Using a biomechanical model for tongue tracking in ultrasound images

Matthieu Loosvelt, Pierre-Frédéric Villard, Marie-Odile Berger

Using a biomechanical model for tongue tracking in ultrasound images

Matthieu Loosvelt, Pierre-Frédéric Villard, and Marie-Odile Berger

LORIA/CNRS, Université de Lorraine, INRIA, France
firstname.lastname@loria.fr

Abstract. We propose in this paper a new method for tongue tracking in ultrasound images which is based on a biomechanical model of the tongue. The deformation is guided both by points tracked at the surface of the tongue and by inner points of the tongue. Possible uncertainties on the tracked points are handled by this algorithm. Experiments prove that the method is efficient even in case of abrupt movements.

Keywords: tracking, biomechanical model, tongue, ultrasound.

1 Introduction

The shape and dynamics of the tongue during speech provide valuable information on the speech production system. Currently, ultrasound imagery is widely recognized as the best way to acquire information on tongue movements at a fast frame rate [2, 14]. As speech applications require the exploitation of rather long speech recording runs, automatic extraction procedures are required. Automatic tracking in US images is known to be very difficult due to low signal-to-noise ratio, high speckle noise, acoustic shadowing, mirroring... This often results in missing parts in the observed contour, especially for fast tongue movements. The interested reader may refer to [11] for an overview of segmentation techniques dedicated to US data. To cope with these problems, there have been many attempts to incorporate prior information on the tongue shape variations to guide the detection. These solutions can be classified into two broad categories: Deformable model-based techniques proposed various extensions of the *Snake* model where contours are guided towards points with high intensity gradients and are submitted to spatial and sometimes temporal regularization constraints to deal with noisy images [10, 1, 15]. On the other hand, learning based techniques make use of manual delineations or of markers glued on the tongue to predict the tongue position in images [12, 4]. However such methods require to manually extract the tongue on large data sets: in [12], 700 X-ray images were considered for training.

In all these tracking methods, the main difficulty is to incorporate only plausible tongue shape deformations within the tracking framework. Regularization techniques brings too general smoothing constraints to be efficient for fast tongue motions whereas learning-based methods need tedious delineation tasks. We propose in this paper a tracking framework based on a biomechanical model which

appropriately characterizes the elastic properties of the tongue and allows tracking even for fast motions.

2 Related works

The use of mechanical models to improve tracking performance is a subject of increasing interest in the computer vision community [16, 8]. The main interest of such models is to improve tracking on parts of the objects with unobserved features. Constraints deduced from the elastic properties of the material generally provide better prediction than general regularization constraints. However, tracking based on mechanics requires to choose the appropriate mechanical models along with its elastic parameters. Fortunately, the mechanical properties of the tongue have been extensively studied [6] and it has been proven that a hyperelastic model is quite appropriate. We thus investigate in this paper the use of a biomechanical model for tongue tracking. To the best of our knowledge, this approach has not been considered in the challenging context of tongue tracking. Past methods are in fact only dedicated to track the tongue contour whereas we aim to track the tongue volume, taking into account keypoints inside the tongue in addition to contour points. Another contribution of the paper is to properly handle uncertain or false point matchings which are common in such noisy images.

The way we extract correspondences over time is described in section 3. The mechanical model which takes into account these features is described in section 4. Results and comparison with existing methods are highlighted in section 5.

3 Extracting point correspondences over time

We describe in this section how do we extract point correspondences over time. The procedure starts with the delineation of a closed contour which contains the upper contour of the tongue and defines the physical region which has to be tracked (see Fig. 1.a). First, the Harris detector is used to find the most prominent corners in the first image of the sequence within this region. These corners are sorted by their quality measure in the descending order. The first 100 features are displayed in Fig. 1.a. The displacement of these points in each subsequent frame is computed thanks to the Farneback's optic flow algorithm [5]. In this approach, the displacement field is supposed to only be slowly varying so that the displacement is computed over a neighborhood of each point with a least square criteria. In addition, we compute the covariance on the estimated displacement at each point. This can be done easily since the estimated covariance on the least squares estimate $\hat{x}$ of the problem $Ax = b$ is given by $\Lambda_x = (A^t A)^{-1}$. Computing this covariance is important since points inside the tongue features in general non ambiguous correspondences and thus a relatively small covariance matrix. On the other hand, since the curvature of the tongue is rather small with homogeneous intensities on both sides of the contours, the points extracted on the tongue contour have a tangential inacurracy in the contour direction and

