A Robust Lip Tracking System for the Acoustic to Articulatory Inversion

Jingying Chen, Yves Laprie, Marie-Odile Berger

A Robust Lip Tracking System for the Acoustic to Articulatory Inversion

Jingying Chen *Yvs Laprie* *Mario-Odile Berger*

Lorraine Laboratory for Research into Information Technology and its Applications (LORIA), INRIA-Lorrain, 615 rue du jardin botanique, 54600 Villers-lès-Nancy, France
Chenj@loria.fr

Abstract

The acoustic to articulatory inversion of speech which refers to the mapping from the acoustic signal to the articulatory, is an interesting problem. Given the acoustic signal, the recovery of the articulatory state is considered difficult. The reason is the "one-to-many" nature of the acoustic-to-articulatory inversion problem: a given articulatory state has always only one acoustic realization but an acoustic signal can be the outcome of more than one articulatory states. In order to solve the "one-to-many" problem of the inversion, visual information complementary to acoustic signal is used. Hence, a robust lip tracking system to provide visual information (such as the width and height of mouth) for the acoustic-to-articulatory inversion is developed in this paper. The proposed approach uses a combination of motion, color and structure information of the mouth area to track lip feature points. This technique is designed to be effective and robust. It has the advantages to detect the lip feature points automatically and recover the feature points lost during tracking process. Encouraging results have been obtained using the proposed approach.

1. Introduction

The acoustic to articulatory inversion of speech poses an interesting problem which has attracted the interest of researchers worldwide during the last three decades. The concern of the problem is the mapping from the acoustic signal to the articulatory. Given the acoustic signal, the recovery of the articulatory state is considered difficult. The reason is the "one-to-many" nature of the acoustic-to-articulatory inversion problem: a given articulatory state has always only one acoustic realization but an acoustic signal can be the outcome of more than one articulatory states. Despite the difficulties in the inversion, the possible applications are promising. The most interesting application is the use of the additional articulatory information derived from the inversion to improve the performance of current speech recognition systems, especially in cases such as with noisy, spontaneous or pathological speech. Other possible applications include speech synthesis, building visual aids for teaching hearing impaired people how to speak and as a means of study in phonetics and phonology [1].

In order to solve the "one-to-many" problem of the inversion, visual information complementary to acoustic signal is used in this study. Since lip features play an important role in the inversion, lip tracking is proposed to provide visual information of lip movement for the inversion.

Works on lip tracking range from purely image-based approaches [2] to sophisticated model-based approaches, e.g. active contour models (or snakes [3]) and active shape models [4]. Each of these approaches has its own strengths and limitations. The lip tracking approach using a single cue about the image sequence is insufficient for reliable tracking. For example, the active contour method often converges to the wrong result when lip edges are not obvious or when lip color is very close to the face color. Also, the method is quite computationally expensive as they require many iterations to fit the lip contour properly. The active shape model method needs a large set of training data to learn patterns of typical lip deformation. Hence, it is believed that a reliable and effective tracking system should use as much knowledge about the image sequence as possible to handle all the sources of variability in the environment. Kaucic *et al* [5] tracked lip contour using a combination of B-splines and Kalman filters. Petajan [6] imposed anatomical constraints to implement an inner contour tracker using color thresholds and teeth templates. Rao [7] developed a color-based deformable template method that combining shape and color information, however, it fails when there is a shadow area near the lip or the lip color is similar to that of the face. A lip

tracking method combining color, shape and motion, was proposed by Tian *et al* [8]. In their method, three mouth states are introduced: open, relatively closed and tightly closed. First, the lip template is manually located in the first frame, and lip color information is modeled as a Gaussian mixture. Then, the key points of the lip template are automatically tracked using Lucas-Kanade tracker in the image sequence. The mouth states are determined by the color and shape information. Encouraging results were reported in their work. However, the tracking failure may still happen caused by sudden head orientation or variation of lighting conditions during tracking process. In order to develop a robust lip tracking system to provide visual information (such as the width and height of mouth) for the acoustic-to-articulatory inversion, a new approach using a combination of motion, color and structure information of the mouth area is proposed in this paper. This technique is designed to be effective and robust. It has the advantages to detect the lip feature points automatically and recover the feature points lost during tracking process. Encouraging results have been obtained using the proposed approach. This suggests a strong potential for lip tracking system.

