Software available: http://tonguetrack.cs.sfu.ca

Graph-based tracking of the tongue contour in ultrasound sequences with adaptive temporal regularization

Lisa Tang, Ghassan Hamarneh
Medical Image Analysis Lab, Simon Fraser University, Canada
{lisat,hamarneh}@cs.sfu.ca

Abstract

We propose a graph-based approach for semi-automatic tracking of the human tongue in 2D+time ultrasound image sequences. We construct a graph capturing the intra- (spatial) and inter-frame (temporal) relationships between the dynamic contour vertices. Tongue contour tracking is formulated as a graph-labeling problem, where each vertex is labeled with a displacement vector describing its motion. The optimal displacement labels are those minimizing a multi-label Markov random field energy with unary, pairwise, and ternary potentials, capturing image evidence and temporal and smoothness regularization, respectively. The regularization strength is designed to adapt to the reliability of images features. Evaluation based on real clinical data and comparative analyses with existing approaches demonstrate the accuracy and robustness of our method.

1. Introduction

Ultrasound (US) imaging is the most effective technique to capture the motion of the human tongue during speech. Analysis of tongue US to extract information about the tongue's shape and dynamics has numerous applications, e.g. studying the effects of aging and glossectomies on speech, studies in linguistics and phonetics, and speech perception and tongue modeling [4,6,21]. A crucial component to tongue shape analysis is the extraction of tongue contours from an US sequence. However, as the number of frames of a typical US sequence is very large, manual segmentation of the tongue contour from each frame for subsequent shape analysis is unrealistic and thus techniques for automatic segmentation are urgently needed.

The automatic segmentation of US images and tracking of shapes within are generally difficult due to known problems [21]: 1) low signal-to-noise ratio; 2) high speckle noise corruption; 3) general US artifacts (e.g. acoustic shadowing, mirroring, and refraction [4]); and 4) weak inter-frame relation due to partially decorrelated speckle noise. In US, parts of the tongue will often disappear between consecutive frames, such that matching based on anatomical landmarks, e.g. apex of the tongue, cannot be reliably performed. Further, structures within the tongue (e.g. tendons and blood vessels) may function as US wave reflectors, causing not only sporadic disappearances of the entire tongue contour, but also occurrences of bright profiles in non-interest regions [13,4], rendering intensity information unreliable. Furthermore, unlike echocardiographs and breast US images, textures useful for matching are lacking in US tongue images, making the automatic segmentation of tongue contour in these images a challenging problem.

Various segmentation approaches for US sequence have been proposed in the literature, including active contour approaches (e.g. snakes); optical flow or registration-based; and tracking. Snakes depend on external and internal energies to pull the contour estimate towards matching image features while constraining the contour to be smooth. In the tongue segmentation algorithm of Akgul et al. [4], the energy functional used includes a similarity term that forces the final contour to be similar to a template contour, and two smoothness terms that either restrict the angles between consecutive points on the contour or constrain the stretching of the contour. In [5], Aron et al. introduced a snake-based approach that relies on preprocessing of US frames to enhance edge information, and the use of electromagnetic sensors for contour-initialization based on an optical-flow formulation. More recently, Roussos et al. [24] incorporated models of shape variations and active appearance models. The use of these models, however, required additional information: X-ray videos gathered during the same US scan were used to construct a motion model and expert's manual annotations of US frames were used to build a texture model. Besides the need for a laborious annotation process, which may be erroneous as landmarks are lacking in US images, the acquisition of X-ray images is not always available, making their approach fairly impractical. In [18], Li et al. also developed a snake-based tongue contour extraction software, called EdgeTrak, that accounts for local edge gradient, local region-based intensity information, and contour orientation. Due to its public availability, this software has become a popular tool [7, 22] for tongue tracking in US and for comparative analyses. e.g. [5, 24]. However, as noted in [24], because EdgeTrak uses a local approach (the contour of one frame is used to initialize the next), it tends to lose track of the contour after having segmented few frames and so, manual refinement is usually needed.

Pairwise image registration of consecutive US frames has also been proposed. The obtained deformation fields are then applied to an input contour to generate the segmentations of all frames. For example, Ledesma-Carbayo et al. [16] applied parametric elastic registration to align consecutive cardiac US frames. Nielsen [20] extracted speckle patterns for use in a block-matching registration algorithm to track the left-ventricle (LV) in US. Duan et al. [8] applied correlation-based optical flow to track LV endocardial surfaces in volumetric echocardiography. More recently, Leung et al. [17] proposed and validated a feature-matching based registration algorithm for real-time applications.

In Bayesian tracking, segmentation is usually formulated as an estimation of a posterior probability of a segmentation contour given all past observations (i.e. image frames). Two popular techniques include Kalman filtering and particle filtering. In [28], Lin et al. proposed particle filtering for tracking the LV in echocardiography using a shape model built via Principle Component Analysis (PCA). They employed particle filters to sample and constrain the allowable space of shape transformations. The likelihood of each sample shape given the observed image data is evaluated to estimate the target shape by a weighted combination of all shape samples. However, they have tested their method on a single US echocardiography only. Furthermore, it is unclear how many manual segmentations are needed to construct a reliable PCA model, as it is not stated, and whether a linear (Gaussian) shape model is indeed valid. In [1], Abolmaesumi et al. extracted artery contours from US images using an edge-based algorithm that was incorporated with temporal Kalman filtering. However, no analyses were done to quantify the accuracy and robustness of their approach. Lastly, Qin et al. [22] proposed a semi-automatic approach to estimate mid-sagittal tongue contour using a learned radial basis function network that nonlinearly maps three landmark locations to a contour estimate. Parameters of a spline-interpolation were then optimized via numerical minimization of reconstruction errors computed based on their groundtruth segmentation data. Nevertheless, their test dataset was limited to an US sequence of one speaker. Their approach also required a high degree of intervention (selection of 3 or more landmarks per frame) and a separate training procedure per US sequence.

