achieve lip segmentation for frontal faces, and it can get an accurate lip contour.

Despite many works have been proposed for mouth location and lip segmentation, most of the existing methods limit their use in indoor situations with controllable lighting condition and their experiments are based on full-frontal faces in good quality images. Few of them mentioned possible solutions to robust lip localization on multi-view faces in such as teleplay or movie.

Lip is a deformable object. In literatures, elliptical segments, B-spline curve, quadratic, cubic or quartics have been used to model lip shape [4]. However, it is hard to define a precise and uniform shape model for various lip occurrences on multi-view faces in teleplay or movie video. Therefore, we use a surrounding box to model the mouth and lip region. Although it may not sufficient to represent the real shape of various lips, the model could capture the main movement of the lip and be sufficient for some video analysis applications, such as talking face detection.

Our proposed method is demonstrated in Fig.3. We get lip Region-of-Interest (ROI) using the information got from the face detection module [4] and the facial feature points detection module [5]. Then we make a color space transformation to get a uniform lip description. Thus the problem turns to separate the lip pixels from the surrounding ones in the new color space. In [3] Liew et al. formulate this problem as a two-class clustering problem. However, we consider it as a binary classification problem which can use prior knowledge better. Due to the complicated condition in the video we need a dynamic and adaptive classifier. It is easily updated on-line. Based on these basic considerations we use a histogram based classifier, which exploits the anthropometrical properties of the human face, to distinguish the lip pixels. This classifier is simple and efficient to achieve our lip localization task.

1-4244-1437-7/07/$20.00 ©2007 IEEE

Finally, we propose a novel projection-cut algorithm by Spatial Deviation Analysis (SDA) to extract the lip surrounding box, which is effective to deal with the background clutters. Experimental results on teleplay videos demonstrate that the proposed approach is efficient and robust for lip localization.

The paper is organized as follows. Section 2 describes an algorithm for lip localization in video. The experimental results are demonstrated in Section 3, followed by the conclusion in Section 4.

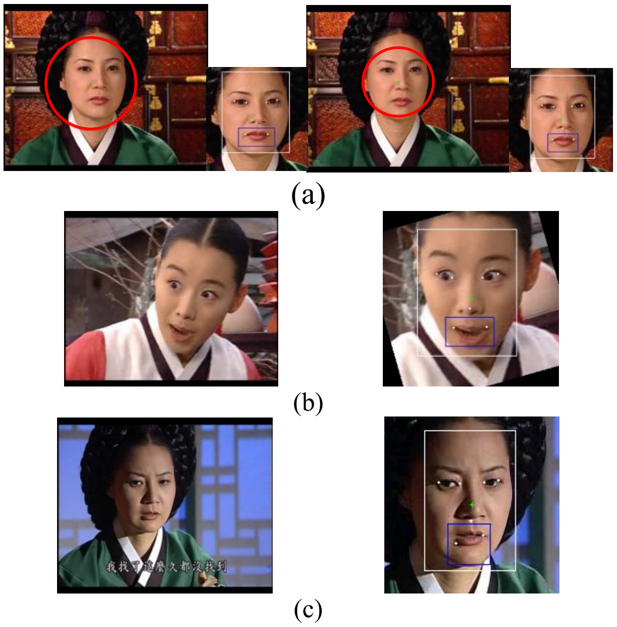

Figure 1: Lip ROI extraction. (a) The scale instability of the face detector outputs for two consecutive frames and faces with rectangle normalization. (b) The rotated face image. (c) The instability of the mouth corners. The green point is the face center from the origin face circle. The seven facial features are shown in white points and the blue rectangle is the lip ROI.

2. LIP LOCALIZATION

The lip localization is performed in two steps. The first step extracts the lip Region-Of-Interest (ROI) in each frame. The ROI is initialized by a set of roughly detected facial feature points. In the second stage the lip localization is achieved by binary classification and a projection-cut scheme.

2.1. Lip ROI Extraction

A multi-view face detector described in [4] is first applied on video frames. The face detector outputs the face's center location, scale and in-plane rotation angle. Then the face region is rotated according to the angle for the convenience of the following steps (Fig.1b).