The outline of the paper is as follows. A system overview is described in Section 2. Lip feature points detection and tracking are presented in Section 3 and 4, respectively. Section 5 describes the results while Section 6 presents the conclusions and directions for future work.

2. System overview

The proposed lip tracking system comprises two modules, i.e. lip feature points detection and tracking. The lip feature points that we are interested in are the two lip corners and the mid-points of upper and lower lips (see Figure 1). Since the purpose of the lip tracking algorithm is to help the acoustic-to-articulatory inversion, we are interested in the vocal tract shape so that the inner lip contour is used here.

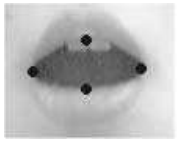


Figure 1: Lip feature points on the inner lip contour.

The lip tracking system is illustrated in Figure 2. The feature points detection operates on the first frame of the image sequence using the color and structure information of the mouth area, and it is assumed that the mouth is horizontal in the first frame. Details are given in the following section. Then, the feature points tracking is performed to follow the detected lip feature points. The proposed lip feature points tracking algorithm is based on the Lucas-Kanade (LK) tracker [9], which is invariant to mouth orientation. The proposed tracker also incorporates color and structure information of the mouth area to recover the lost points which might be caused by sudden mouth opening, head orientation or variation of lighting conditions during tracking process.

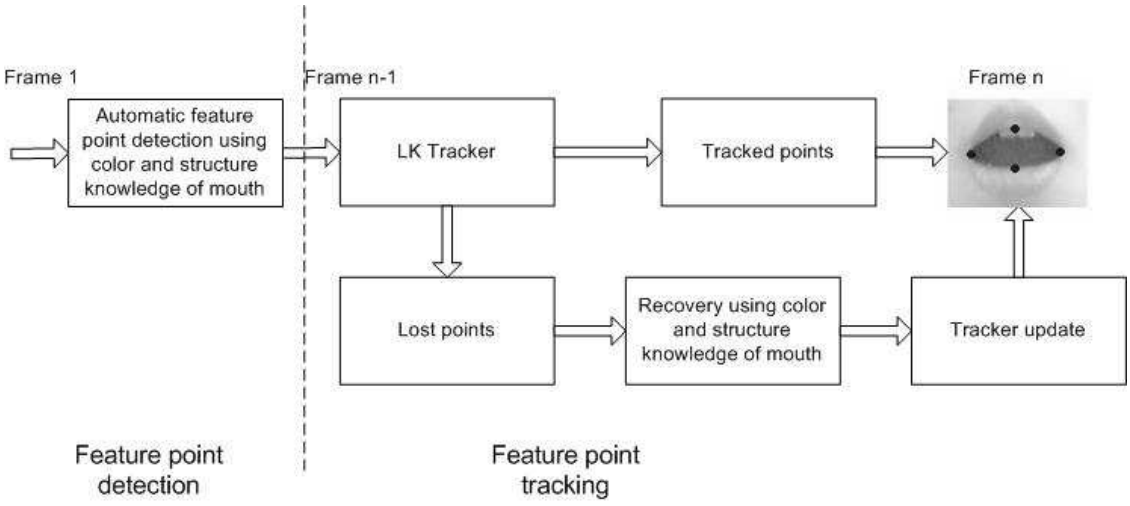


Figure 2: The lip tracking system.

3. Lip feature points detection

Lip localization

Before lip feature points detection, the approximate location of the speaker's lips is estimated using color information. First, the image captured from camera is transformed from RGB color space to HSI color space which separates hue (H) and saturation (S) from intensity (I). Then, the hue value is used to calculate the candidate lip pixel because hue is relatively insensitive to the lighting variations. Finally, connected components consisting of pixels with hue value that lies within the range of value typical of the lips are formed, which is identified as the region of interest (ROI) in this study (see Figure 3).

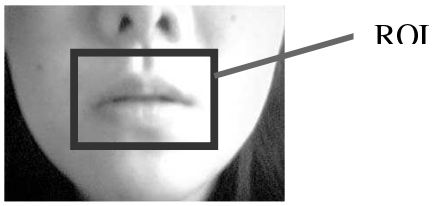


Figure 3: Lip localization.

