Real Time Tracking for 3D Realistic Lip Animation

Brice Beaumesnil, Franck Luthon

Real Time Tracking for 3D Realistic Lip Animation

Brice Beaumesnil and Franck Luthon
IUT Informatique de Bayonne
LiUPPA
Château Neuf, Place Paul Bert 64100 Bayonne, France
beaumesn@iutbayonne.univ-pau.fr, franck.luthon@univ-pau.fr

Abstract

This article deals with facial segmentation and liptracking with feedback control for real-time animation of a synthetic 3D face model. Classical approaches consist in two successive steps : video analysis then synthesis. We want to build a global analysis/synthesis processing loop, where the image analysis needs the 3D synthesis and conversely. For that, we fit a generic 3D-face model on the speaker's face in our analysis algorithm for using synthesis information (like 3D information or face shape). This approach is inspired from control systems theory with feedback loops.

The contribution of the paper is to use simple image processing techniques on available data, but to improve segmentation through the feedback loop. Moreover, we propose a robust lip corners tracking based on estimation motion algorithm. The speaker is only asked to be in front of the camera with the mouth closed at the beginning of the video session (neutral position). This allows to do a quick initialisation step in order to fit the 3D-face model. Results show that real-time (30Hz) and robust performances are achievable under real-world conditions, which are two key issues for face and lip tracking applications.

1. Introduction

We present a complete real-time analysis/synthesis framework allowing lip tracking for animation of a clone with a single camera in unconstrained environment (typ. webcam in the office). The approach is based on lip segmentation from a hue component computed within a non-linear color space that is robust to luminosity variations. A robust lip corner detection helps to initialize two active contours that converge on the lip borders. Then inner and outer borders of the lip are extracted, and interpreted to make a real-time realistic animation of a clone's mouth.

More than 80% of visual information during a conversation between two people is due to lip motion. Those visual clues are essential for a better understanding of the speech [4]. For synthetic talking head animation, realistic rendering of lip motion is the key point. The purpose of this paper is to link in real-time the face image analysis with the synthesis of an animated 3D model of the head. We focus on the speaker's lip video segmentation from a mono-camera (webcam or motorized camera) for on-line animation of the mouth of a clone, without dealing here with the sound information (no speech processing). In future developments of course, speech processing should be coupled with image processing for optimal audiovisual rendering. But here, we want to show what one can do with the video only. Our aim is to get a realistic rendering of the mouth movements, but not necessarily to get the most precise lip contour extraction. It means that image analysis needs not to be very sophisticated, but just good enough for our application (i.e. get an acceptable rendering for a realistic animation). We propose to compensate for some defaults in the analysis by implementing a feedback loop form the 3D synthesis towards the input segmentation process.

2. Description of the Processing Loop

Our framework is divided into five parts :

1. low-level color segmentation (analysis) : it works in the LUX [6] space that is little sensitive to lighting variations and exhibits very distinctly skin and lip hue areas.
2. active contour positioning (estimation) : to delineate both inner and outer lip contours.
3. 3D model regularization (synthesis and feedback loop) : error measurement and active contour repositioning (in loop with step 2).
4. transmission of geometrical parameters (communication) : via Internet to a distant computer.

0-7695-2521-0/06/$20.00 (c) 2006 IEEE

5. clone animation (interpretation/synthesis) : it uses the lip contours extracted and adjusted at step 2 and 3 and transmitted at step 4.

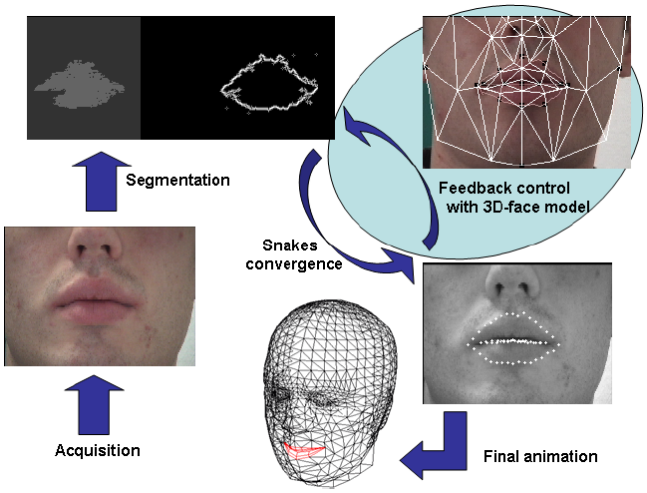

FIG. **1.** Complete real-time analysis/synthesis framework.