In the next step, seven facial feature points are located in the face region [5], including the left and right corners of each eye, the tip of the nose, and the left and right corners of the mouth (Fig.1). We use the facial feature points to help initialize the lip ROI because the scale (illustrated as red circle in Fig.1a) got from the face detector sometimes is not very accurate. Figure 1 shows examples of the face detection and feature points localization. As illustrated in Fig.1a, the face scale from the detector is not stable or credible enough. What we can trust from the detector is that the mouth region is always included in the detected face region and below the face center. Utilizing the facial feature points we can generate a normalized face rectangle. After normalization, the scales of the similar faces are almost the same, as shown in Fig.1a. Also, the detected mouth corners provide us the initial guess of the exact positions. Typically the detected mouth corners are located close to the true positions (see Fig.1c). Thus the lip ROI can be identified according to the mouth corners and nose tip, as shown in a blue rectangle. This ROI region could effectively remove many background clutters, which is very helpful especially for profile faces.

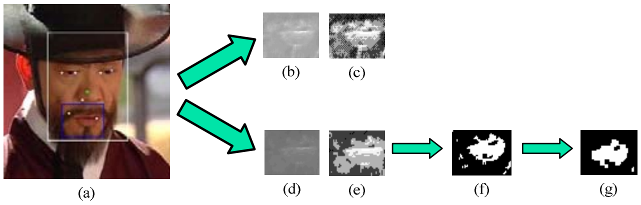

Figure 2: Color space transformation. Face image with lip ROI (a), pseudo hue image of lip ROI (b), contrast enhancement on image b (c), u-component image of lip ROI (d), contrast enhancement on image d (e), binary image of image d (f), image d after morphological operations (g).

2.2. Color space transformation

Although a number of researchers [1] have suggested that the skin hue is fairly consistent across different people, the colors of the lip and skin region usually overlap considerably. In [2] the pseudo hue color plane is used based on the observation that there is more green than blue in the skin while for lips these two components are almost the same. However, the pseudo hue value defined as R/(R+G) of beard and shadow can be very similar as that of the lip, as shown in Fig.2. The reason lies in that the pseudo hue value may be very high when all components of RGB are low. Therefore, pseudo hue plane is not appropriate for our purpose to distinguish the lip from its surroundings, such as beard.

Liew et al. [3] utilize two approximately uniform color spaces, CIELAB color space and CIELUV color space, to describe the lip. In our experiments the two color space transformations are almost the same and the A-component of CIELAB and U-component of CIELUV are more distinctive. For simplicity, we use the transformation from RGB color space to U-component plane of CIELUV color space. The result of the color transformation of the lip ROI is shown in Fig.2d. Although the contrast of the lip to its surrounding is very low in Fig.2d, it can remove the beard's influence and the classifier in the next stage can distinguish the lip very well. The difference between the lip and its surroundings can be significantly improved when enhance the contrast of the transformed image by histogram equalization (Fig.2e).

2.3. Histogram-based classifier

Now our aim is to separate the lip pixels from the surroundings in the transformed image. In [3] Liew et al. formulate this problem as a two-class clustering problem. However, we consider it as a binary classification problem. Due to the complicated condition in the video we need a dynamic and adaptive classifier and taking into account the efficiency: the classifier should be easily updated on-line and not too complicated. Here a histogram based classifier is used, which exploits the anthropometrical properties of the human face with the help of face scale normalization, to distinguish the lip pixels.

Through statistical study, we found that the lip-face area ratio is usually below 0.06. This anthropometrical property of face is very useful for classifier construction. We use this property to select an adaptive threshold. Specifically, we calculate the histogram (not normalized) in the lip ROI, and start to search the threshold. We select a point (bin position) as the threshold when the sum of the point to the whitest point is just below six percent of the number of pixels in the normalized face rectangle. The binary image after classification is shown in Fig.2f.

Let the histogram be represented as a d-dimensional vector $\{h_i\}_{i=0}^{d-1}$, where d is the number of bins, and let S_{face} denote the area of the face rectangle. Thus we can define the threshold selection criteria as:

$$\sum_{j=i}^{d-1} h_j - r \times S_{face} < 0; \quad \left(i \in \left[\arg\max_{i\in[0,255]} h_i , d-1 \right] \right)$$

where r is the lip-face ratio which we use in our study is 0.06. The objective is to find the first j satisfying the criteria.

