EXTRACTION OF TONGUE CONTOUR IN X-RAY VIDEOS

Minghao Yang, Jianhua Tao, Dawei Zhang

The National Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences, Beijing, China
{mhyang, jhtao, dwzhang}@nlpr.ia.ac.cn

ABSTRACT

In spite of the development of new image techniques, X-ray remains an important technique in studying speech production phenomena. In this study, we propose an automatic contour extraction method of tongue in X-ray videos. At first, we take a region gradient based edge-detector to find the initial boundary points. As X-ray is high noise image and tongue is frequently occluded by teeth, there are high ratio outliers in the initial boundary point set. To solve this problem, we propose a cluster based point-to-point distance ratio filter to remove the outliers, which greatly reduces the iteration times of later RANSAC and B-Spline approximation for the final boundary points. Our method is nearly full-automatic, and obtains tongue's accurate contour. The experiments show that the proposed method could be an effective tool for tongue's continuous motion analysis.

Index Terms—X-ray detection, maxlikehood boundary estimation, outliers rejection, RANSAC

1. INTRODUCTION

As a technique of recording the tongue's continuous movement, X-ray is appropriate in studying speech production phenomena [1, 2]. Many works have been proposed for studying the speech production mechanism, including discussing the tongue position and speech production of vowels [3-7], organs' physiological structure and voices classification [8, 9], tongue's continuous motion and speech generation [10] with X-ray data.

To understand the relationship between tongue positions and acoustic signals, researchers tried to extract the contour of tongue in X-ray images. With the hypothesis that tongue moves in a predefined interest region, snake [11], fuzzy active contour models with minimal user interaction [12] and edge-based template matching [13] have been proposed to extract the contour of tongue in X-ray data. Recently, Julie proposed a semi-automatic method to extract the vocal tract contours [2]. In this method, approximately 10% frames were needed to mark by hand [2, 10]. As X-ray is high noise images and tongue is often occluded by teeth in pronunciation, it is still an open problem to extract the contours of tongue in X-ray sequences automatically and accurately.

This work is supported by the National Natural Science Foundation of China (NSFC)(No.61273288, No.61233009, No.61203258), and the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM (CSIDM) Programme Office.

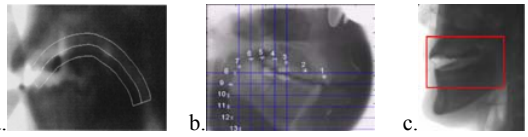


Fig.1. Determine the tongue's movements area manually. (a) the initial contour hypothesis adopted in [11]; (b) the manual step applied on full-size images adopted in [2]; (c) the manual step in our method.

In this study, we present a nearly full-automatic method to extract the contour of tongue. Our method consists of three steps: (1) detecting the initial boundary points in tongue movement area; (2) removing the outliers from the initial boundary points set; and (3) obtaining the final accurate contour with RANSAC and B-Spline approximation. The rest of this paper is organized as follows: we introduce how to detect the initial boundary points of tongue in section 2; the detail discussion on how error boundary points are removed and how to extract the tongue's accurate contour are described in section 3; the experiments are given in section 4; finally, we conclude this study in section 5.

2. DETECTION OF THE INITIAL BOUDARY POINTS

2.1. Registration of the tongue movement range

A predefined range helps to find the initial boundary points quickly in the whole image. In [11], tongue's movement range was limited in a predefined area around a arched contour(Fig.1(a)). In [2], users need to mark about 13 tongue boundary points on near 150 key frames(Fig.1(b)). In our method, users draws a rectangle as the tongue's movement range at the first frame of X-ray video(Fig.1(c)). Then the system is subsequently able to process large image sequences without further interaction automatically. Our method is relatively convenient for user interaction and the rectangle is wide enough to detect the significant changes of tongue positions between frames.

2.2. Detection of the initial boundary points

We aim to find the initial boundary points in tongue movement area quickly. As X-ray is high fuzzy image and tongue is often occluded by teeth, it is difficult to obtain the boundary points with traditional Gaussian kernel edge-detector. In this step, we use a kind of region gradient based edge-detector (maxlikehood boundary estimation [14]) to find the initial boundary points. This algorithm is effective to find the move edge in high noise images and occlusion

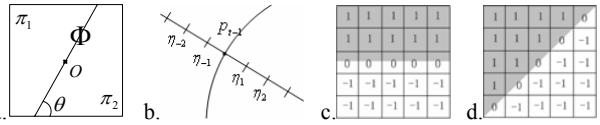


Fig.2. Maxlikehood boundary estimation algorithm. (a) region Φ is divided into two areas π_1 and π_2; (b) exploring the location of p_t along perpendicular direction of p_{t-1}; (c),(d) the 5*5 estimation masks along 0 degrees and 45 degrees.