All of the aforementioned methods also have their limitations. Snake-based approaches demand good initialization and, in the absence of training data and presence of high levels of noise corruption, require interactive procedures to refine the obtained segmentations. Accuracy is thus generally sensitive to initialization and model parameters (e.g. number of curve segments or edge points used). Pairwise registration and optical-flow based approaches, which propagate results from one frame to initialize the segmentation in another, can easily accumulate errors and eventually lose track of the subject. Lastly, Bayesian tracking and general tracking methods based on velocity models or parametric models often require training data [21, 19].

In this paper, we propose a graph-based approach that performs tongue contour-tracking semi-automatically without the need for training data and tedious refinement procedures. Using a single input of a contour extracted from the start of an US sequence, our method can generate accurate tongue contour segmentations for all subsequent frames in the whole sequence. Our approach casts the tracking problem as a graph-labeling problem wherein each vertex of a graph corresponds to the final position of a control point on a tongue boundary in a frame. Our goal is then to assign to each vertex a displacement label that maps this particular vertex to a reference contour point. In solving this labeling problem, we define data-likelihood terms designed specifically for US images and spatio-temporal regularization terms that ensure each tongue contour estimate is smooth and that it evolves consistently over time. The amount of regularization is also designed such that it adapts according to the quality of the image data in each frame.

Our proposed method may be seen as most similar to [10] where Friedland and Adam performed ventricular segmentation on US image sequences. In their method, a 1-dimensional cyclic Markov Random Field (MRF) was used to describe a configuration of radii values, each of which gives a distance of a point on the detected cavity boundary to a pre-detected centroid. Simulated annealing was used to optimize an energy function that: 1) forced each detected boundary point to be centered on an optimally detected edge; 2) ensured smoothness in the radii values of neighbouring boundary points; 3) guided the overall segmentation to avoid secondary boundaries; and 4) enforced one-way temporal continuity (the current and next frame) in the radii configuration. While shown to be effective for US echocardiographs, their method assumed that the target object is generally elliptical and that a consistent centroid location can be calculated, assumptions which are inappropriate for the segmentation of the non-elliptical tongue contour that also deforms irregularly and in biased directions. Additionally, we model the problem with a special graph and adapt the graph-cuts optimization software of Ishikawa [12] in order to solve our exact problem globally. While the use of graph-cuts had been proposed by Xu and Ajika [27] for active contour-based image segmentation, it was not for segmenting a sequence of images. As only image gradients were used, their approach can only be applied to natural images and, as confirmed by our experiments with their tool, would fail when applied to ultrasound images, which have relatively worse signal-to-noise ratio. We also note that Freedman and Turek [9] had previously proposed a graph-cuts based illumination-invariant tracking algorithm. As we shall present later, our method differs in that we employ a combination of domain-specific regularization terms (spatial and temporal) and incorporate adaptive regularization so that the amount of temporal regularization is increased or decreased depending on the absence or presence of reliable local image features.

To this end, our contributions are: 1) we propose a graph-based approach for US sequence tracking that enforces both spatial and temporal regularization on the contour segmentations; 2) we develop effective ways to encode temporally-varying regularization; 3) we adapt the algorithm of Ishikawa in order to solve the exact problem globally; 4) we examine various approaches to capture image evidence for reliable contour-tracking; and 5) we conduct thorough validation and comparisons with [18, 25] on real asclinical data.

2. Methods

Let there be an image sequence $\mathbf{I}_{0:T-1}$, where the subscripts denote frame numbers 0 to $T-1$, and an initial segmentation contour $\mathbf{x}_0$ that is represented by a vector of spatial coordinates of length N, i.e. $\mathbf{x}_0 = \{\mathbf{x}_{0,1}, \cdots, \mathbf{x}_{0,N}\}$, where $\mathbf{x}_{0,i}$ denotes the spatial coordinates of the i-th control point of $\mathbf{x}_0$. Our goal is to find, for each frame t, a contour $\mathbf{x}_t$ that segments the tongue shape. We shall reach this objective via a graph-labeling approach where we represent the spatial coordinates of *all* segmentation contours in frames 0 to t with a graph and seek to label each node i in the graph with a displacement vector $\mathbf{d}_{t,i}$ such that upon the termination of our algorithm, the spatial coordinates of each control point $\mathbf{x}_{t,i}$ is calculated as $\mathbf{x}_{0,i} + \mathbf{d}_{t,i}$. Just as in many snake-based formulations [4, 18], the optimality of each label assignment is computed as the weighted sum of data-driven and regularization penalties, where the former attracts a segmentation to the desired features in each frame and the latter penalizes excessive bending and discontinuities in the estimated contours. Additionally, we impose temporal regularization to ensure that contours from consecutive frames would not deviate significantly. With a graph-based approach, these penalties are encoded in a MRF energy in which unary potentials capture image evidence while pairwise and ternary potentials capture temporal and spatial regularization, respectively. In contrast to [18,4], where either the solution from a previous frame is used to initialize the next or where temporal regularization is ensured via post-processing, our method solves the tracking problem globally such that it considers all frames in a sequence together and tries to find the most globally plausible set of displacement vectors under *both* spatial and temporal constraints. We now present the details of our graph-based tracking approach.