thus an elongated covariance matrix (see Fig. 1.b). The interest of computing the covariance matrix on the estimated displacement is twofold. First, it allows us to remove features which are not reliable according to the eigenvalues of the matrix (practically matrices with two large eigenvalues of the same order are removed). Second, these uncertainties are integrated within the biomechanical model (Fig. 2). In order to have a relatively small number of evenly distributed features, a minimal distance d between selected features is imposed (15-30 pixels in practice). Features are chosen in increasing order of their covariance matrix. This explains why features with a relatively large covariance matrix are kept in regions where few features were initially detected.

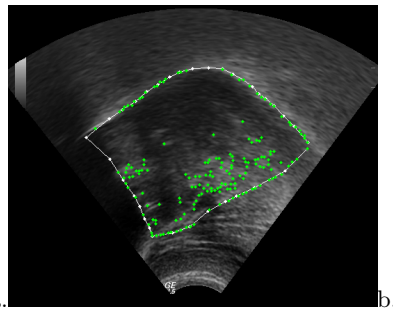
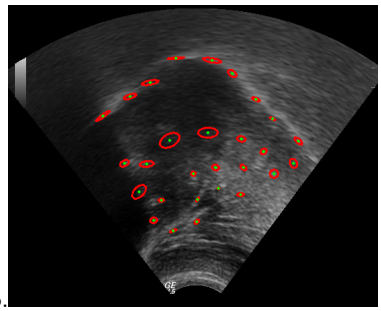
a.  b. 

Fig. 1. (a) The tongue region delineated on the first image of the sequence and the extracted features. (b) Covariance computed on the selected features with $d = 20$ pixels on another image of the sequence.

4 The biomechanical model

Biomechanical models consist in computing 3D object deformations using physical laws. In order to have accurate results we choose to model our problem using the continuum mechanics formulation (conservation laws, matter continuity, etc.). Various numerical techniques exist to solve this set of equations. We use the finite element technique that allows to compute the equations of continuous mechanics. We first define the discrete geometry for dividing the complex problem into small elements. The problem is a mechanical system expressed in ordinary differential equations. We then establish the boundary conditions that the system must satisfy. One of the system equations is the constitutive law that depends on the material properties. Finally we point out how we solve the system.

The geometry

For speech applications, modeling the tongue in the sagittal plane is sufficient. We use a 2D mechanical model with a 2D geometry where all the motions and

deformations occur only in a plane. We thus adopt a quasi-plane strain model, which precisely consists in only focusing on deformations in 2D.

Contours are extracted on the ultrasound image by delineating the upper tongue contour and closing the shape with anatomical landmarks (Fig. 1.a). This 2D polyline is meshed using the "meshAdapt" algorithm [7] while specifying a minimum and maximum edge length of respectively 0.5 and 2 mm. We also impose the points where those displacements have been tracked (see §3) to belong to the mesh. See the resulting mesh on Fig. 2.

The boundary conditions and the constitutive law
Defining boundary conditions in biomechanical problems is not an easy task as they often result from a complex mechanism of muscle contractions and relaxations. Tongue material does not follow a basic elastic law. As in [6], we chose to model its behavior with a hyper-elastic constitutive equation. We used a first order Mooney-Rivlin law converted from a Yeoh strain-energy function with the elasticity parameter values found in [6]. The model is guided by image-based external forces described in the next section. It must be pointed out that a more sophisticated model of the tongue has been proposed in [3] where the tongue is described by 11 groups of muscle geometries. Besides the fact that the anatomy of the speaker is required, such a model is not appropriate for tracking since it would require too many feature points to be controlled.

The forces monitoring the model
The strategy is as follows : instead of modeling the muscle actions we simulate the effects of some key vertices in the tongue as if these points were manipulated by inserted pins. Forces are then dynamically applied to each mesh node n corresponding to the tracked points. The force is linearly transmitted to the neighbors nodes with a kernel factor providing the number of affected neighbors. We chose this factor to be equal to five because it is the best trade-off between shape regularization and mechanical result accuracy.