Our work assumes that we have at our disposal an automatic tool for face detection and tracking, so that an optimal framing of the speaker's face is at our disposal (which is required in the case of videoconferencing for example). If this is not the case, the only constraint is that the speaker should stay in front of the camera with little motion (normal behavior in front of a webcam). In this paper, we simply ask the speaker to seat directly in front of the camera so that his face covers the major part of the image. Moreover, we ask him to have the mouth closed (neutral position) on the first frame of the sequence, in order to make a good initialisation of his mouth proportions (3D modelisation).

Lighting conditions are not calibrated nor constant : they correspond to realistic office environment (non uniform lighting, light sources that may be added or suppressed depending on the daylight).

2.1 Extraction of Lip Area

For face and lip segmentation, we use the LUX color space defined in [6]. This color space is non linear with respect to the RGB color components. It helps to reinforce the color contrast while being relatively insensitive to lighting variations. In this color space, most of the color information relative to a human face is coded by the U component (red chromaticity) in the particular case when $R > L$ (L being the luminance Eq.1). Therefore, we derive from LUX space the simplified hue component U, Eq.2, that is more discriminating than RGB or HSL for face and lip segmentation.

$$L = (R+1)^{0.3}(G+1)^{0.6}(B+1)^{0.1} - 1 \qquad (1)$$

$$U = \begin{cases} 256\frac{L+1}{R+1} & \text{if } R > L, \\ 255 & \text{otherwise or if } L < \lambda. \end{cases} \qquad (2)$$

The threshold λ is introduced in order to detect very dark areas like the inner side of the mouth or nostrils. This will amplify the various gradients computed on the hue.

Since the hue difference between face and lips is more contrasted in this space, we can easily classify pixels as lips or face. For that purpose we use the k-means classification algorithm [7] with three hue classes and a lip detection based on a non linear filter (Eq.3).

$$M(i) = \begin{cases} 1 + \sum_{j \in \nu(i)} a(j) M(j) & \text{if } U(i) \in \text{“}lips\text{”} \\ 0 & \text{otherwise} \end{cases} \qquad (3)$$

Where M is a matrix of size $L \times C$ (image size) initialised to 0 and the $a(j)$ (with $j \in v(i)$) are the coefficients attributed to the different pixels belonging to the causal and connex neighborhood $v(i)$ of the current pixel i.

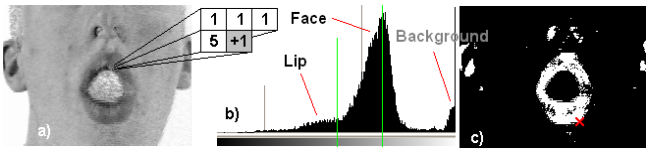

FIG. **2.** (a) Hue computed from Eq.2 with nonlinear filter mask ; (b) Hue Histogram ; (c) Lip classe in white and P (the maximum value of M) in red

This simple technique gives an approximate bounding box (BB) for mouth positioning in the image plane (obtained by a simple scan of the connected component lip-hue area that includes P : point giving the maximum value of M (see Eq.3)). However, it may fail when lip corners are not well detected or when lower and upper lips form two separated areas (this may happen e.g. in case of the French phoneme [a]). To cope with those situations, a lip tracker is added : it is based on Lucas-Kanade algorithm applied on a few relevant points detected on the outer lip contour. Since the speaker is supposed to have his mouth closed at the first image, we are guaranteed to have a single lip area at the beginning, so that the outer lip contours can be tracked properly afterwards, even if the two lips are no more connected areas in the video sequence. Using this simple motion estimation technique instead of motion detection as in [6] yields a better robustness to lighting variations.

2.2 Lip corners

For initializing the active contours on the lips outer border, we need lip corners position on all images. To have it, we use two methods :

1. On the first image : we use L composant (luminance) and find on each column of our bounding box BB the darkest pixel. We obtain a curve that passes between lower and upper lips and on the lip corners. We adjust the result to have lip corners on the same line (we ask to the speaker to seat directly in front of the camera) see Fig.3(a).
2. Otherwise : we use the same result and combine it with a lip tracking by Lucas-Kanade algorithm. The tracking helps to find a reduced search zone for lip corners position, and the curve helps to place precisely lip corners in this zone (see Fig.3(b)).