2.1. Contour tracking via graph-labeling

We begin by constructing a two-dimensional graph $\mathscr{G}(\mathscr{V}, \mathscr{E})$, where the set of control point coordinates $\mathbf{x}_{t,i}$ constitutes the set of graph vertices $\mathscr{V}$. Our graph-labeling approach aims to find a set of label assignments that represent a set of displacement vectors $\mathbf{D}$, where each of its element $\mathbf{d}_{t,i}$ spatially maps the i-th point on $\mathbf{x}_t$ to the i-th point on $\mathbf{x}_0$. To enforce regularization on the assignments, the set of graph edges $\mathscr{E}$ has a grid-like topology such that vertices along row t represent the segmentation contour of frame t, and edges along each column represent correspondence between two points in consecutive frames, e.g. ($\mathbf{x}_{t,i}$, $\mathbf{x}_{t+1,i}$). For brevity, we denote the column (temporal) edges as $e^t \in \mathscr{E}^T$ and row (spatial) edges as $e^s \in \mathscr{E}^S$ such that $\mathscr{E}^T \bigcup \mathscr{E}^S = \mathscr{E}$. e^s represents intra-frame connectivity between two controls points and is used to regularize the assigned displacement vectors of neighbouring points within each contour. e^t represents inter-frame connectivity and is used to regularize the assigned displacement vectors of a particular control point across time.

As we introduced previously, the optimality of assigning a displacement vector to a vertex will be measured by data and regularization energies. The data energy encourages each displaced control point $\mathbf{x}_{t,i}$ to become attached to image edges while the regularization energies ensure each contour remains smooth and continuous and that contours of adjacent frames move coherently. With a graph-labeling approach, the optimality of $\mathbf{D}$ can then be easily captured by the MRF energy of labeling $\mathscr{V}$. The unary, pairwise, and ternary potentials of the MRF energy represent external, temporal regularization, and spatial regularization energies, respectively.

Specifically, in edge-based formulations [4, 21, 3, 18], a unary potential may be designed to capture image forces that attract a control point $\mathbf{x}_{t,i}$ to locations with high image gradient. In this case, we wish to penalize the assignment of $\mathbf{d}_{t,i}$ to $\mathbf{x}_{t,i}$ if $\mathbf{d}_{t,i}$ displaces $\mathbf{x}_{t,i}$ to locations of low image gradients:

$$E_{image}(\mathbf{x}_{t,i}, \mathbf{d}_{t,i}) = \exp(0.2|\nabla I(\mathbf{x}_{t,i} + \mathbf{d}_{t,i})|^{-1}) \qquad (1)$$

Conversely, in template-based, feature-matching formulations, where image features around a contour estimate are matched to those of a template, the unary potential may instead be based directly on image-features. Thus, if $\mathbf{F}(\mathbf{x}_{t,i}, t)$ denotes a set of image features extracted at $\mathbf{x}_{t,i}$ in frame t, then we could treat the input contour $\mathbf{x}_{0,i}$ as the template and seek to minimize the dissimilarity between $\mathbf{F}(\mathbf{x}_{0,i}, 0)$

and $\mathbf{F}(\mathbf{x}_{t,i},t)$. Furthermore, if the dissimilarity between two feature sets is defined as their L^2-norms, then each unary potential may alternatively be defined as:

$$E^F_{image}(\mathbf{x}_{t,i}\ ,\ \mathbf{d}_{t,i}) = \|\mathbf{F}(\mathbf{x}_{0,i},0) - \mathbf{F}(\mathbf{x}_{t,i} + \mathbf{d}_{i,t},t)\| \qquad (2)$$

The appropriateness of these alternative data terms is explored further in Sec. 3.

Similar to the approach of [4], pairwise potentials are used to impose temporal constraints over the segmentation contours in two consecutive frames and is defined as:

$$E_{temp}(\mathbf{x}_{t,i}\ ,\ \mathbf{x}_{t+1,i}\ ,\ \mathbf{d}_{t,i}\ ,\ \mathbf{d}_{t+1,j}) = \|\mathbf{d}_{t,i} - \mathbf{d}_{t+1,i}\|, \qquad (3)$$

which measures the L^2-norm between two displacement labels that have been assigned to temporal neighbours $\mathbf{x}_{t,i}$ and $\mathbf{x}_{t+1,i}$ that are connected by e^t.

Finally, ternary potentials are used to impose smoothness on each segmentation contour by measuring the amount of bending due to three displaced neighbouring points on a given contour estimate, and is estimated as [4, 3]:

$$E_{spat}(\mathbf{x}_{t,i}\ ,\ \mathbf{x}_{t,j}\ ,\ \mathbf{x}_{t,k}\ ,\ \mathbf{d}_{t,i}\ ,\ \mathbf{d}_{t,j}\ ,\ \mathbf{d}_{t,k}) = 1 - \frac{\mathbf{u}_{ij} \cdot \mathbf{u}_{jk}}{|\mathbf{u}_{ij}||\mathbf{u}_{jk}|} \qquad (4)$$

where $(\mathbf{x}_{t,i}, \mathbf{x}_{t,j}, \mathbf{x}_{t,k})$ is an ordered triplets of spatially adjacent vertices connected by e^s and $\mathbf{u}_{ij} = \mathbf{x}_{t,j} + \mathbf{d}_{t,j} - \mathbf{x}_{t,i} - \mathbf{d}_{t,i}$.