environments [15, 16]. One of the advantages is that it does not require any prior edge extraction. Only point coordinates and image intensities are manipulated. Its basic idea is given as (1). In (1), $\phi_\theta^{p_{t-1}}$ is a line which divides a region Φ into two areas π_1 and π_2 at point p_{t-1} along angle θ (Fig.2(a)). n_i is the number of pixels in π_i and $\hat{c}_i$ is the mean values of intensity for region π_i ($i \in (1,2)$). Given an edge point p_{t-1} at frame t-1, then the corresponding edge point at next frame could be estimated by $p_t = \arg\max(\xi(\phi_\theta^{p_{t-1}+\eta_i}))$ (i=±1,±2,±3……), where $p_{t-1}+\eta_i$ is the new location along the perpendicular direction of tangent at p_{t-1} with the moving step η_i (Fig.2(b)).

$$p_t = \arg\max(\xi(\phi_\theta^{p_{t-1}+\eta})), \qquad \xi(\phi_\theta^{p_{t-1}+\eta_i}) = \frac{n_1 n_2}{2(n_1+n_2)}(\hat{c}_1 - \hat{c}_2)^2 \qquad (1)$$

In X-ray image, most parts of tongue, especially the front parts, shift between 0 and 45 degrees in pronunciation. Then the initial boundary points could be obtained by (2). In (2), M is the tongue movement area; $\xi(\phi_{0,45}^M)$ presents that each pixels in M is convolved with 0 and 45 masks. Function H(.) takes the pixels which has local maximum value of $\xi(\phi_{0,45}^M)$ as the boundary points. Then we obtain the initial boundary points set S_0. Fig.3(c)(d) presents two 5*5 masks along 0 and 45 degrees.

$$S_0 = H(\xi(\phi_{0,45}^M)) \qquad (2)$$

It is a time consuming procedure that each pixel in M convolves with 0 and 45 masks. Supposing that the intensity variance σ of π_1 is equal to that of π_2, and $n_1 = n_2$, then for a point p_i($p_i \in$ M), $\xi(\phi_{0,45}^{p_i})$ could be simply written as (3), where C is a constant and α ($0<\alpha<1$) is a weight factor for 0 degree mask. In this way, with the help of integral image technique[17], we could obtain the convolution of p_i with two masks in two subtraction operation.

$$\xi(\phi_{0,45}^{p_i}) = C^*(\alpha * \|(\hat{c}_1 - \hat{c}_2)\|_0 + (1-\alpha) * \|(\hat{c}_1 - \hat{c}_2)\|_{45}) \qquad (3)$$

3. EXTRACTION OF THE ACURATE TONGUE COUTOUR

With the maxlikehood boundary estimation, we obtain tongue's initial edge points set S_0. Supposing that the ratio of correct boundary points is denoted as u_s(u_s = w/h), where h is the number of total points in S_0 and w is the number of correct boundary points, the value of h is about 120 and 160, and u_s is between 50% and 70% in our experiments. To obtain the accurate contour points, we propose a cluster based point-to-point distance ratio filter to remove the outliers from S_0, and then combine RANSAC with B-Spline approximation to fit the final tongue contour.

3.1. Outliers rejection

The points close to each other are taken together as a cluster. For example in Fig.3, the points in rectangle A and B makes of two clusters. There are two similar steps in the cluster based point-to-point distance ratio filter. Fig.3 gives the procedure of the first step.

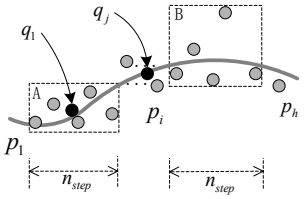


Fig.3. Cluster based point-to-point distance ratio filter.

