TRACKING ARTICULATORS IN X-RAY IMAGES WITH MINIMAL USER INTERACTION: EXAMPLE OF THE TONGUE EXTRACTION

Marie-Odile Berger and Yves Laprie

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

ABSTRACT

Vocal tract X-ray image sequence are used to study articulatory phenomena and to design approximate articulatory models. The purpose of this paper is to describe an automatic tracking tool for extracting the contours of the tongue which is the most important articulator.

Tracking the tongue in X ray images is an arduous task because it appears as a weak contour and above all because it is immersed among a lot of contours (teeth, palate,...). Hence we have developed a robust algorithm that makes a snake based method and a motion based method cooperate.

Significant results show the strength of the approach.

1. INTRODUCTION

X-ray moving pictures are widely used in articulatory phonetics to investigate how sounds are articulated, and to design approximate articulatory models. The fact that articulator contours must be extracted by hand, which is indeed a tedious task, explains why a vast number of X-ray moving images is still unexploited and probably why some articulatory phenomena are not better known. Recent advances in computer vision have allowed contour extraction and tracking algorithms to be developed which could be of a great help to automatically analyze X-ray images. This work deals with tongue tracking because the tongue has probably the most important articulatory role. Tracking the tongue, i.e. detecting its contour in every frame, is difficult because this contour is weak and its characteristics strongly vary depending on whether it lies in the upper jaw or in pharynx. This explains why other attempts only based on *snakes method* [1] cannot achieve the task properly. To take into account contour characteristics, our tracking algorithm makes a velocity field method and a snake method cooperate, each of them used according to the location of the contour in the vocal tract. This approach has been successfully applied to cine-radiographic movies.

2. TRACKING THE TONGUE CONTOUR IN THE PHARYNX REGION WITH THE SNAKE METHOD

The basic idea behind the snake method [2] 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, such an energy function can be written :

$$E = \frac{-1}{|C|}\int_C |\nabla I|(x(s), y(s))ds$$

where $C = (x(s), y(s))$ is the contour that belongs to a particular class of regular functions, $|C|$ its length, ∇I is the gradient, s is the curvilinear abscissa. The minimization of E is performed using the Euler-Lagrange dynamics equation and must be computed iteratively from an initial curve since $|\nabla I|$ is only known at the discrete grid points of the image. From this initial curve, the snake moves towards the nearest contour under the influence of the force field created by the energy. Thus, snakes can be favorably used for tracking purpose [3, 4, 5] : once the snake is initialized by hand on the tongue contour in the first frame, it will automatically track the contour from frame to frame by using the contour in a frame as the initialization in the next frame.

Note that knowledge on the shape of the tongue can be incorporated in the snake model because the curve C can be searched for in a particular function class. Hence, because of the good results obtained by Liljencrants [6] to describe tongue contours by means of a Fourier series, we first tested our method with this representation. Unfortunately, this approach failed due the lack of relevant gradient information in the possible

filling regions. Therefore, we accepted B-spline snakes [7] which impose smoothness constraints on the curve and yield good results everywhere.

This method is widely used for medical imaging because it allows weak contours to be properly detected and tracked even for noisy images [8]. Nevertheless, this method succeeds only if the contours to be tracked are in some sense isolated. Indeed, if another contour lies in the vicinity of the contour, the snake may be attracted by spurious contours.

Concerning our application, the snake method is inappropriate for tracking the tongue in the upper jaw because the dental contours prevent the snake from reaching the tongue contour. We have thus designed a motion based method in this case.

3. TRACKING THE TONGUE CONTOUR IN THE UPPER JAW REGION WITH A VELOCITY FIELD 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 [9]. 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 Xray 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. 1):

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

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

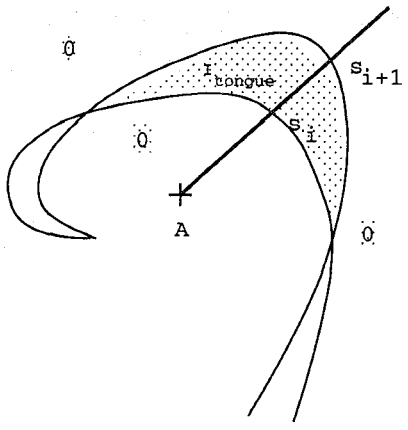

Figure 1: Modelling the tongue displacement.