Lip feature points detection

The lip feature points detection includes two steps: (1) find the lip corners and (2) find the mid-point of the upper and lower lips.

Step 1: First, a horizontal integral projection [10, 11] is applied on the intensity image in the ROI to find the vertical position of the shadow line between the lips. The shadow line is the darkest horizontally extended structure in the ROI, its vertical position can be found where the horizontal integral projection value P_h is the global minimum (see figure 4). The position can be considered as the approximate vertical position of the lip corners.

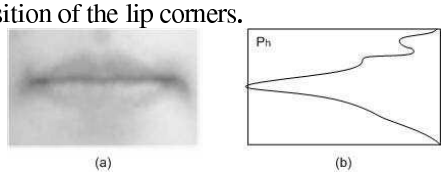


Figure 4: Finding the vertical position of the shadow line between lips using the horizontal integral projection in the ROI, (a) ROI and (b) horizontal integral projection value P_h plotted against corresponding rows.

Second, a vertical integral projection is performed on the horizontal edge map in the ROI. The horizontal edge map can be obtained by applying the Sobel horizontal edge detector. The horizontal positions of the lip corners are estimated by examining the vertical integral projection values P_v, the locations where the values exceed or fall below a certain predefined threshold are considered as the estimated horizontal positions of the lip corners (see figure 5). Then, the vertical position of the lip corners is adjusted by finding the darkest pixel along their horizontal position.

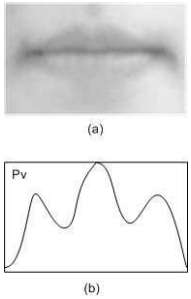


Figure 5: Finding the horizontal position of the lip corners using the vertical integral projection in the ROI, (a) ROI and (b) vertical integral projection value P_v plotted against corresponding columns.

Finally, the positions of the lip corners are refined. We search for the darkest pixels with maximum contrast around the positions (found above) of the left and right corners in the two small search windows (see figure 6).

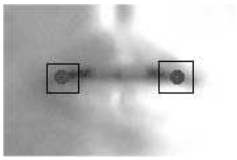


Figure 6: Refining the lip corners in the two search windows.

Step 2: The horizontal position of the mid-points of the upper and lower lips is computed as the middle between the left and right lip corners. For the vertical position of the mid-points, two states are considered: closed mouth and open mouth. If the mouth is closed, the mid-points should lie in the shadow line. If the mouth is open, the mid-points should lie either between the teeth and lip

flesh or between the oral cavity and lip flesh. Teeth can be separated easily from other parts of the mouth because they have low saturation and high intensity, and oral cavity has low intensity. Hence, the vertical position of the mid-points can be found using the structure information of the mouth and their color characteristics. Examples are given below (see Figure 7).

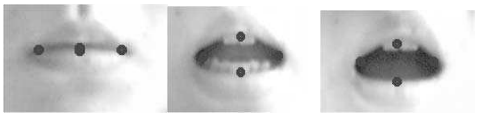


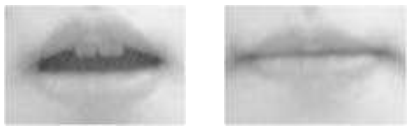
Figure 7: The lip feature points of the different mouth states.

4. Lip feature points tracking

After feature points detection, the Lucas-Kanade (LK) algorithm [9] is performed to track the feature points. The algorithm detects the motion through the utilization of optical flow. Mase and Pentland [12] developed their lip reading system using optical flow analysis. Optical flow is defined as the distribution of apparent velocities in the movement of brightness patterns in an image. Here, it is assumed that the intensity values of the neighborhoods of the feature points being tracked do not change, but have only translation movement from image I_t at time t to the next image $I_{(t+1)}$ at time $t+1$. $I_t(\mathrm{x})$ and $I_{t+1}(\mathrm{x})$ are the intensity values of the image I_t and $I_{(t+1)}$ at the location $\mathrm{x} = [x, y]^T$, where x and y are the two pixel coordinates of an image point. Consider an image point $\mathrm{a} = [a_x, a_y]^T$ on the image I_t. The goal of feature point tracking is to find the location $\mathrm{b} = \mathrm{a} + \mathrm{d} = [a_x + d_x, a_y + d_y]^T$ on the image $I_{(t+1)}$ such as $I_t(\mathrm{a})$ and $I_{t+1}(\mathrm{b})$ are "similar". The **d** is the displacement vector chosen to minimize the residue factor ε. In order to solve the *aperture problem*, the residue factor ε computation is based on neighborhood R, around the point **x,** as follows:

$$\varepsilon(\mathrm{d}) = \sum_{R} (I_t(\mathrm{x}) - I_{t+1}(\mathrm{x} + \mathrm{d}))^2 \quad (1)$$