Each node n displacement information is converted into a spring force $\mathbf{F}_t^n$ dependent on time t with eq. (1). k is a material constant that affects the convergence rapidity, $\mathbf{P}_F^n$ is the final position of a point n given by Farneback's method and $\mathbf{P}_t^n$ is the node position at a given time t after a biomechanical simulation. On Fig. 2 $\mathbf{P}_F^n$ are the green squares and $\mathbf{P}_t^n$ are the blue ones.

$$\mathbf{F}_t^n = k(\mathbf{P}_F^n - \mathbf{P}_t^n) \tag{1}$$

The simulation stops when the $\mathbf{P}_t^n$ gets reasonably closed to the $\mathbf{P}_F^n$.

Handling possible erroneous features
Due to the high level of noise in US images, some features may be tracked with a rather large inaccuracy which is estimated by our algorithm. However, those features are kept and used to guide the mechanical model since they introduce information in areas where few features are available. To take into account the covariance matrix Λ_F^n on the features P_F^n, a Mahalanobis distance is used to compute the potential energy of the spring systems, which gives rise to the

modified forces:

$$\mathbf{F}_t^n = k(\Lambda_F^n)^{-1}(\mathbf{P}_F^n - \mathbf{P}_t^n) \quad (2)$$

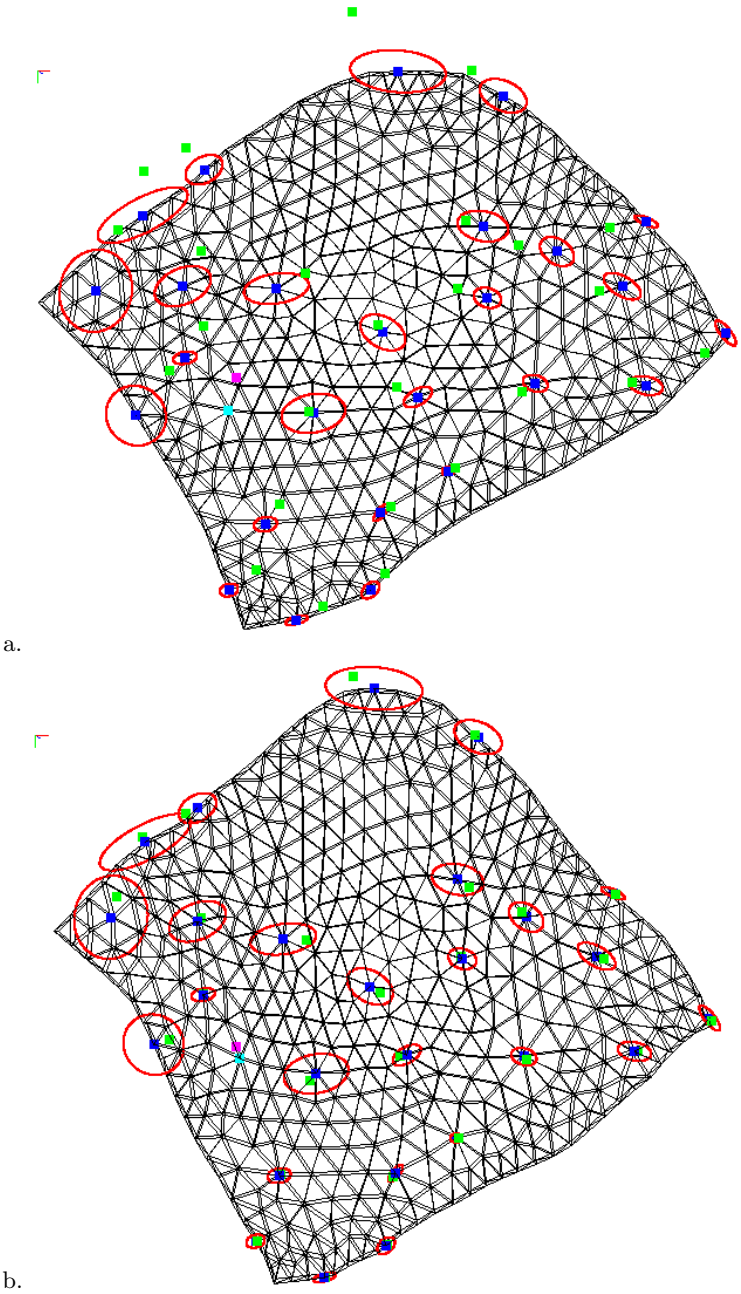
The system resolution
Ordinary differential equations are solved using the finite-element method. They are integrated through an implicit backward Euler method [13] that is reasonably stable for small time step. It is not unconditionally stable because it uses a semi-implicit integration rather than a fully implicit integration. Tongue shape undergoes significant changes during speaking. Large deformation are then taken into account by using the Inversible Element model [13]. Practically our method has been implemented in C++ using Vega FEM library [13]. The following parameters are used: $k = 20$, time step $= 0.1$, the simulation ends when the mean distance between the current and the final point is less than 0.1 mm (note that the size of the pixel in our ultrasound images is 0.18mm). An exemple of convergence of the method is shown in Fig.2.