The overall MRF energy representing the above energies of the set of contour segmentations implied by $\mathbf{D}$ is then defined as:

$$\begin{aligned} E(\mathbf{D}) = &\sum_{\mathbf{x}_{t,i} \in V} \alpha\, E_{image}(\mathbf{x}_{t,i}\ ,\ \mathbf{d}_{t,i}) \\ &+ \sum_{(\mathbf{x}_{t,i}\ ,\ \mathbf{x}_{t+1,i}) \in \mathscr{E}^T} \beta\, E_{temp}(\mathbf{x}_{t,i}\ ,\ \mathbf{x}_{t+1,i}\ ,\ \mathbf{d}_{t,i}\ ,\ \mathbf{d}_{t,j}) \\ &+ \sum_{(\mathbf{x}_{t,i}\ ,\ \mathbf{x}_{t,j}\ ,\ \mathbf{x}_{t,k}) \in \mathscr{E}^S} \gamma\, E_{spat}(\mathbf{x}_{t,i}\ ,\ \mathbf{x}_{t,j}\ ,\ \mathbf{x}_{t,k}\ ,\ \mathbf{d}_{t,i}\ ,\ \mathbf{d}_{t,j}\ ,\ \mathbf{d}_{t,k}) \end{aligned} \qquad (5)$$

where α, β, γ are weights for the respective energies.

In solving for $\mathbf{D}$, from which we compute the set of all segmentation contours, we minimize (5) using the graph-cuts optimization technique of Ishikawa [12].

By formulating the contour tracking problem as a graph-labeling problem and representing intra-frame and inter-frame relationships via e^s and e^t, we can now track the entire US sequence without reinitialization of the contour. The entire tracking problem is now performed in a single optimization procedure, that ensures an exact and global solution [12]. Furthermore, as we explain in the next section, our framework easily allows for adaptive regularization such that contour tracking in unreliable, noise-corrupted frames can now be guided by the more reliable adjacent frames.

2.2. Adaptive temporal regularization

As motivated in the introduction, image features like ridges and edges cannot be reliably extracted in US images as the tongue contour may partially or completely disappear in a frame. In these cases, we advocate the idea of adaptively adjusting the amount of temporal regularization such that when the region of a control point $\mathbf{x}_{t,i}$ lacks informative image features, meaningful information available in adjacent frames is used instead.

In our formulation, an increased emphasis on temporal information is effectively done by increasing the amount of temporal regularization to be enforced on a graph node such that its label assignment is forced to become similar to the assignments obtained in adjacent frames. Encoding adaptive regularization in a graph-based approach is done by adjusting the weights of edges in $\mathscr{E}^T$ according to the image features available around the local region of $\mathbf{x}_{t,i}$. Specifically, for every node $\mathbf{x}_{t,i}$, we examine the distribution of all unary potentials of $\mathbf{x}_{t,i}$ and assign a high weight when its variance σ is high or assign a low weight otherwise:

$$\lambda(\mathbf{x}_{t,i}) = \eta_1 \sqrt{\sigma(\mathbf{x}_{t,i})} + \eta_2 \qquad (6)$$

where η_1 and η_2 are constants used to rescale λ to [0,1]. The weight β in (5) is then replaced by λ.

2.3. Estimating data penalty with local-phase gradients or image features

What remains to be detailed in (5) is how exactly we calculate E_{image}. We explored two ways of capturing image-based penalty: one that attracts segmentation estimates to image edges and one that match features extracted around estimates to those extracted from a reference contour. In the former, instead of calculating the gradients on the original images, we calculated the gradients on their local-phase features, which can be interpreted as a qualitative description of salient regions in images such as edges or ridges that are invariant to changes in illumination or image contrast [14, 26, 21]. These are based on a model that postulates that image features are perceived at points in an image where the Fourier components are maximal in phase. These features have shown to be useful for boundary detection of LV in echocardiographs [21] and recently proposed for bone segmentation and fracture detection in US images [11]. In extracting these features, we employ the implementation of [15] that uses monogenic filters.

In the feature-based approach (Section 1), we adopted the work of [17], where Leung et al. extracted the following set of attributes for feature-based registration of US liver images: pixel intensity, gradient magnitude and the Laplacian of Gaussian (LoG). According to the authors, LoG helps locate features that surround smooth or faded soft transitions while the gradient magnitude helps detect image edges and ridges. Based on our experiments, we found that matching the gradient and LoG features extracted with a 3×3 spatial mask worked best. Prior to feature extraction, we applied anisotropic diffusion filtering using the approach of [8] where a linear model was used to control the gradient weight of the diffusion function. This allowed us to obtain more reliable features.

3. Results

3.1. Data acquisition & groundtruth segmentations

A General Electric Logiq Alpha 100 MP ultrasound scanner (General Electric Medical Systems, Milwaukee, Wisconsin) with a model E72 6.5 MHz transducer that uses a 114° microconvex array was used for acquisition of 8 datasets. The tongue movement of each participant was recorded using the protocol described in [23]. The video output from the US scanner was digitized to a video camera with a capture rate of 30 frames per second (fps) at a resolution of 240×320 pixels, each of size 0.48^2 mm^2.

Two data sets were created from the collected sequences. The first, denoted by Φ_{dense}, contains 2 US sequences randomly selected from the entire set. For every sequence in this set, a dense set of 60 segmentations was created (every 2^{nd} frame was segmented) to allow for precise assessment of tracking accuracy. The second, denoted as Φ_{sparse}, contains the remaining 6 sequences from which only a sparse set of segmentations were created (every 5^{th} to 8^{th} frame). This set was used to assess the robustness of tracking.

