EXTRACTION OF TONGUE CONTOURS IN X-RAY IMAGES WITH MINIMAL USER INTERACTION

Yves Laprie and Marie-Odile Berger

CRIN-CNRS & INRIA-Lorraine
BP 239 54506 Vandœuvre-lès-Nancy FRANCE
email:{laprie,berger}@loria.fr

ABSTRACT

In spite of the development of new imaging techniques, X-ray images still keep a prominent place to studying articulatory phenomena. Indeed, they are still unsurpassed to obtain an overall view of the moving vocal tract. However X-ray images require that articulators contours are extracted by hand which is a tedious task. This paper describes an approach towards the automatization of the tongue contour extraction. The "Snake" method introduced in computer vision to extract contours is unable alone to achieve the task. Therefore we make "Snakes" cooperate with an optical flow method applied where contours are not sufficiently isolated from spurious contours. Our experiments have shown that the tongue is tracked successfully when it is visible, and that interaction with the user remains necessary when the tongue is obscured.

1. INTRODUCTION

1.1. Specificities of X-ray vocal tract images

Studying articulatory phenomena as well as developing new articulatory synthesis approaches require dynamical measures of the vocal tract. Although modern imaging techniques give better images like MRI, or involve less bio-hazard like EMA, X-ray imaging remains the most appropriate tool to studying speech articulators because it allows shooting at a fast rate with a sufficient resolution to recover the shape of articulators. However, before being exploited these images require that contours are extracted by hand which is a tedious task. This is all the more difficult because X-rays come through the head and thus project on the image several contours which may intersect together and whose saliency depends on the opacity of surrounding tissues. Often this obliges the speech scientist to take into account two consecutive images to enhance contour perception.

Therefore we are working with the aim of developing a software which can tract the articulators with minimal user interaction. The algorithm we designed rests on recent advances in computer vision which have allowed contour extraction and tracking techniques (by means of the active contour method called "Snake") to be developed.

The paper is organized as follows: first we give a brief overview of the snake method and we explain why this concept cannot allow the tracking to be performed properly especially in the upper jaw region. Then we present the motion based algorithm we have developed to track the tongue in the upper jaw. Then we describe the whole algorithm. Finally we exhibit examples that prove the relevance of our method.

1.2. Contribution of the image processing techniques

Visual tracking is one of the most fundamental problem in the vision community and numerous works have been devoted to this task.

When the shape to be tracked contains salient or characteristic points or features (corners, typical contours, colors...), token matching methods can be used [1]. Otherwise, researchers often resort to optical flow methods: provided that the motion of the feature to be tracked is not too important between two frames, the velocity of each point in the image can be computed with more or less accuracy. The problem we address is very difficult for many reasons:

- the images are very noisy and the tongue appears as a weak contour in the image. Moreover, the tongue may be obscured by superimposed structures like the teeth or the dental fillings. To be convinced of that, the reader can look at the edge map of an X-ray image(Fig. 1).
- The displacement of the tongue between two frames can be very important (for instance fin the region of the tongue tip) and above all, it is non rigid. Hence, classical algorithms using rigidity constraints on the object to be tracked cannot be used.

To cope with theses problems specific to curve tracking, numerous tracking algorithm make use of *active contour models* (also known as Snakes) [3]. In this concept, a snake is a deformable curve evolving under the influence of an energy term that pushes it towards the nearest edge. This concept id widely used, especially for medical imaging: once the snake is initialized on the contour in the first frame, it will automatically track the contour from frame to frame.

At first sight, this method is well suited for our purpose and experiments on this topic are related in [5]. Nevertheless, because other feature are superimposed on the tongue, tracking cannot be achieved

Figure 1: The edge map of an X-ray image.

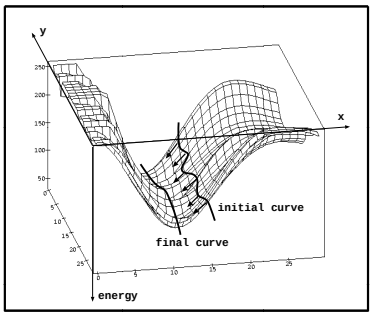

Figure 2: The snake process.

alone with snake methods. We have therefore designed a tracking tool that makes a snake method and a motion based method cooperate together; the tracking task is therefore achieved properly with minimal user interaction.