Algorithm 1:

Input: S_0', h, n_{step}, ratio;
Output: S_1, h_1;
  h_1=0; t=0; j=0;
  while (t≤h/n_{step}){
    Put points p_i($p_i \in S_0'$, t*n_{step}≤i≤(t+1)*n_{step}) into cluster Q_t;
    The points in Q_t are denoted as q_j($q_j \in Q_t$ and 0≤j＜n_{step});
    Let D^t_{max} =max(||q_u-q_v||), D^t_{min} =max(1, min(||q_u-q_v||))
    where (0≤u,v≤n_{step} and u≠v);
    if (D^t_{max} / D^t_{min} ≤ratio)
    {
      $q_j = \sum_{k=1}^{n_{step}} q_k / n_{step}$; Add point q_j into S_1; j++; h_1=j;
    }
    t+=1;
  }

In the first step, all points in S_0 are sorted by x-coordinate, and we denote this new set as S_0', where $x(p_0) < x(p_1) < x(p_2) < \cdots < x(p_i) < \cdots x(p_h)$ (x(.) presents the x-coordinate value of point p_i($p_i \in S_0'$)). Algorithm 1 gives the detail process for the first step of distance ratio filter.

In algorithm 1, parameter "n_{step}" presents the point number of cluster Q_t. Parameter "ratio" is a threshold used to determine whether the "n_{step}" data are correct boundary points or not. If the "n_{step}" points in current cluster Q_t are all "good" points(rectangle A in Fig.3), then their center points "q_j" is a correct boundary point and will be put in output boundary set S_1. Otherwise, Q_t contains outliers(rectangle B in Fig.3), all the points in Q_t are removed.

In algorithm 1, "ratio" determines whether a center point of a cluster is kept as a "good" point. As the size of S_0 and the points' position change greatly between frames, we try to determine the optimal value of "ratio" automatically. Histogram based maximum between-class variance algorithm is able to determine the optimal threshold between classes [18]. Supposing that there are total N(=h/n_{step}+1) clusters in S_0, r^t_{max} is the value of D^t_{max} / D^t_{min} for the t^{th} cluster, and R= arg(max(r^t_{max})) (1≤t≤N), then for every pair q^t_u and q^t_v (0≤u,v≤n_{step} and u≠v) in cluster Q_t, let $r_k = D^t_{max}$ /(max(1,|| q^t_u - q^t_v ||) (1≤k≤n_{step}*(n_{step} -1)/2), then we obtain K distance ratios, where K=N*n_{step}*(n_{step} -1)/2. Based on these K values, we build a distance ratios histogram. The x-coordinate of histogram presents bin's index calculated by (4), where r_k is normalized to r_k' ($r_k' \in$ [0,G] and G>>R). This normalization procedure makes all distance ratios distribute more widely in x-coordinate of histogram, which helps to find the threshold accurately with maximum variance criterion. The value of y-coordinate is the hit number of point pairs' distance ratio located in a normalized bin. Depended on the maximum between-class variance algorithm [18], we obtain the value of "ratio"

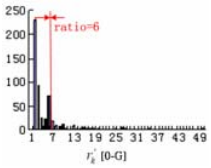


Fig.4. The normalized distance ratio histogram for the 23rd frame of phoneme "a.

automatically with (5). In (5), u_T is the expectation of all distance ratios in a frame, and w_i and $(1-w_i)$ present the probability of distance ratios which belong to (0-i] and (i-G] respectively. Fig.4 presents the normalized histogram of S_0 for the 23rd frame of phoneme "a". The value of G is set to 50 and the value of "ratio" is 6.

$$r_k^{'} = G * (r_k / R) \tag{4}$$

$$\text{ratio} = \arg\max_{i} \left(\max_{i=1}^{G} \left(\frac{(u_T w_i - w_i)^2}{w_i (1 - w_i)} \right) \right) \tag{5}$$

The second step of distance ratio filter is similar to the procedure of the first step. The only difference between them is: the first step is processed in the whole tongue area along x coordinate, and the second step is processed in the front tongue area along y coordinate. With the two steps of cluster based point-to-point distance ratio filter, we obtain new boundary points set S_1 and S_2 respectively.

3.2. Contour approximation with B-Spline and RANSAC

In our experiments, the point number of S_2 is between 8 and 15, and the value of u_s for S_2 is between 82.20% and 97.31%. We further use RANSAC [19] and control points through B-spline approximation to eliminate the final outliers, and obtain the accurate boundary point set S_3. To determinate the tongue tip point, we use the dynamic program algorithm to obtain its position between frames [11]. The root point of tongue is viewed to be fixed at the tip of throat, and is figured out by user at the first frame. Finally, we obtain the whole tongue contour..

4. EXPERIMENTS

4.1. Extraction results

Our algorithm runs on a Chinese female announcer pronunciation X-ray video, which contains 20 phonemes and 181 syllables. The resolution for every frame is 640*480 and time cost for each frame is about 15ms. The computer is dual CPU 2.11G and 2.0G RAM.