Two experts were recruited to manually create the segmentations; each created either Φ_{dense} or Φ_{sparse}. For each frame of an US sequence, a segmentation was manually created by positioning 15 to 23 points which the expert believed would represent the contour of the tongue. A total of 16-23 frames were segmented per sequence.

Intra and inter-rater reliability of the experts were assessed by repeating the extraction session on a segment of a data set. The intra-rater mean measurement error of the expert who created Φ_{dense} was found to be 0.92 mm (0.72 mm std) while the mean error of the other expert was found to be 0.77 mm (0.69 mm std). Inter-rater difference was statistically insignificant (mean positional difference of 0.92mm ±1.1 std). Because the accuracy of the US scanner was estimated to be within 61 mm and that no single intra-rater mean measurement error exceeded 3 mm, their segmentations were regarded as acceptable and thereafter treated as groundtruth.

3.2. Validation

Following [24], we performed validation by comparing the tracked contours obtained by the proposed method with the groundtruth segmentations. The segmentation error on each tracked contour was defined as the mean Euclidean distance (MED) between each point on a segmentation solution and the closest point on the groundtruth contour.

We used the QPBO-based graph-cuts optimization algorithm of Ishikawa [12]. Due to the small size of the graph (< 30 control points for every frame in $T < 60$ frames), the run-time of the algorithm depended mostly on the size of label set (label set of displacement vectors). In defining this label set, we sampled the $\Re^2$ space in polar coordinates within a bounded range. The discretization of this space is thus parameterized by the number of steps of along the radial axis, N_{rad}, and the number of steps of along the angular axis, N_{ang}, and the maximum radius R_{max}. For 2D US sequences, we observed that the maximal displacement required is no more than half the size of the image width. Thus, for all our experiments, we set $R_{max} = 30$ mm, $N_{ang} = 8°$, and $N_{rad} = 2$ mm to obtain < 100 labels. For a set of 150 labels, the optimization algorithm converged usually in < 2 min, when run on a PC with 2.66 GHz Intel® Core™ 2 Duo CPU. The overhead on the overall run-time of our method thus chiefly depended on the calculation of the unary potentials (pre-filtering and feature-extraction).

We first examined the effects of different parameters on the obtained solutions in the case of uniform temporal regularization. Figure 1 presents qualitative and quantitative results generated when the amounts of spatial and temporal regularization were varied. Not surprisingly, the sensitivity of the results on the choice of parameters can be high, since these effectively change the MRF energy and its global minimum solution. As expected, when low spatial regularization was enforced, the obtained segmentation contours were more irregular. Similarly, the higher the temporal regularization was, the smoother the contour evolution over time (Figure 2).

Based on initial experiments, we empirically determined a range for each of these values for subsequent tests: $\alpha = [0.3, 0.8]\frac{1}{6}$, $\beta = [0.3, 0.8]\frac{1}{3}$, and $\gamma = [0.3, 0.8]\frac{1}{2}$. In the case of adaptive temporal regularization (ATR) (Section 2.2), the only parameter that required tuning is γ (as λ, which balances between temporal and the data term, is a function of $\mathbf{x}_{t,i}$). In these cases, we used $\gamma \in [0.2, 0.5]$.

With these parameters empirically determined, we performed validation experiments on our method, testing four variations of the energy in (5): 1) with E_{image} as the data term (gradient of local phase feature) with ATR; 2) E_{image} without ATR; 3) E^F_{image} as the data term (feature-based) with ATR; and 4) E^F_{image} without ATR. For each of the above variants, we repeated segmentations on the datasets, with each trial using a different parameter setting (e.g. different combinations of α, β, and γ). Results of these experiments will be presented shortly.

For comparative analysis, we also performed segmentations with two other methods. The first method employed