5 Results

Experiments have been conducted on several ultrasound sequences of natural speech. The size of the images is 532x434 while the size of the pixel is 0.18mm.

Method accuracy relative to the point location
Leave-one-out cross validation is used to evaluate the accuracy of the method. We took the case where a minimal distance $d = 15$ pixels is applied. We have alternatively removed every point $\mathbf{P}_F^n$ of the mechanical model and computed the euclidean distance between this point moving without constraint and $\mathbf{P}_F^n$. Fig. 2 shows an illustration of this method : light blue squares represent the position of the point calculated by the mechanical model, while pink squares represent the final position of the point, estimated with Farneback method. Fig. 3.a shows the error for every point. Two points display an error superior to 3 mm (#1 and #15). They correspond to locations close to dark areas where the computed variance is high. The other distances are low and show that our model is stable relatively to the constraint point locations. Fig. 3.a also shows that the median error is 0.7 mm : this range of accuracy is quite compatible with the dimensions of our target, namely building a dynamic model of the vocal tract from US images.

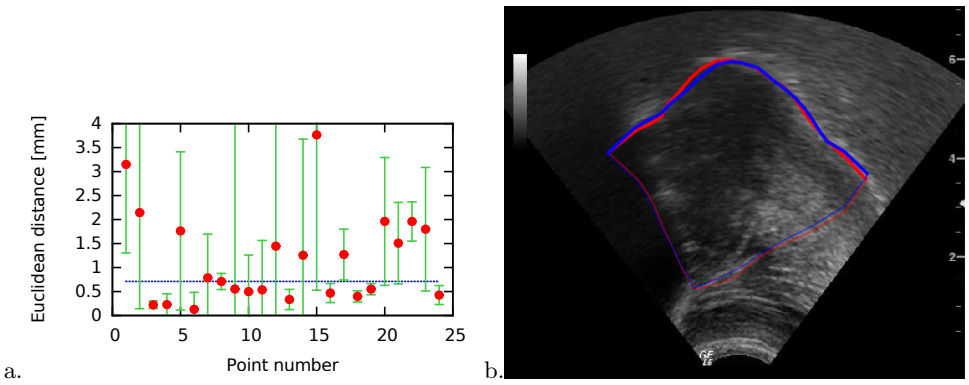
Method accuracy relative to the number of points
Our method also depends on the number N of points tracked by Farneback's method. We have tested the influence of this number N by applying our method with two different minimal distances $d = 30$ pixels and $d = 15$ pixels. Fig. 3.b shows two tongue contours computed with both datasets. There is a slight difference due to the too low constraint number with $d = 30$, which is not enough to guide the model. More accurate results are obtained when N increases ($d = 15$). However, we cannot use too small d distances since it may produce mechanical incoherences.



a.

b.

Fig. 2. Convergence of the biomechanical model: (a) initial position: points detected in the first image are in blue and their associated convergence is in red. They are integrated into the mesh. Green points are the points detected in the final image (b): Position of the mesh after convergence. All the green points are now within the covariance areas of the blue points.



a. b.

Fig. 3. Euclidean error for every tracked points. Median (blue line) is 0.71 mm. The largest eigenvalue of Λ_F^n is also displayed in green with a 50% scale. (b)Influence of d on the accuracy: more accurate results are obtained with d=15 pixels (red curve) than with d=30 pixel (red).

Method validation on ultrasound sequences

Fig. 4 shows nine images extracted every over image in one ultrasound sequence. Upper tongue contours have been computed with our method. They have been compared with a ground truth, which is a sequence of manually segmented contours. The errors ϵ between both contours for each image have been computed and are displayed in pixels and converted into millimeters. The errors remain below 1mm.