2. TRACKING THE TONGUE WITH SNAKES: USEFULNESS AND LIMITS

2.1. Overview of the snake method

The basic idea behind the "snake" method [3] is to design an energy functional whose local minima comprise the set of contours to be searched for. As contours are characterized by gradient local maxima, the energy function integrates the opposite of the gradient information. The contours are reached by minimizing this functional: from an initial curve, the "snake" moves towards the nearest contour under the influence of the force field created by the gradient. Hence the snake minimizes the energy term:

$$E = \int_C (\alpha|v'|^2 + \beta|v''|^2) - \int_C (|\nabla I(v(t))|)dt$$

where I is the intensity of the X-ray image.

The first two terms impose regularity constraints on the curve while the third one compels the snakes to reach curves with high gradient. Since the energy E is not convex, there may be many local minima of E. The Euler Lagrange Equation (1) [4] allows us to characterize any such local minimum:

$$-\alpha v'' + \beta v^{iv} - \frac{\partial|\nabla I(v(t))|}{\partial v} = 0 \qquad (1)$$

This equation can only be solved through an iterative way from a rough estimation of the contour position. (See [3] for further details). Fig. 2 illustrates the snake process: from the initialization, the curve is submitted to the force field created by the potential E and reaches the minimum of the energy while preserving smoothness constraints on the solution.

Hence, because of the regularization term, the snake process allows contours to be detected contours even for noisy images provided that an initial guess of the contour position is given. This explains why snake methods are often used for tracking task.

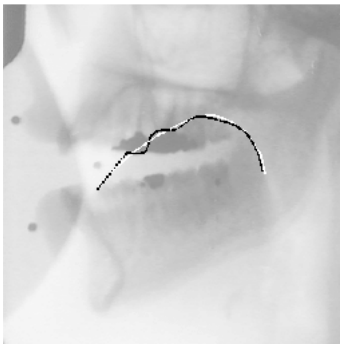

Figure 3: The snake is attracted by points presenting high gradient.

2.2. Why the snakes cannot alone perform tongue tracking

However, this approach is restricted to the case that the contour to be tracked is sufficiently isolated. Indeed, the snake may be attracted by spurious contours and especially by contours with high gradients. Fig. 3 illustrates the excessive flexibility of the snake for our particular application. In the upper jaw region, some contours as the dental fillings are superimposed on the tongue. Then, starting from curve that roughly delineates the tongue (in white), the snake will be attracted by the highest gradient points and will converge towards the dental fillings (in black). This proves that an approach only based on snakes is doomed to failure. Other authors proposed to increase the snake rigidity in order to avoid these problems [6]. However, our experiments showed us that such methods have no efficiency if surrounding contours are very strong.

To summarize, we claim that tongue tracking can be achieved with snakes in the pharynx region because the contour is sufficiently isolated. But other methods must be developed in the upper jaw regions because superimposed contours hide the tongue contour.

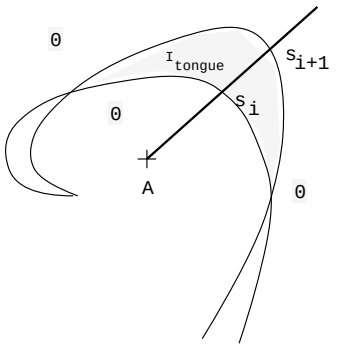

Figure 4: Modeling the tongue displacement.

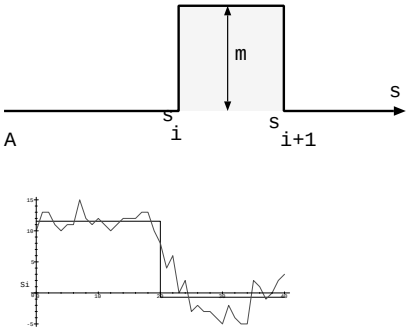

Figure 5: (a): theoretical difference profile along a ray (b): real difference profile.