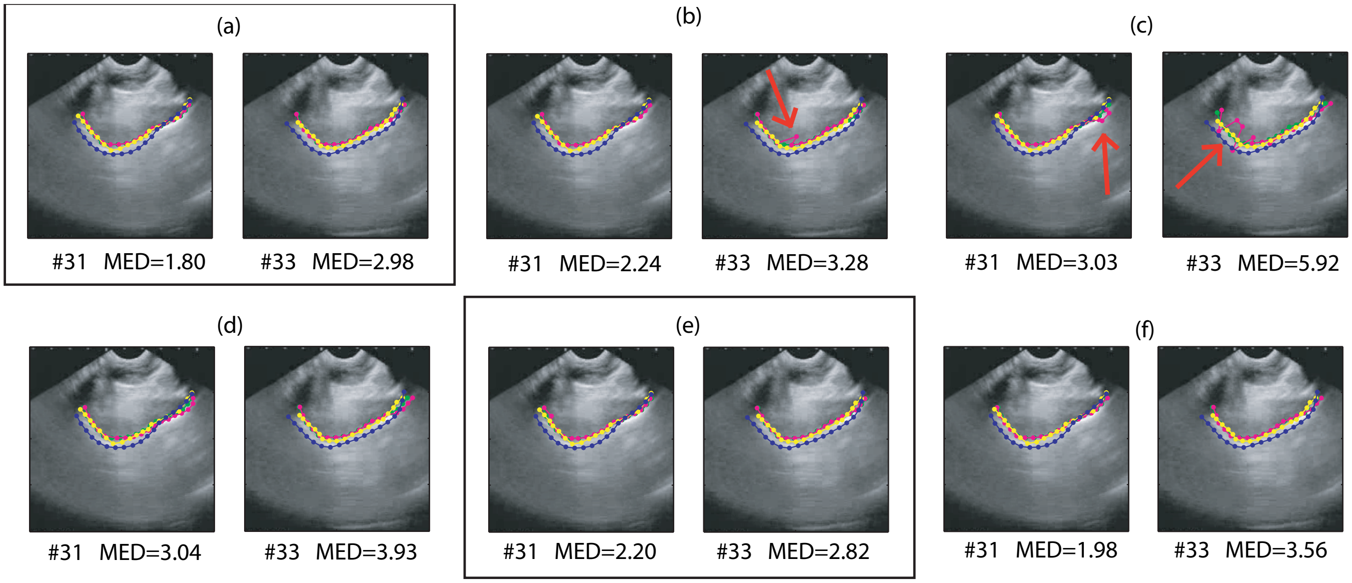


Figure 1. Segmentations of pairs of two consecutive frames (frame number and errors are given below each image). Contours obtained by our method are shown in magenta; EdgeTrak in blue; and groundtruth in yellow. The top row (a-c) depicts the effect of changing the amount of spatial regularization in our method; β is varied with fixed $\alpha = 0.8$ and $\beta = 0.2$: (a) $\gamma = 0.8$, (b) $\gamma = 0.2$, (c) $\gamma = 0.1$. The bottom row (d-f) shows the effect of changing the amount of temporal regularization with $\alpha = 0.8$ and $\gamma = 0.4$: (d) $\beta = 0.1$, (e) $\beta = 0.6$, (f) $\beta = 1$. Note that the smoothness of the contour is ensured in (a), but not in (b) and (c), as shown with red arrows, due to insufficient regularization. For (d-f), because of higher temporal regularization, the inter-frame difference between the two contours in (e) is more subtle than that in (d). The bounded cases appear to reflect the optimality of the respective parameters. The errors of EdgeTrak are 5.53 mm and 7.31 mm for the respective frames.

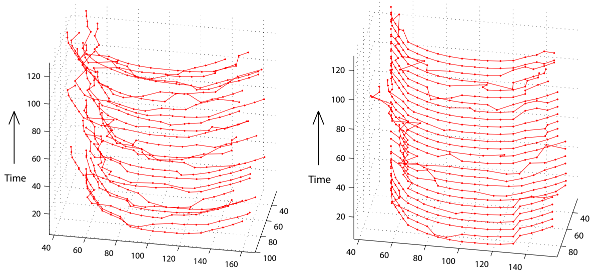


Figure 2. Contours extracted from a US sequence are stacked along the vertical time axis. Results obtained with (left) minimal temporal regularization ($\beta = 0.05$) and (right) adequate temporal regularization ($\beta = 0.48$).

was the snake-based EdgeTrack[1] software developed by Li et al. [18]. The second method we performed was pairwise diffeomorphic demons registration using the algorithm developed by Vercauteren et al. [25] that is based on histogram-matching of local regions. As the latter accounts for spatial uncertainty on pixel-wise correspondences, we believe that it is more robust than the algorithms presented in [2], where optical-flow registrations were performed on consecutive US frames. Results on pairwise registration will help us approximate the performance of the correlation-based optical-flow method of [8].

In the demons algorithm, the parameters involved are $\zeta_{demons1}$ and $\zeta_{demons2}$, which respectively control the amount of smoothing applied on the update field and on the deformation field [25]. These were tuned empirically ($\zeta_{demons1} = [1,5]$ and $\zeta_{demons2} = [2,4]$). For EdgeTrak, the only parameter needed for tuning was the balance, ζ_{ET}, between the smoothness and image terms. Based on initial experiments on 4 US sequences, we found that EdgeTrak performed best when $\zeta_{ET} \leq 0.1$ and failed miserably when $\zeta_{ET} > 0.3$. We also found EdgeTrak highly sensitive to the number of points provided by the input contour; its performance stabilized only when we interpolated the input segmentation to twice its original resolution. Thus, we did this interpolation for all subsequent tests and repeated the tracking session with $\zeta_{ET} = \{0.01, 0.03, 0.05, 0.1, 0.3\}$.

In all methods, only the contour extracted from the first frame of the groundtruth segmentations of each sequence was used as input; the groundtruth segmentations of all subsequent frames were used for validation.

Overall accuracies of all methods evaluated using Φ_{dense} are presented in Table 1. For all methods, only the registration trial performed on each US sequence that gave the lowest MED was used in the comparison. Note that the results of EdgeTrak had the highest MED while the demons

[1] We could not validate against the work of [24,5] as we do not have additional multi-modal data (e.g. electromagnetic sensors or X-ray images).

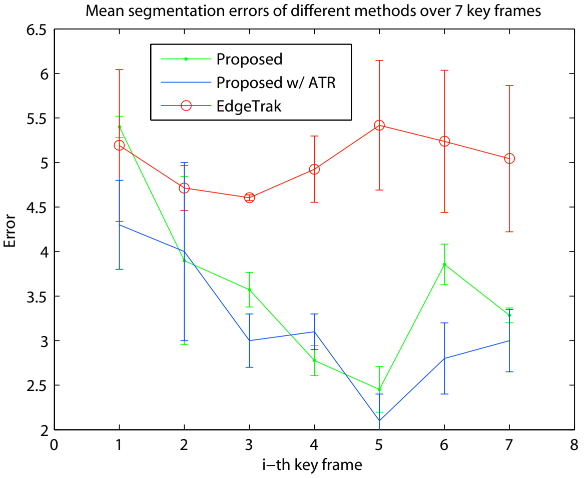


Figure 3. Mean segmentation errors obtained by the proposed method (with and without ATR) and EdgeTrak calculated over 7 key frames. Note the accuracy of our approach (green) is higher than that of EdgeTrak (red), with our method's accuracy slightly improved when ATR was used (blue).