Fig.5(a) and Fig.5(b) present the extraction results for the phoneme "ai" and syllable "nu" respectively. There are about 30 frames in the video of phoneme "ai". In this video, the positions of the tongue, especially the tongue tip and tongue mid, change frequently between frames(Fig.5(a)). In the syllable video of "nu", these are about 32 frames. Tongue tip and tongue mid are mostly occluded by teeth (Fig.5(b)). To evaluate the effectiveness of our method, we give the comparison between the extraction results with those of ground truth marked manually(Fig.5(a)(c-f) and Fig.5(b)(c-d), where the white curve is our tracking result, and the green curve is the ground truth marked manually. From the comparison of our method with ground truth, it could be seen that our method is

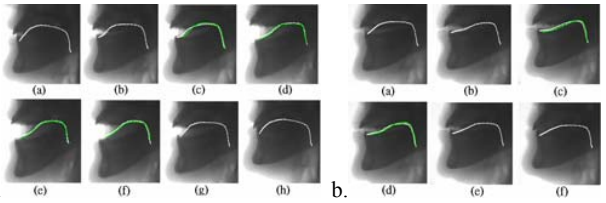


Fig.5. Extraction results: (a) the extraction result of tongue contour for phoneme "ai"(every third frame of the sequence is shown); (b) the extraction result of tongue contour for syllable "nu" (every fourth frame of the sequence is shown).

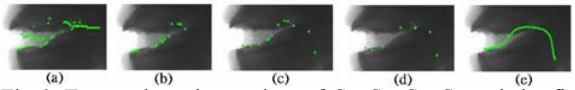


Fig.6. Tongue boundary points of S_0, S_1, S_2, S_3 and the final tongue contour for the 19th frame of Chinese syllable "kuai".

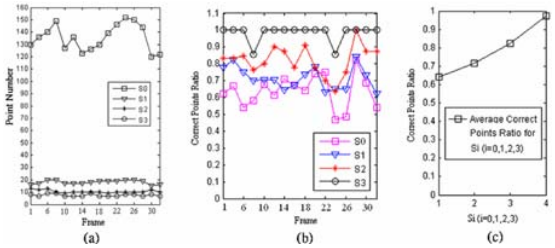


Fig.7. The statistics data for the sequence of syllable "kuai"(total 32 frames). (a): the points number contained in S_0, S_1, S_2 and S_3; (b) the correct boundary points ratio "u_s" of S_0, S_1, S_2 and S_3 (in (a) and (b), every second frame of the sequence is shown); (c) the average correct ratio "u_s" for S_0, S_1, S_2 and S_3.

effective to extract the tongue contour from X-ray videos. The further precision analysis of our method for more phonemes and syllables is discussed in section 4.3.

4.2. Outlier rejection

In our experiments, all the X-videos were processed with the same parameter settings. 11*11 masks are adopted in maxlikehood boundary estimation algorithm and α in (3) is set to 0.5. G is set to 50 and the values of "n_{step}" are set to 5 and 3 for the first and the second step of distance ratio filter.

Fig.6(a)-Fig.6(e) present outliers rejection results of our method for S_0(the initial boundary point set), S_1, S_2 (the boundary point set for the first and the second distance ratio filter), S_3(the point set for RANSAC and B-Spline approximation) and the final tongue contour for the 19th frame of Chinese syllable "kuai" respectively. And more statistics information, including point number and correct boundary points ratio "u_s" of S_0, S_1, S_2, S_3 for syllable "kuai", are shown in Fig.7. It could be seen from Fig.7(a) that with the first distance ratio filter, the range of point number in S_0 and S_1 is decreased from [120-160] to [15-25]. Meanwhile, the average correct point ratio increases from 63.93% to 71.46%(Fig.7(b)(c)). And after the second filter procedure(S_1-S_2), about 9-15 points are left in the boundary point set, and about 82.20%-97.31% points are "good" points, which is benefit to speed up the process of RANSAC

Table 1: The Chinese phonemes and syllables used in precision analysis(denoted by Mandarin pinyin)

phonemes	a, o, e, i, u, ai, ei, ao, ou
syllables	ba, ma, man, ka, kai, kuai, ha, hua, nu, nan, nuo

to obtain the final boundary points set S_3. And in S_3, 97.54% points are correct, with point through B-Spline approximation, the accurate tongue contours are obtained.