Comparison with existing methods

In order to prove the effectiveness of our approach, we have compared our method with the state-of-the art method for tongue tracking [15]. We also compare it with the well established CPD method which enables point set non-rigid registration (see [9] and the publicly available code). These methods were compared to manually detected contours on a challenging sequence with moving upward compression motion of the tongue (Fig.5, top). Comparison of these three methods are shown in (Fig.5, bottom). Tang's method (red curve) in unable to detect the right contour through the 15 frames sequence when sudden abrupt large motions occur. This is probably due to the temporal regularization term which is unable to cope with non-uniform temporal variations. CPD method (green curve) gives better results, but these results are dependent on several parameters (regularization and size of the Gaussian kernel). Our method (white curve) displays the best results despite the fact that there are few points on the right part of the tongue. This is due to the use of the elastic mechanical model which allows for better interpolation in areas where features are missing or sparse.

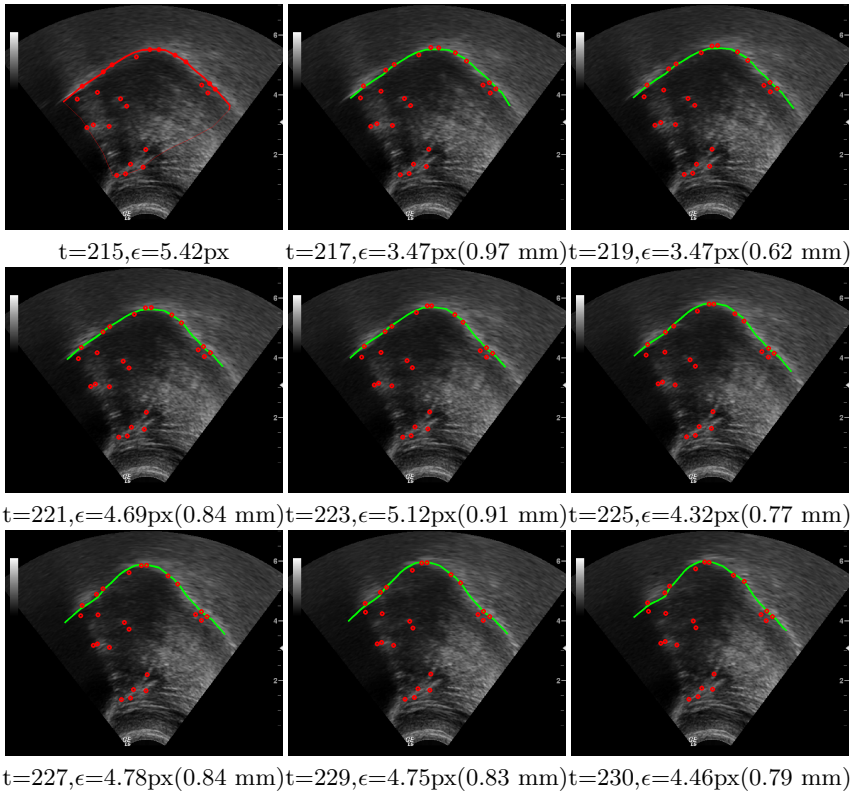


Fig. 4. Ultrasound image sequence: initial contour is in red, tracked points are with red circles and computed contours are in green. The errors ϵ show the mean distance (expressed in pixels and millimiters) between our method and manually segmented contours.

6 Conclusion

We have proposed a new approach for tongue tracking based on a biomechanical method. Using both contour points and inner points along with their uncertainties allows us to efficiently guide the mechanical model with a small number of points. Experiments show that the method is reliable and more efficient than existing methods to handle non uniform tongue movements. Future work will concern the pre-processing step in order to allow detection in few-textured areas and the incorporation of M-estimators to cope with possible outliers. We will also improve our mechanical model by adding collision detection with the palate in order to reproduce its crucial interaction with the tongue.

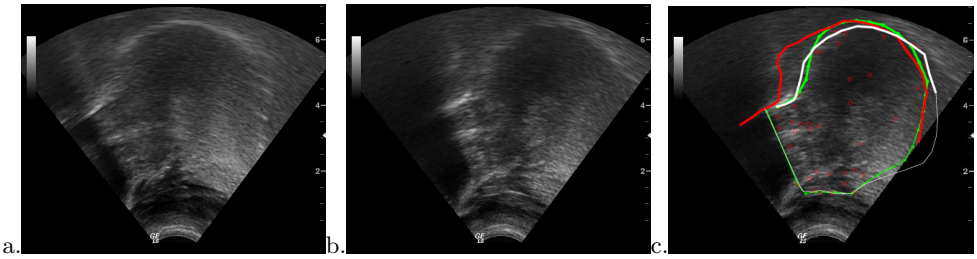


Fig. 5. a,b : Snaphots 1 and 15 of a 15 frames tongue sequence with a rearword motion. c : Tongue tracking with different methods: our method (white), Tang's method (red), CPD method (green). The points tracked in the sequence are drawn with red circles.