Resolution of equation (1) is detailed in [13]. In order to track large motions (such as sudden head movements and mouth opening) without losing sub-pixel accuracy, a pyramidal implementation of the LK tracking method is used. In this study, each image is decomposed into 4 levels, from level 0 (the original finest resolution image) to level 3 (the coarse resolution image), a 5x5 Gaussian filter is employed to smooth out the noise and 11x11 neighborhood is used for all levels. The LK tracker can track the feature points properly under most conditions, however, when the neighborhood around tracked point varies too much between image I_t and $I_{(t+1)}$, the point might be lost. For example, if the lip movements are fast (see Figure 8), the mouth changes from open state with teeth appearance to closed state, the mid-points of the upper and lower lips can be difficult to track. Hence, the recovery of lost points is necessary. Similar to the feature points detection, the recovery includes two steps: (1) recover the lip corners and (2) recover the mid-point of the upper and lower lips.



(a) Open state (frame n-1) (b) Closed state (frame n)

Figure 8: Mouth state transition.

Step 1: Instead of using the integral projection to estimate the initial position of the lip corners, we search for the darkest pixels with maximum contrast around the positions in the previous frame just before the points lost, in the two search windows (see figure 6). The search window size is chosen heuristically to be big enough to include the possible new positions.

Step 2: Along the direction perpendicular to the line connecting the two lip corners, we search for the new position of the point in a search window centered on the position in the previous frame just before the points lost, using the color and structure information of the mouth as described in Step 2 of Section 3.

5. Experiment results

Experiments were carried out using a set of color image sequences of lip movement under natural lighting condition. The length of sequences ranges from 30 to 60 frames. The image size is 320x240 pixels at the frame rate of 30 frames per

second. Some results using the proposed tracking algorithm are shown in Figure 8. Three short typical image sequences of lip movement are given below: (i) open mouth with upper teeth appearance, (ii) the transition from open mouth to closed mouth and then to open mouth with upper teeth appearance, (iii) the transition from open mouth with both upper and lower teeth appearance to closed mouth. From the results, one can see that the four lip feature points have been tracked correctly under different mouth states. Table 1 gives an estimate of the tracking accuracy of the proposed tracking algorithm. The measurements were taken using 60 frames. Manually determined position of the lip feature points were used as reference for measuring displacement error. The error was computed as the distance between the reference points and the points obtained by the proposed method in x and y directions. It can be seen that the tracking is quite accurate.

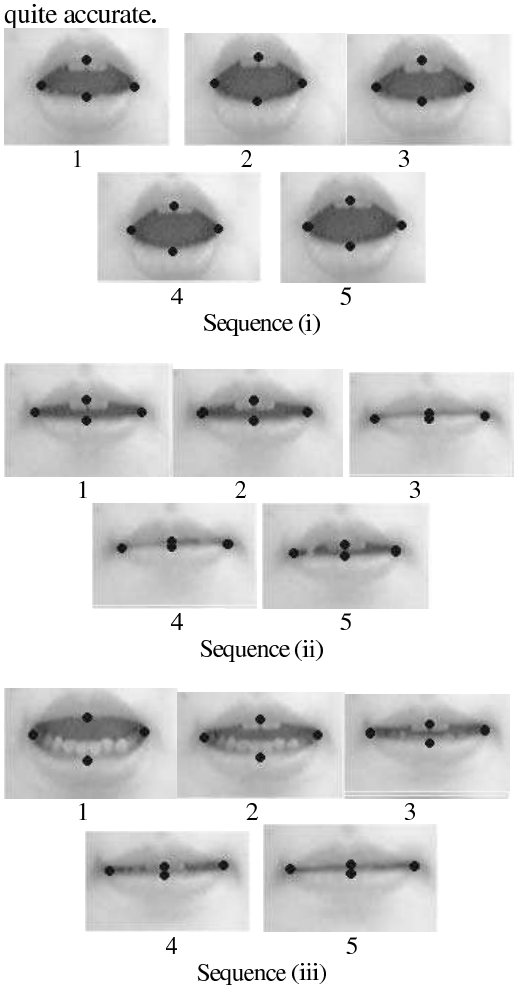


Figure 8: The results of lip feature points tracking under different mouth states.