3. TRACKING THE TONGUE IN THE UPPER JAW REGION WITH MOTION BASED METHOD

Instead of detecting the contour in a static image using the contour detected in the previous one, we rather attempt to compute the motion of the tongue contour using grey level variations between two consecutive frames [2]. The advantage of such a method is that other contours lying in the vicinity of the tongue do not disturb motion retrieval. On the other hand, the tongue position is only recovered with little accuracy.

Before using motion based methods, the images of the sequence must be registered. Indeed, the X-ray film we have at disposal undergoes some jittering effect. Moreover, since the radiation dose of X-ray is not constant, the intensity may change noticeably between two frames.

The tongue is roughly an homogeneous organ. Thus, since we are concerned with X-ray images, we can consider that if the tongue overlaps a point in the image, the intensity at this point increases with a constant value I_{tongue} [1]. Hence, in the ideal case, given two consecutive images I_i and I_{i+1}, we have (Fig. 4):

$$I_{i+1}(x,y) - I_i(x,y) = \begin{cases} 0 \ if (x,y) \ does \ not \ belong \ to \ the \ tongue \ in \ I_i \\ nor \ in \ I_{i+1}, \ or \ if \ (x,y) \ belongs \ to \ the \\ tongue \ both \ in \ I_i \ and \ I_{i+1} \\ \\ +/- I_{tongue} \ \ if \ (x,y) \ belongs \ to \ the \ tongue \\ in \ I_{i+1} \ but \ not \ in \ \ I_i, \ \ \ or \ if \ (x,y) \ belongs \\ to \ the \ tongue \ in \ I_i \ but \ not \ in \ \ I_{i+1} \end{cases}$$

It must be noticed that this equation is fulfilled except for points belonging to the dental fillings.

Let A be an origin point lying inside the tongue contour and chosen by the practitioner in the first image of the sequence. We consider the rays passing through A and any point belonging to the tongue contour in I_i. Hence, the theoretical profile $Diff_{ray}$ of the difference image on such a ray is a hat (Fig.5.b). The tongue position in I_{i+1} is then computed by searching the step edge fitting at best the difference profile on each ray (Fig .1.c): let s_i be the tongue position on the ray; we search for s_{i+1}, m_r, m_l minimizing

$$\sum_{s_i \leq s < s_{i+1}} (Diff_{ray}(s) - m_l)^2 + \sum_{s_{i+1} \leq s \leq s_{max}} (Diff_{ray} - m_r)^2$$

where s_{i+1} is the contour position on the ray in I_{i+1},s_{max} is the length of the ray, m_l is the hat height and m_r is the step height after s_{i+1}.

m_r is theoretically equal to 0 but due to the noise, this value is not null (Fig. 5). (Fig 6.a) exhibits the predicted points obtained with this method on the difference image (the tongue contour in I_i is also shown). Among them, some are erroneous and are removed on a statistical point of view: let m be the average distance between s_i and s_{i+1} on each ray, and let σ be the associated standard deviation. The points for which $s_i s_{i+1} < m - \sigma$ are removed from the predicted curve as well as the points belonging to the fillings (Fig 6.b).

At last, the contour in the upper jaw is recovered by fitting a B spline curve to these predicted points (Fig 2.c). The final contour of the tongue in the upper jaw is shown on the intensity image I_{i+1}.

4. ALGORITHM LAYOUT

We now describe the complete algorithm we have designed.

Initialization:
The practitioner outlines the tongue in the first image and supplies a rectangle containing a motionless region (from the top of the image up to the dental fillings in the upper jaw). He also chooses an origin for the tongue representation (point A in Fig. 4.a). The rectangle and the origin are updated according to the registration process in the subsequent images. The basis of the rectangle is used as separator between the upper jaw and the pharynx region.

For each new image I_{i+1} do:

[1] This hypothesis is not fulfilled for the fillings because the intensity is saturated at these points.