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

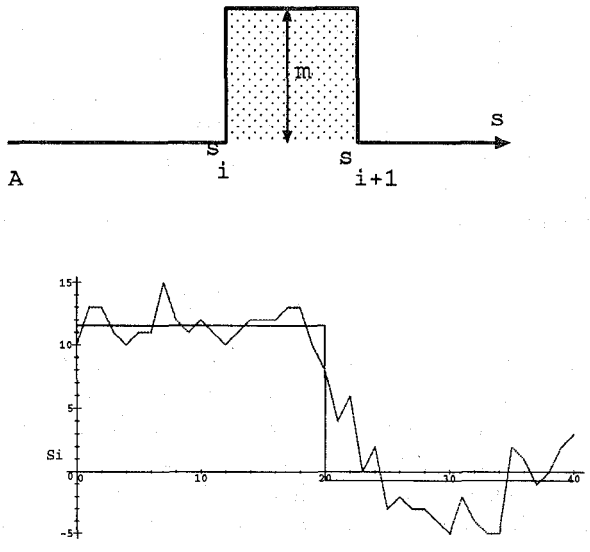

Figure 2: (a): theoretical difference profile along a ray (b): real difference profile.

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.2.a). 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 .2.b): 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. 2). (Fig 3.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 3.b).

At last, the contour in the upper jaw is recovered by fitting a B spline curve to these predicted points (Fig 3.c). The final contour of the tongue in the upper jaw is shown on the intensity image I_{i+1}.

4. OVERVIEW OF THE ALGORITHM

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. 1.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. **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. 3) 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. PERSPECTIVES

Although a totally automatic articulator tracking is hard to imagine (the tongue is not always visible, especially when it moves too fast) the X-ray material, in spite of the poor quality of the images used, validates successfully this approach.

We are aware that our method must be adapted if the motion between two frames is very important (for particular sounds) or if the mouth is closed. Indeed, our algorithm depends on two parameters: the number of iterations in the snake process and the length of the ray on which the homologous point is searched for. If the displacement between two frame is great, these two parameters must be increased; otherwise the predicted contour will be wrong.

Since the tongue contour is very weak, the contour given by the algorithm can hardly be assessed automatically. Thus we plan to use the knowledge of the sound uttered by the speaker to assess, or at least, to guide the tracking process. Approximate articulatory model are indeed available [10] and could be used to detect incompatibilities between the computed tongue motion and the uttered sound.

6. REFERENCES

[1] M.K. Tiede and E. Vatikiotis-Bateson. Extracting articulator movement parameters from a videodisc-based cineradiographic database. In *Proc. of ICSLP, Yokohama (japon)*, pages 45–48, Mai 1994.

[2] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active Contour Models. *International Journal of Computer vision*, 1:321–331, 1988.

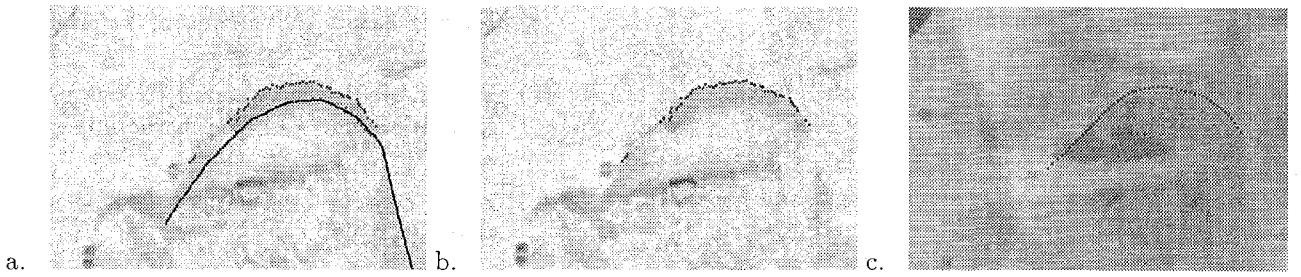

Figure 3: Predicting the tongue contour in the upper jaw region: (a) the tongue contour in I_i and the predicted points in the upper jaw, (b) the predicted points after statistical filtering (c) the tongue contour in the upper jaw.