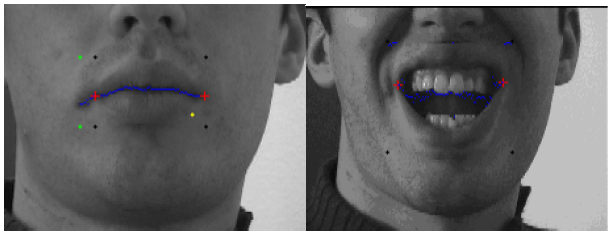

FIG. **3.** (a) Lip corners (in red) on the first image of the sequence ; (b) Lip corners with Lucas-Kanade estimation (in black : BB corners, in blue : per column darkest pixel)

This method can reduce the BB defined by the lip map, if lip corners are not on the lip map extremity (see Fig.3(a)).

2.3 Active Contours and Clone Animation

For estimating the lips outer border, one active contour (or snake [5]) is initialised with the BB detected as explained above.

This snake is made of a finite number of control points (it depends of the BB width) that are forced to undergo only vertical displacements during iterations. The points are initialised on cubic curves computed from the BB and from the lip corners position.

The forces used for snake convergence are the following :

- an internal force that controls elasticity and curvature (very classically defined)
- an external force that specifies the features that should attract the snake (namely spatial gradient computed both on hue and on luminance maps)
- a constraint force that is specific of the problem at hand (the snake is forced to converge towards the gravity center of the BB)

After convergence of the outer snake, we initialize a 3D model of the mouth (and all the face) with the control points of the snake using the method explained in [3].

We use the 3D-face model named CANDIDE-3 [1] in order to model the speaker's head. On the first image of the sequence, the 3D-face model is fitted on the speaker's face, and for the rest of the sequence this "3D-face fitted model" is deformed by its animation parameters.

The 3D-face model is extracted and then helps to limit the face deformations degrees and the mouth deformation degrees. Thus, the 3D-face model allows to constrain the previous segmentation solution (obtained thanks to active contours) and then to keep the mouth position in an acceptable solution space.

Thanks to the "3D-face fitted model", we can initialize another snake (inner one) to detect precisely inner contour. The points of the snake are initialised on cubic curves computed from "3D-face fitted model" and from the lip corners position. The constraint force changes to force points stay near 3D inner contour.

After convergence of this second snake towards the inner lip contour, the interpretation step for our application purpose is readily done : we are able to compute various geometrical features from the control points of the two snakes and our "3D-face fitted model", that are then used as input parameters for the 3D head model. This computation of animation parameters is of course dependent of the 3D model used. The talking head used is a 3D clone, it is built with 275 mobile points that allow realistic synthesis of mouth motion thanks to six parameters (dedicated to visemes and phonemes of the French language) [2]. In our case, a simple system of linear equations transforms the MPEG4 points coordinates of our 3D-face model into animation parameters that are taken as input by the talking head.

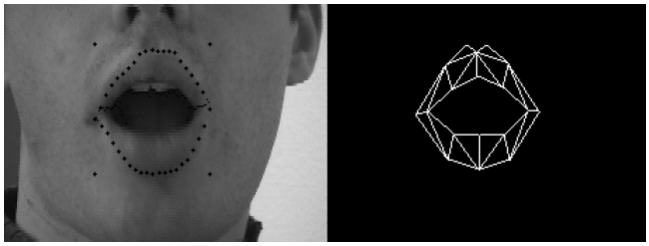

FIG. **4.** (a) After convergence outer snake ; (b) lip of the "3D-face fitted model"

3 Experimental Results

The keypoint is that even if lip contours are not very well estimated (poor initial segmentation), it is not redhibitory for a realistic animation. Indeed the comparison between the estimated snake locations and the lip of the "3D-face fitted model" positions allows to make a backward correction (through the feedback loop).