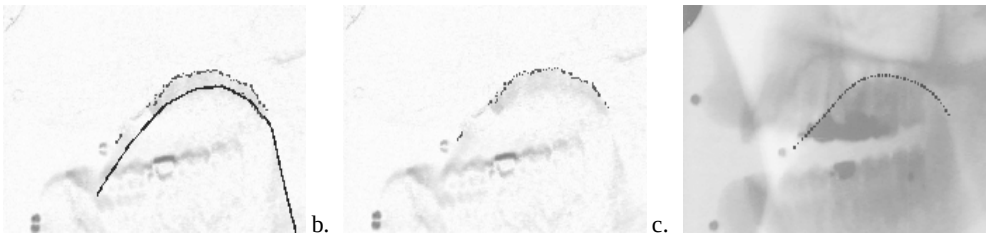

Figure 6: Predicting the tongue contour in the upper jaw region: (a) the tongue contour in I_i and the predicted points in the upper jaw, (b) the predicted points after statistical filtering (c) the tongue contour in the upper jaw.

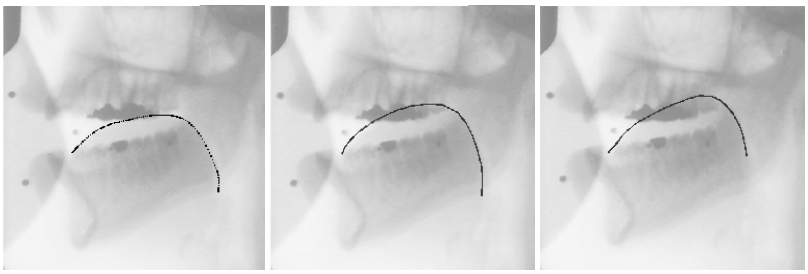

Figure 7: Example of tongue tracking in three consecutive frames

1. **Registration**
   The consecutive images I_i and I_{i+1} are registered in order to match the upper jaw, which may be considered motionless. Note that registration must allow both for motion and average energy variation between the two images.
2. **Tracking the different parts of tongue**
   Considering the tongue contour obtained in I_i, regions corresponding to the three types of contour mentioned above are searched for. This is achieved by studying the contour location according to the contour tooth region of the upper jaw and the filling region. According to the contour type the following methods are triggered:
   (a) B-spline based active contour for isolated contour (pharynx region),
   (b) contour prediction achieved by computing the motion of the tongue in the upper jaw region,
   (c) hidden contour regions can be filled in by continuity.
3. **Recovering the whole tongue contour**
   The global contour id deduced from the sub-contours using a B-spline approximation.

(Fig. 7) exhibits tongue tracking results in three consecutive frames. Though the tongue undergoes a large displacement between two frames, the tongue contour is properly recovered.

5. CONCLUSION

This tracking algorithm has been tested on a short film transferred to video format and then digitized. Our experiments have shown that the tongue is tracked successfully when it is visible, and that interaction with the user remains necessary when the tongue contour is too obscured or the tongue moves too fast. It seems to us that possible improvements might come from two aspects: (**i**) a physiological model of the tongue which allows deformations produced by muscles to be controlled, (**ii**) an articulatory model which gives the space of possible tongue shapes. The physiological model could be used to predict the tongue shape from the current contour and the articulatory model to validate contours. However, one must keep in mind that this necessitates a preliminary speaker normalization which is not possible accurately before the tongue contours are available. Therefore prediction and validation as well must tolerate a sufficiently large error.

6. REFERENCES

1. R. Deriche and O. Faugeras. Tracking line segments. *Image and Vision Computing*, 8(4):261–270, November 1990.
2. E. C. Hildreth. Computations Underlying the Measurement of Visual Motion. *Artificial Intelligence*, 24:309–354, 1984.
3. M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active Contour Models. *International Journal of Computer Vision*, 1:321–331, 1988.
4. R.S. Schechter. *The variational Method in Engineering*. McGraw-Hill Book Comp., New York, 1967.
5. M.K. Tiede and E. Vatikiotis-Bateson. Extracting articulator movement parameters from a videodisc-based cineradiographic database. In *Proc. of ICSLP, Yokohama*, pages 45–48, 1994.
6. N. Ueda and K. Mase. Tracking Moving Contours using Energy Minimising Elastic Contour Models. In *Proceedings of Second European Conference on Computer Vision, Santa Margherita Ligure (Italy)*, volume 588 of *Lecture Notes in Computer Science*, pages 453–457, 1992.