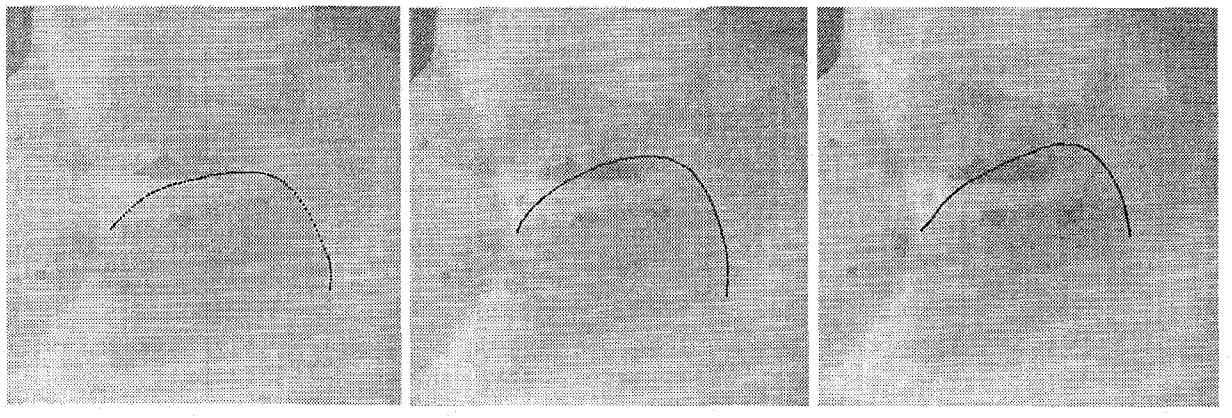

Figure 4: Example of tongue tracking in three consecutive frames.

[3] M.-O. Berger. How to Track Efficiently Piecewise Curved Contours with a View to Reconstructing 3D Objects. In *Proceedings of the 12th International Conference on Pattern Recognition, Jerusalem (Israel)*, volume 1, pages 32–36, 1994.

[4] B. Bascles, P. Bouthemy, R. Deriche, and F. Meyer. Tracking Complex Primitives in an Image Sequence. In *Proceedings of the 12th International Conference on Pattern Recognition, Jerusalem (Israel)*, 1994.

[5] F. Leymarie and M. Levine. Tracking Deformable Objects in the Plane Using an Active Contour Model. *IEEE Transactions on PAMI*, 15(6):617–634, June 1993.

[6] J. Liljencrants. Fourier series description of the tongue profile. *STL, QPSR*, (4):9–18, 1971.

[7] B. Bascles and R. Deriche. Stereo Matching Reconstruction and Refinement of 3D curves Using Deformable Contours. In *Proceedings of 4th International Conference on Computer Vision, Berlin (Germany)*, pages 421–430, 1993.

[8] I. Hunter, J. Soraghan, J. Christie, and T. Durrani. Detection of Echocardiographic Left Ventricule Boundaries using Neural Networks. In *Computers in Cardiology, Londres, 1993*, pages 201–204, 1993.

[9] E. C. Hildreth. Computations Underlying the Measurement of Visual Motion. *Artificial Intelligence*, 24:309–354, 1984.

[10] S. Maeda. Un modèle articulatoire de la langue avec des composantes linéaires. In *Actes 10èmes Journées d'Etude sur la Parole*, pages 152–162, Grenoble, Mai 1979.