Method	Max	Min	Mean	Std
E_{image} (Eq. 1)	**7.81**	1.02	4.70	2.11
E_{image} (Eq. 1) + ATR	8.19	**0.98**	**4.49**	**1.62**
EdgeTrak	16.34	3.46	6.64	3.75
Demons	9.34	1.73	5.67	4.25

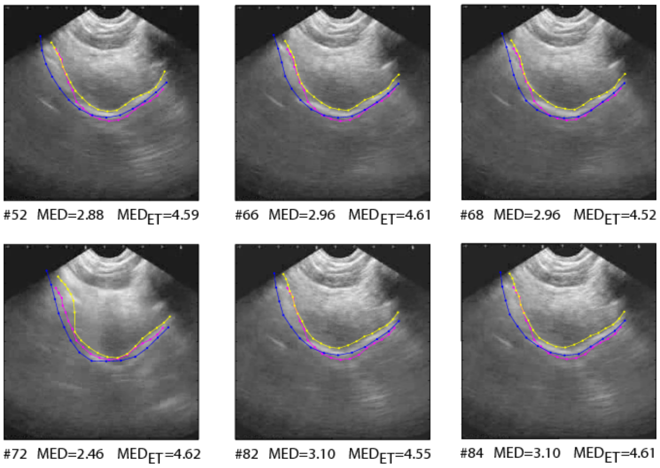
Table 1. Overall accuracy of different methods assessed using Φ_{dense}. Shown are the maximum, minimum, mean, and standard deviation of the measured errors over all frames in all sequences. Evidently, the accuracy generally improved when adaptive temporal regularization (ATR) was enforced. Bolded numbers indicate lowest errors per measurement group.

approach had larget variance in MED. The overall accuracies evaluated using Φ_{sparse} are next presented in Table 2. Due to the extended duration each sequence in Φ_{sparse} spanned, validation on this set included only our proposed method and EdgeTrak. We also visually compared the accuracy of individual frames in Figure 3, which shows that our method was also robust over time while the performance of EdgeTrak was least satisfactory. Figure 4 also shows an example comparison of qualitative results between the two methods.

In summary, our approach gave more accurate segmentations than those obtained with EdgeTrak and pairwise demons registration, and as demonstrated in our quantitative analyses (Table 1, Table 2, Figures 3 and 4), the use of adaptive temporal regularization further improved the accuracy of our method.

Method	Max	Min	Mean	Std
Phase-based (Eq. 1)	7.21	**0.92**	3.28	2.25
Phase-based + ATR	**6.03**	1.16	4.10	2.78
Feature-based (Eq. 2)	7.48	1.20	**2.23**	**1.47**
Feature-based + ATR	7.13	0.96	4.10	1.81
EdgeTrak	30.68	3.45	7.32	4.89

Table 2. Overall accuracy of different methods assessed using Φ_{sparse}. The first four rows show results obtained from different variants of our proposed method: phase-based versus feature-based (Section 2.3) and the use of uniform versus adaptive temporal regularization (ATR). Bolded numbers indicate lowest errors per measurement group.


#52 MED=2.88 MED_{ET}=4.59 #66 MED=2.96 MED_{ET}=4.61 #68 MED=2.96 MED_{ET}=4.52

#72 MED=2.46 MED_{ET}=4.62 #82 MED=3.10 MED_{ET}=4.55 #84 MED=3.10 MED_{ET}=4.61

Figure 4. Qualitative and quantitative comparions of the results generated using the proposed method (contour in magenta) and EdgeTrak (blue) on a sequence in Φ_{dense}. Frame number and errors are given below each image (error of results obtained by EdgeTrack is denoted as MED_{ET}). The groundtruth contour is coloured in yellow.

4. Conclusions and Future Work

In this paper, we have proposed and tested a graph-based approach for the semi-automatic extraction of the tongue contour in US image sequences that does not require re-initialization nor training data. Given an input contour, we reformulated contour-tracking as a graph-labeling problem where we seek to find a set of displacement vectors that would spatially map each segmentation solution to the initial contour in an optimal way. Optimality of the segmentations was then measured via the corresponding MRF energy where unary, pairwise, and ternary terms each respectively represents data, temporal regularization, and spatial smoothness energies. In exploiting temporal coherency exhibited in image sequences, we have also developed an adaptive regularization approach wherein the segmentation of a featureless region of an image frame is guided by information available in adjacent frames. We are currently examining ways to automatically adjust the involved parameters and working on incorporating texture-based image features to devise an even more robust data energy term.

5. Acknoweldgments

We are grateful to Dr. Tim Bressmann from the Department of Speech-Language Pathology, University of Toronto, for provision of the US data. We also thank Bianca Herold and Matt Toom for assistance in data preparation.

References

[1] S. Abolmaesumi, P. Abolmaesumi, M. R. Sirouspour, and S. E. Salcudean. Real-time extraction of carotid artery contours from ultrasound. In *Computer-Based Medical Systems*, pages 181–186, 2000.