Morphological closing and opening with an eight-neighborhood structuring element are used to smooth the binary image and eliminate small erroneous blobs. The morphological closing is realized by performing a dilation operation, followed by an erosion operation. This operation will fill up small holes and gaps. The morphological opening is realized by reversing the dilation and erosion operations and has the effect of opening up small-connected regions and protrusions. The final lip localization result is shown in Fig.2g.

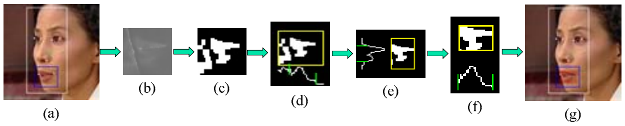

Figure 3: Fast lip localization using Projection-Cut method. Face image with lip ROI (a), color transformed image of lip ROI (b), classified image of image b (c), x-projection of first iteration (d), y-projection of first iteration after x-cut (e), x-projection of second iteration (f), face image with final lip localization (g). The white curve is the projection of pixels in the yellow box (lip ROI) and the two green lines are in the cut points. The localization result is represented by a red rectangle.

2.4. Lip localization by Spatial Deviation Analysis (SDA)

When processing face images with clean background, a binary image with only one big white connected component representing the lip can be acquired using above classifier, thus to locate the lip by Connected Component Analysis (CCA) is straightforward. However, in the presence of background clutters in the lip ROI, the binary image would have big erroneous blobs and it is difficult to get the lip localization through CCA. Here, we propose a novel projection-cut algorithm by Spatial Deviation Analysis (SDA) to solve the problem.

We utilize an appearance descriptor - *projection vector* and a spatial deviation descriptor - *variance vector* to model the lip localization problem. The projection vector is defined as the number of white pixels in each row (column) and the variance vector is defined as the standard variance of the positions of white pixels in each row (column).

We formulate the *y*-projection vector as an example. The *x*-projection vector is similar. Let the width and the height of the ROI be represented as *w* and *h* respectively, and the *y*-projection vector is represented as $\{p_i\}_{i=0}^{h-1}$. Let $\{v_i\}_{i=0}^{h-1}$ denote the standard variance of the positions of white pixels in the *i-th* row. Thus we can define $\{p_i\}_{i=0}^{d-1}$ as follows:

$$p_i = \begin{cases} \frac{1}{255}\sum_{j=0}^{w-1} I(i,j) & t_1 < v_i < t_2 \\ 0 & others \end{cases}$$

where

$$v_i = \frac{1}{255}\sum_{j=0}^{w-1}\left[\left(j - m_i \right)^2 \times I(i,j) \right]$$

$$m_i = \sum_{j=0}^{w-1} \left[j \times I(i,j) \right]$$

$I(i,j)$ is the intensity of the pixel (i,j). t_1 and t_2 are the thresholds of v_i which we use in our study is 0.2*w and 0.8*w respectively.

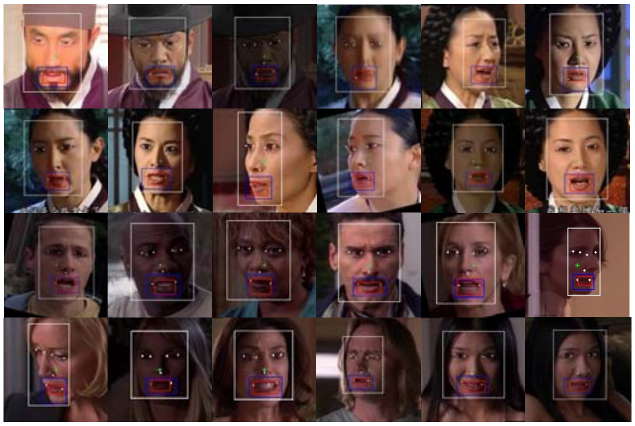

Figure 4: Example results in videos with complicated condition. (Top two rows) The result images of Dae. (Bottom two rows) The result images of Housewives.