References

1. Michael Aron, Anastasios Roussos, Marie-Odile Berger, Erwan Kerrien, and Petros Maragos. Multimodality Acquisition of Articulatory Data and Processing. In *16th European Signal Processing Conference - EUSIPCO 2008*, 2008.
2. Tim Bressmann, Parveen Thind, Catherine Uy, Carmen Bollig, Ralph W Gilbert, and Jonathan C Irish. Quantitative three-dimensional ultrasound analysis of tongue protrusion , grooving and symmetry: data from 12 normal speakers and a partial glossectomee . *Clin Linguist Phon*, 19(6-7):573–88, 2005.
3. Isabelle Buchaillard, Stéphanie, Pascal Perrier, and Yohan Payan. A biomechanical model of cardinal vowel production: muscle activations and the impact of gravity on tongue positioning. *Journal of the Acoustical Society of America*, 126(4):2033–2051, October 2009.
4. Mohsen Farhadloo and Miguel Á. Carreira-Perpiñán. Learning and adaptation of a tongue shape model with missing data. In *ICASSP*, pages 3981–3984, 2012.
5. Gunnar Farnebäck. Two-frame motion estimation based on polynomial expansion. In *Proceedings of the 13th Scandinavian Conference on Image Analysis*, SCIA'03, pages 363–370, Berlin, Heidelberg, 2003. Springer-Verlag.
6. J. M. Gerard, J. Ohayon, V. Luboz, P. Perrier, and Y. Payan. Non-linear elastic properties of the lingual and facial tissues assessed by indentation technique. Application to the biomechanics of speech production. *Medical Engineering and Physics*, 27(10):884–92, December 2005.
7. Christophe Geuzaine and Jean-François Remacle. Gmsh: A 3-D finite element mesh generator with built-in pre- and post-processing facilities. *International Journal for Numerical Methods in Engineering*, 79:1309 – 1331, 2009.
8. Nazim Haouchine, Jérémie Dequidt, Igor Peterlik, Erwan Kerrien, Marie-Odile Berger, and Stéphane Cotin. Image-guided Simulation of Heterogeneous Tissue Deformation For Augmented Reality during Hepatic Surgery. In *ISMAR - IEEE International Symposium on Mixed and Augmented Reality 2013*, Adelaide, Australia, October 2013.
9. Bing Jian and Baba C. Vemuri. Robust point set registration using gaussian mixture models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(8):1633–1645, August 2011.
10. M. Li, X. Khambhamettu, and M. Stone. A level set approach for shape recovery of open contours. In *ACCV*, pages 601–611, September 2006.

11. J. A. Noble and D. Boukerroui. Ultrasound image segmentation: a survey. *IEEE Transactions on Medical Imaging*, 25(8):987–1010, 2006.
12. Anastasios Roussos, Athanassios Katsamanis, and Petros Maragos. Tongue tracking in ultrasound images with active appearance models. In *ICIP*, pages 1733–1736, 2009.
13. Funshing Sin, D. Schroeder, and Jernej Barbic. Vega: Non-linear fem deformable object simulator. *Comput. Graph. Forum*, 32(1):36–48, 2013.
14. M. Stone, G. Stock, K. Bunin, K. Kumar, M. Epstein, C. Kambhamettu, M. Li, and J. Prince. Comparison of speech production in upright and supine position. *Journal of The Acoustical Society of America*, 122(1), 2007.
15. Lisa Tang, Tim Bressmann, and Ghassan Hamarneh. Tongue contour tracking in dynamic ultrasound via higher-order {MRFs} and efficient fusion moves. *Medical Image Analysis*, 16(8):1503 – 1520, 2012.
16. Stefanie Wuhrer, Jochen Lang, and Chang Shu. Tracking complete deformable objects with finite elements. In *3DIMPVT*, pages 1–8, 2012.