As regards the inner lip contour, teeth are not a problem : they even amplify the gradients, ensuring a better convergence (see Fig.5). The proposed feedback processing solve

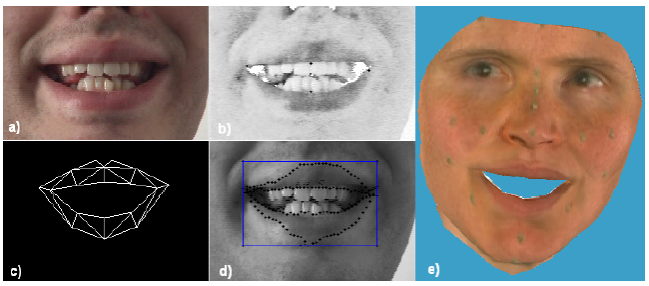

FIG. **5.** (a) Color image ; (b) Hue computed from Eq.2 with left and right lip corners, central upper and lower lip positions in black ; (c) lip of the "3D-face fitted model" ; (d) Outer and inner snakes after convergence with bounding box (in blue) ; (e) Animated clone

TAB. **1. Algorithm CPU time**

algorithm features	CPU pourcent for the first image	CPU pourcent for the rest of the sequence
color transform	28%	49%
lip detection	1%	
liptracking		1%
clustering	57%	25%
active contours and clone animation	14%	25%

problem of the tongue appearing (that has almost the same hue as lips) and the presence of reflected highlights. Indeed, its force the lip to have the same form during all the sequence. For exemple, if we do not have any gradients (due to the tongue appearing) the "3D-face fitted model" inner lip contours and the inner snake contours are very different (error distance control). In this case, we use our "3D-face fitted model" to animate the clone.

As a matter of fact, outer lip contours are very well estimated when the mouth is wide open. On the opposite, inner lip contours are well estimated when the mouth is closed or in the case of protrusion. We may use this fact in conjunction with the 3D result to guide the processing of re-computation where the segmentation initially failed.

Table 1 represents CPU needs during a test sequence (computed on the sequence used in Fig.5 where the image size is 230×172 pixels and lip width is 110 pixels). On an $i386$ processor at $1.4GHz$, processing rate is better than $30Hz$. The difference between the first image and the rest of the sequence is due to the k-means algorithm : it is computed on whole the image on the first image to find the lip position. For all the sequence just the color transform is executed on whole the image, other processings are execute just in our bounding box.

4 Discussion and Future Work

As a conclusion, we have shown that with a few points estimated by a rapid segmentation scheme and by the use of a feedback loop to correct some segmentation defaults, one can achieve real-time realistic animation of a synthetic talking head. Currently only vowels were tested carefully. The clone reproduces the speaker's lip motion with a very nice precision.

The whole analysis algorithm, implemented in non optimised C-code on an $i386$ processor at $1.4GHz$, works in real-time (i.e processing rate better than $30Hz$).

Another direction of our current research is to use all 3D face information (like the depth) and exploit other face features (namely nostrils, eyes, eyebrows and ears). This will help us to get other relevant points for 3D model scaling (cf. user-dependent facial geometry taken into account at initialisation step) and to take account of 3D face position.

Thus, having more information on the whole face may enable a better understanding of some spoken phonemes, and the 3D-pose of the model can be use to have a better lip animation (if the face are not vis-a-vis the camera, we will not have any deformation).

Finally, as we want to be able to animate various clones and propose a generic solution, we are also working on the use of MPEG4-compliant 3D models (using FDP and FAP, facial animation parameters).

Références

[1] J. Ahlberg. Real-time facial feature tracking using an active model with fast image warping. In *International Workshop on Very Low Bitrate Video, VLBV*, pages 39–43, Oct. 2001.

[2] C. Benoît, T. Lallouache, T. Mohamadi, and C. Abry. A set of french visemes for visual speech synthesis. *Talking Machines : Theories, Models, and Designs*, pages 485–503, 1992.

[3] M. Chaumont and B. Beaumesnil. Robust and real-time 3d-face model extraction. In *IEEE International Conference on Image Processing, ICIP'2005*, pages 461–464, Sept. 2005.

[4] T. Chen and R. Rao. Audio-visua integration in multimodal communication. In *Processings of the IEEE*, volume 86, pages 837–852, May 1998.

[5] M. Kass, A. Witkin, and D. Terzopoulos. Snake : Active contour models. *International Journal of Computer Vision*, 1 :321–331, 1987.

[6] M. Lievin and F. Luthon. Nonlinear color space and spatio-temporal mrf for hierarchical segmentation of face features in video. *IEEE Trans. on Image Processing*, 13 :63–71, Jan. 2004.

[7] J. Macqueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, Berkeley, 1967.