[2] B. Achmad, M. M. Mustafa, and A. Hussain. Inter-frame enhancement of ultrasound images using optical flow. In *International Visual Informatics Conference*, pages 191–201, 2009.

[3] Y. Akgul and C. Kambhamettu. A coarse-to-fine deformable contour optimization framework. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25:174–186, 2003.

[4] Y. Akgul, C. Kambhamettu, and M. Stone. Extraction and tracking of the tongue surface from ultrasound image sequences. *IEEE Conference on Computer Vision and Pattern Recognition*, page 298, 1998.

[5] M. Aron, A. Roussos, M. odile Berger, E. Kerrien, and P. Maragos. Multimodality acquisition of articulatory data and processing. In *European Conference on Signal Processing*, 2008.

[6] T. Bressmann, C. Heng, and J. Irish. The application of 2D and 3D ultrasound imaging in speech-language pathology. *Speech-Language Pathology and Audiology*, 29:158–168, 2005.

[7] L. Davidson. Comparing tongue shapes from ultrasound imaging using smoothing spline analysis of variance. *Acoustical Society of America*, 120(1):407–415, 2006.

[8] Q. Duan, E. Angelini, and O. Gerard. Comparing optical flow based methods for quantification of myocardial deformations on RT3D ultrasound. In *IEEE International Symposium on Biomedical Imaging*, pages 173–176, 2006.

[9] D. Freedman and M. W. Turek. Illumination-invariant tracking via graph cuts. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 10–17, 2005.

[10] N. Friedland and D. Adam. Automatic ventricular cavity boundary detection from sequential ultrasound images using simulated annealing. *IEEE Transactions on Medical Imaging*, 8(4):344–353, 1989.

[11] I. Hacihaliloglu, R. Abugharbieh, A. Hodgson, and R. Rohling. Bone segmentation and fracture detection in ultrasound using 3D local phase features. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, pages 287–295, 2008.

[12] H. Ishikawa. Higher-order clique reduction in binary graph cut. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 20–25, 2009.

[13] K. Iskarous. Detecting the edge of the tongue: A tutorial. *Clinical Linguistics and Phonetics*, 19(6):555–565, 2005.

[14] P. Kovesi. Image features from phase congruency. *Videre: A Journal of Computer Vision Research*, 1(3):2–26, 1999.

[15] P. D. Kovesi. MATLAB and Octave functions for computer vision and image processing. URL: <http://www.csse.uwa.edu.au/∼pk/research/matlabfns/>.

[16] M. J. Ledesma-carbayo, J. Kybic, M. Desco, A. Santos, S. Member, M. Shling, S. Member, P. Hunziker, and M. Unser. Spatio-temporal nonrigid registration for ultrasound cardiac motion estimation. *IEEE Transactions on Medical Imaging*, 24(9):1113–1126, 2005.

[17] C. Leung, K. Hashtrudi-Zaad, P. Foroughi, and P. Abolmaesumi. A real-time intrasubject elastic registration algorithm for dynamic 2-D ultrasound images. *Ultrasound in Medicine and Biology*, 35(7):1159 – 1176, 2009.

[18] M. Li, C. Kambhamettu, and M. Stone. Automatic contour tracking in ultrasound images. *Clinical Linguistics and Phonetics*, 19(6):545–554, 2005.

[19] S. Mitchell, J. Bosch, B. Lelieveldt, R. van der Geest, J. Reiber, and M. Sonka. 3-D active appearance models: segmentation of cardiac MR and ultrasound images. *IEEE Transactions on Medical Imaging*, 21(9):1167–1178, 2002.

[20] K. E. Nielsen. Practical use of block-matching in 3D speckle tracking. Master's thesis, Norwegian University of Science and Technology, Department of Computer and Information Science, 2009.

[21] J. A. Noble and D. Boukerroui. Ultrasound image segmentation: a survey. *IEEE Transactions on Medical Imaging*, 25(8):987–1010, 2006.

[22] C. Qin, M. Carreira-Perpin, K. Richmond, A. Wrench, and S. Renals. Predicting tongue shapes from a few landmark locations. In *Proc. Interspeech*, pages 2306–2309, 2008.

[23] O. Rastadmehr, T. Bressmann, R. Smyth, and J. C. Irish. Increased midsagittal tongue velocity as indication of articulatory compensation in patients with lateral partial glossectomies. *Head and Neck*, 30(6):718–726, 2008.

[24] A. Roussos, A. Katsamanis, and P. Maragos. Tongue tracking in ultrasound images with active appearance models. In *IEEE International Conference on Image Processing*, pages 1733–1736, 2009.

[25] T. Vercauteren, X. Pennec, A. Perchant, and N. Ayache. Diffeomorphic demons using itk's finite difference solver hierarchy. In *Insight Journal – ISC/NA-MIC Workshop on Open Science at MICCAI 2007*, October 2007. Source code available online.

[26] J. Woo, B.-W. Hong, C.-H. Hu, K. K. Shung, C. C. Kuo, and P. J. Slomka. Non-rigid ultrasound image registration based on intensity and local phase information. *J. Signal Process. Syst.*, 54(1-3):33–43, 2009.

[27] N. Xu, N. Ahuja, and R. Bansal. Object segmentation using graph cuts based active contours. *Computer Vision and Image Understanding*, 107(3):210–224, 2007.

[28] L. Yang, B. Georgescu, Y. Zheng, P. Meer, and D. Comaniciu. 3D ultrasound tracking of the left ventricle using one-step forward prediction and data fusion of collaborative trackers. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.