After acquiring the projection vector, we search one local minimum from each end of the projection vector respectively. Then the part between these two minima is kept and others are removed. After that the lip ROI is updated. The algorithm iterates until no new cuts. It is efficient and converges in 2~4 iteration steps typically. The algorithm is demonstrated in Fig.3.

3. EXPERIMENTAL RESULTS

The test is carried out on 2 teleplay videos, including *episode 31 of Dae Jang Geum* and *episode 21 of Desperate Housewives season 2*. In the following we use *Dae* and *Housewives* for short to denote the two teleplay videos, respectively.

We sample result images randomly for the convenience of statistical analysis (as shown in Table 1). To justify if the detected lip position is correct we annotate the lip locations by hand, allow two pixels error to the center of the lip bounding box and four pixels error to the width and the height of the box. Table 1 shows the face number we get from the face detector and statistical results of lip localization. Error happens always in the images that the left and right lighting condition of the lip is changed intensively. If we consider the continuity of videos, the error would decrease dramatically. Some examples of results are shown in Fig.4.

	Left-profile	Frontal	Right-profile
Dae	6218(19%)	18029(55%)	8766(26%)
	6/124(4.8%)	9/360(2.5%)	9/175(5.1%)
Housewives	5469(17%)	19475(61%)	7174(22%)
	8/109(7.3%)	18/389(4.6%)	11/143(7.7%)

Table 1: (First row of each entry) The face number we get from the face detector and its percentage of the total face number in the bracket. (Second row of each entry) Statistical results of lip localization: the ratio of incorrect located image number to sample images and its percentage in the bracket.

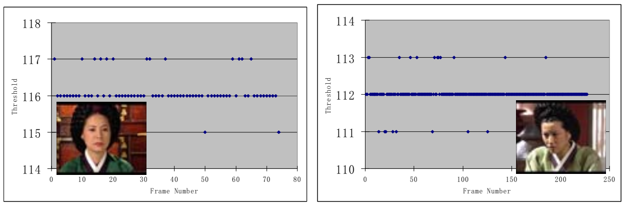

Figure 5: The stability of the color threshold in a shot and the diversity between shots. Y-axis: threshold. X-axis: frame number. Sample images are attached to each chart.

Figure 5 shows the stability of the color threshold in a shot and the diversity between shots. We can see that the threshold is almost the same in a shot and our classifier can handle different lighting condition. The result also justifies that the classifier using the anthropometrical properties of the human face is effective and robust.

4. CONCLUSION

Lip localization on multi-view faces in video is a difficult problem due to the complicated condition in the video and the weak color contrast between the lip and the face region. In this paper, the lip localization problem is formulated as a binary classification problem. The proposed classifier is able to exploit the anthropometrical properties of the human face with the help of face scale normalization. It can distinguish the lip from its surroundings effectively and we can get accurate lip localization through SDA. The proposed method is tested on two teleplay videos and the experimental results demonstrate that the approach is efficient and robust for both frontal faces and profile faces in different lighting condition.

5. ACKNOWLEDGEMENTS

This work is partially supported by the National Hi-Tech Research and Development Program of China (863 Program, Grant No. 2006AA01Z315), and the National Natural Science Foundation of China (Grant No. 60121302).

6. REFERENCES

[1] X. Zhang, R.M. Mersereau, “Lip Feature Extraction Towards an Automatic Speechreading System,” in *ICIP2000*.

[2] N. Eveno, A. Caplier, P.Y. Coulon, “A Parametric Model for Realistic Lip Segmentation,” in *International Conference on Control, Automation, Robotics and Vision*, *ICARCV2002*.

[3] A.W.C. Liew, S.H. Leung, W.H. Lau, “Segmentation of Color Lip Images by Spatial Fuzzy Clustering,” *IEEE Trans. on Fuzzy Systems*, vol. 11, no.4, pp. 542-549, 2003

[4] C. Huang, H.Z. Ai, Y. Li, S.H. Lao, “Vector Boosting for Rotation Invariant Multi-View Face Detection,” in *ICCV 2005*.

[5] L. Zhang, H.Z. AI, S.H. Lao, “Robust Face Alignment Based on Local Texture Classifiers,” in *ICIP2005*.