4.3. Precision analysis

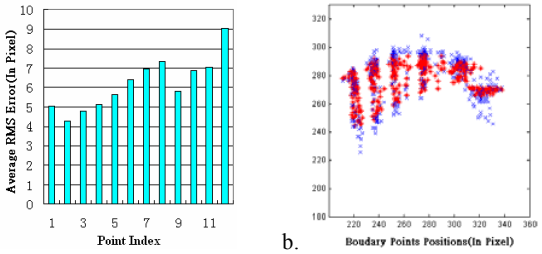
To evaluate the precision of the extraction contour, we compare the results of our method with the contour marked manually on 20 phonemes and syllables random picked from our X-ray database(Table 1). To reduce the errors brought by subjective estimation, the contours are marked by three volunteers(one volunteer is a professor who studies speech production and other two volunteers are medical college students).

Our program first automatically marks 12 points (x^i , y^i), $i \in [0,12)$ on the extracted contour equally from tongue tip to tongue root along x coordinate. The mark procedure is processed every second frames for a given sequence. Each volunteer random choose 5 phonemes or syllables from table 1 and mark the point (x_v^i, y_v^i) which is thought to be ground truth boundary point for (x^i, y^i) one by one. In (x_v^i, y_v^i), v is the index of volunteers($v \in [1,3]$). Then the EMS error for the i^{th} point between our extraction results with that of ground truth is obtain by (6).

$$error^i = \sqrt{(x^i - \bar{x}^i)^2 + (y^i - \bar{y}^i)^2} \quad \text{where} \quad (\bar{x}^i, \bar{y}^i) = \sum_{v=1}^{3} (\bar{x}_v^i, \bar{y}_v^i)/3 \qquad (6)$$

Fig.8(a) present the average RMS errors histogram of the 12 boundary points for the phonemes and syllables listed in table 1. It could be seen from Fig.8(a) that the maximum top 3 errors happen for points 7, 8(tongue dorsum) and point 12(tongue root). In our method we regard the root point as fixed. The head motion in pronunciation makes the tongue root move and generates obvious error from the ground truth position. Fig.8(b) presents the points' location for the extracted boundary points(denoted as red "+") and the marked ground truth(denoted as blue "x") points for syllable "nan". In Fig.8(b), every second points from tip to root point are shown . There are some frames that the ground truth positions of the 1^{st} point are closer to lower teeth and the ground truth positions of the 7^{th} point are nearer to upper jaw than those of our extracted contour. The reason is that tongue tips happen to be occluded by lower teeth and tongue dorsum are occluded by upper teeth. Our algorithm takes the teeth boundary as tongue contour in those frames. However, it could be seen from Fig.8(b) that most of our extraction contour points are close to the positions of ground truth.

In [2], all sentences in the test X-ray video Laval43 are read by a male native speaker of Canadian French. The video resolution is 564*460, and the RMS error for tongue contour is less than 8.5 pixels, where tongue length is about 250 pixels. In [13], the RMS error for tongue is about 15 pixels. While in our X-ray videos, all sentences are read by a Chinese female announcer and the image resolution is 640*480. The tongue length is bout 240 pixels. The estimated EMS error of tongue contour is less than 8.3 pixels, except that the error of tongue root is about 9.0 pixels. As pronouncing X-ray video is important and not easily borrowed resources, we did not compare our results with those of [2] and [13] on same X-ray videos.



a. b.

Fig.8. The precision analysis of extracted tongue contour. (a): the average RMS errors of the 12 boundary points for the phonemes and syllables listed in table 1. (b): location for the extracted boundary points and the marked ground truth points for syllable "nan" (the 1^{st},3^{rd},5^{th},7^{th},9^{th} and the 11^{th} points are shown).

However, from the comparison between our results and those of the ground truth, it could be seen that our method obtain competitive results.

We further calculate the average errors ratio for point 7, 8, 12 with the total length of tongue by $e(p_i)/T$, where $e(p_i)$ presents the average RMS errors for point p_i, and T is the length of tongue(about 240 pixels in our video database). The RMS error ratio for point 7, 8, 12 are 2.92%, 3.0% and 3.75% respectively. With the average error ratios are below 4.0%, our method could be an effective tool for tongue's continuous motion analysis of speech production.

5. CONLUSIONS

In this study, we present an automatic method to extract the contour of tongue. In general, our method consists of two phases: finding the initial boundary points and removing the outliers. To remove the outlier fast and accurately, we first use simplified maxlikehood boundary estimation algorithm to find the initial boundary points in tongue movement's area. To reduce the time cost, related techniques such as integral image is also adopt in our method. Then we propose a cluster based point-to-point distance ratio filter to remove the outliers from boundary point set. The proposed filter is parameter adaptation and is proved to be an effective method to remove outliers in high fuzzy and occlusion X-ray images. Finally, we combine RANSAC with point-through B-Spline technique to obtain the final boundary points. The experiments show that our method could be an effective tool for the continuous motion analysis of speech production in X-ray videos.