	Error x	Error y
Lip corners	2.9	1.5
Middle points of upper and lower lips	0.7	2.7

Table 1: Average feature point displacement errors in x and y directions in pixel

The comparison has been made between the original LK tracker and the proposed tracking algorithm that incorporates the recovery process into the LK tracker. Using the original LK tracker, we found that it could track properly for sequence (i). However, during sequence (ii), the middle points of upper and lower lips were lost when the mouth was closed. For sequence (iii), the two lip corners drifted away when the lower teeth disappeared as shown from the second to third frame, also, the middle points of upper and lower lips were lost when the mouth was closed. While using the proposed tracking algorithm, the feature points are tracked correctly. This suggests a strong potential for lip tracking system.

6. Conclusions

A robust lip tracking system to provide visual information (i.e. the width and height of mouth) for the acoustic-to-articulatory inversion is proposed in this paper. The integral projection method alone is not orientation invariant while the LK tracker alone cannot recover the lost points during tracking process, hence, an approach using a combination of motion, color and structure information of the mouth area is proposed to track lip features points. It has the advantages to detect the feature points automatically and recover the feature points lost during tracking process. Encouraging results have shown that the proposed method outperforms the original KL tracker in terms of recovering lost points. Currently, this study is based on 2D images, we are investigating to apply it on stereovision system to provide high accurate 3D mouth position (such as the protrusion of the upper and lower lips) for the inversion.

Reference:

[1] A. Toutios, K. Margaritis: "Acoustic-to-articulatory inversion of speech: a review", in Proceedings of the International 12th Turkish Symposium on Artificial Intelligence and Neural

Networks (TAINN-2003), Canakkale, Turkey, July 2003.

[2] J. Yang, R. Stiefelhagen, U. Meier and A. Waibel, "Real time face and facial feature tracking and applications", in Proceedings of the International Conference on Auditory-Visual Speech Processing AVSP'98, pp. 207-212, 1998.

[3] M. Kass, A. Witkin and D. Terzopoulus, "Snake: Active contour models", International Journal of Computer Vision, 1(4), pp. 1435-1444, 1992.

[4] J. Leuttin, N.A. Tracker and S.W. Beer, "Active shape models for visual speech feature extraction", Electronic System Group Report No. 95/94, University of Sheffield, UK, 1995.

[5] R. Kaucic, B. Dalton and A. Blake, "Real time lip tracking for audio-visual speech recognition application", in Proceedings of ECCV, Cambridge, UK, pp. 376-387, 1996.

[6] E. Petajan and H. Graf, "Robust face feature analysis for automatic speechreading and character animation", Speechreading by Man and Machine, pp. 425-436, 1996.

[7] R. R. Rao, Audio-Visual Interaction in Multimedia, PHD Thesis, Electrical Engineering, Georgia Institute of Technology, 1998.

[8] Y. Tian, T. Kanade and J. Cohn, "Robust lip tracking by combining shape, color and motion", in Proceedings of the 4th Asian Conference on Computer Vision, January, 2000.

[9] B. Lucas and T. Kanade, "An interactive image registration technique with an application in stereovision", in Proceedings of the 7th International Joint Conference on Artificial Intelligence, pp. 674-679, 1981.

[10] T. Kanade, Picture processing by computer complex and recognition of human faces. Technical report, Kyoto University, 1973.

[11] R. Brunelli and T. Poggio. "Face recognition: Features versus templates", IEEE Trans. PAMI, 15(10), pp. 1042-- 1052, 1993.

[12] K. Mase and A. Pentland, "Automatic lipreading by optical-flow analysis", Systems and Computer in Japan 22(6), pp. 67-76. 1991.

[13]http://www.comp.nus.edu.sg/~cs4243/lab/opencv.pdf