6. RELATION TO PRIOR WORK

We propose a method to extract tongue contour from X-ray videos. Being different from the classical methods of fitting tongue contour to a predefined model [11-13], we extract the accurate boundary points in tongue movement area without obvious shape constraint [11] or prior knowledge based edge selection [13]. It is different from the Discrete Cosine Transform (DCT) method [2] that our method is nearly full-automatic. All the interaction operation for user is to mark a rectangle as tongue's movement range at the first frame. Also the tongue's movement range in our method is wider than the area proposed in [11], which helps to detect the significant contour deformation of tongue between frames. Our work could be an effective tool for the continuous motion analysis of tongue movement.

8. REFERENCES

[1] F. Roers, D. Mürbe, and J. Sundberg, "Predicted Singers' Vocal Fold Lengths and Voice Measures Classification—A Study of X-Ray Morphological," *Journal of Voice,* vol. 23, pp. 408-413, 2009.

[2] J. F. Jallon and F. Berthommier, "A semi-automatic method for extracting vocal tract movements from X-ray films," *Speech Communication,* vol. 51, pp. 97-115, 2009.

[3] R. Harshman, P. Ladefoged, and L. Goldstein, "Factor analysis of tongue shapes," *Journal of the Acoustical Society of America,* vol. 62, pp. 693-707., 1977.

[4] M. T. T. Jackson, "Analysis of tongue positions: Language-specific and cross-linguistic models," *Journal of the Acoustical Society of America,* vol. 84, pp. 124-143, 1988.

[5] S. Wood, "A radiographic analysis of constriction locations for vowels," *Journal of Phonetics,* vol. 7, pp. 25-43, 1979.

[6] K. Honda, "Organization of tongue articulation for vowels," *Journal of Phonetics,* vol. 24, pp. 39-52, 1996.

[7] G. M. Fant, "Acoustic theory of speech production," 1960.

[8] Z. Wu and M. Lin, "An Outline of Experimental Phonetics," 1989.

[9] E. Slud, M. Stone, P. J. Smith, and M. G. Jr, "Principal Components Representation of the Two-Dimensional Coronal Tongue Surface," *Phonetica,* vol. 59, pp. 108-133.

[10] R. Sock, F. Hirsch, Y. Laprie, and P. Perrier, "An X-ray database, tools and procedures for the study of speech production. 9th International Seminar on Speech Production " *International Conference on Intelligent Systems and Signal Processing 2011,* pp. 41-48, 2011.

[11] F. Höwing, L. S. Dooley, and D. Wermser, "Tracking of non-rigid articulatory organs in X-ray image sequences," *Computerized Medical Imaging and Graphics* vol. 23, pp. 59-67, 1999.

[12] M.-O. Berger and Y. Laprie, "Tracking articulators in X-ray images with minimal user interaction: Example of the tongue extraction," *IEEE International Conference on Image Processing,* 1996.

[13] G. L. Thimm and J. Luettin, "Extraction of articulators in x-ray image sequences," *In Proc.EUROSPEECH,* September 1999.

[14] P. Bouthemy, "A maximum-likelihood framework for determining moving edges," *IEEE Transactions on Pattern Analysis and Machine Intelligence* vol. 11, pp. 499-511, 1989.

[15] A. N. Stein and M. Hebert, "Local detection of occlusion boundaries in video," *Image and Vision Computing,* vol. 27, pp. 514-522, 2009.

[16] A. I. Comport, É. Marchand, M. Pressigout, and F. Chaumette, "Real-Time Markerless Tracking for Augmented Reality: The Virtual Visual Servoing Framework," *IEEE Transactions on Visualization and Computer Graphics,* vol. 12, pp. 615-628, 2006.

[17] P. A. Viola and M. J. Jones, "Rapid Object Detection using a Boosted Cascade of Simple Features," *International Conference on Computer Vision and Pattern Recognition,* vol. 1, pp. 511-518, 2001.

[18] N. OTSU, "A Threshold Selection Method From Gray-level Histogram," *IEEE Transactions on Systems, Man, and Cybernetics,* vol. 9, pp. 62-66, 1979.

[19] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM,* vol. 24, pp. 381-395, 1981.