ENHANCED SNAKE BASED SEGMENTATION OF VOCAL FOLDS

Sonya Allin[1]*, John Galeotti*[2]*, George Stetten*[2]

[1]HCII, [2]Robotics Insitute
Carnegie Mellon University
Pittsburgh, PA

Seth H. Dailey

Brigham and Women's Hospital
Harvard Medical School
Boston, MA

ABSTRACT

We present a system to segment the medial edges of the vocal folds from stroboscopic video. The system has two components. The first learns a color transformation that optimally discriminates, according to the Fisher linear criterion, between the trachea and vocal folds. Using this transformation, it is able to make a coarse segmentation of vocal fold boundaries. The second component uses an active contour formulation recently developed for the Insight Toolkit to refine detected contours. Rather than tune the internal energy of our active contours to bias for specific shapes, we optimize image energy so as to highlight boundaries of interest. This transformation of image energy simplifies the contour extraction process and suppresses noisy artifacts, which may confound standard implementations.

We evaluate our system on stroboscopic video of sustained phonation. Our evaluation compares points on automatically extracted contours with manually supplied points at perceived vocal fold edges. Mean deviations for points located on the minor axes of the vocal folds averaged 2.2 pixels across all subjects, with a standard deviation of 3.6.

1. INTRODUCTION

Visual evaluation of the vocal folds plays an important role in the diagnosis of laryngeal disorders. Stroboscopy can reveal critical asymmetries in the pliability of the vocal folds and can help rate glottal closures. Parameterizing this oscillation is the subject of ongoing medical research.

Direct observations are typically made with a stroboscopic endoscope and recorded on video at a rate of 30 and 60 frames per second. Because the vocal folds can vibrate faster than most camera frame rates, a strobe light is used to guarantee that complete cycles of oscillation are captured. The light of this strobe varies slightly in phase from the vibration of the folds, thereby reducing aliasing in resulting stills. Although the images can capture the temporal dynamics of the vocal folds, they can be low contrast and subject to noise. This noise and lack of contrast present challenges to the reliable automatic extraction of fold contours.

Many approaches to vocal fold segmentation make use of active contours with shape priors [1][2]. Generally, priors are used to modify internal energy in the active contour formulation, which is then paired with external energy derived from image pixels. Although priors can yield compelling segmentation results, they can be time consuming to craft and may not generalize well. Shapes of folds during oscillation, for example, rarely manifest during periods of extreme arytenoid abduction, and shapes from healthy individuals may be very different from those with injuries or illnesses. In addition, very accurate shape priors may still become overwhelmed by pervasive image noise.

Our work seeks to enhance the power of existing segmentation methods by focusing, not on the internal energy of snakes, but on external image energy instead. By means of a simple training procedure, we learn a color transformation which optimally separates the pixels in the trachea from pixels on the vocal folds, as well as pixels on the vocal folds from pixels on surrounding tissues. Optimality, in this formulation, is judged relative to the Fisher linear criterion. By leveraging the discriminative power of a learned transformation, we quickly make reasonably accurate segmentations of fold contours in previously unseen images. These segmentations are subsequently refined with an active contour formulation based on the Fourier series.

The Fisher criterion has been employed previously to learn a discriminative global image transformation in formulations such as FisherFaces [3]. Operating locally, Charmicael and Hebert [4] describe a technique wherein textured filter responses are used to separate regions that correspond to an object of interest from "clutter". Our problem permits some simplification. The image regions we consider are more uniform in appearance than components of an object or "clutter," and only subtly textured. As a result, we rely principally on color information to populate our feature space, and still achieve reasonable discrimination results. The inspiration for the set of color transformations we employ is found in recent work of Collins and Liu [5], who use the Fisher criterion to segment background from foreground

This work was supported by the first two authors' NSF Graduate Research Fellowships, NSF IGERT DGE-0333420, NLM and NIBIB

for the purpose of online tracking.

2. METHODS

Active Contours, first introduced by Kass and Witkin[6], are a popular segmentation tool for medical images. Snakes have the desirable property of being able to enforce continuity and closure, allowing them to precisely represent boundaries when portions of the boundaries are obscured. The accurate automatic placement of a contour in an image is typically achieved by coupling terms describing image "energy" with terms that govern a contour's stiffness or elasticity, called internal energy. Internal energy can be composed of heuristic constraints or be learned through training. Image "energy", by contrast, is often formed of intensity-based edge energies, or some measure of coherence in a postulated segmentation.

In our work, we suggest that a good segmentation can be simply achieved by focusing on the construction of image energy. By transforming image energies to facilitate region classification, we extract contours without training priors on shape.

2.1. Segmentation of Vocal Folds

Our boundary extraction method determines a linear combination of image attributes which optimally discriminate between different regions of the larynx. More specifically, we seek a linear function of image attributes that maximizes the difference between image samples taken from different sections of the larynx, while simultaneously minimizing the difference between sample responses from the same portion of the larynx.

In the two class case, this means discovering some scalar function of pixels, $\Phi(x)$, which maximizes the following term:

$$\frac{\frac{1}{N}\sum_{x_i \epsilon A}\sum_{x_j \epsilon B}\|(\Phi(x_i)-\Phi(x_j))\|^2}{\frac{1}{M}\sum_{C \epsilon A,B}\sum_{x_i \epsilon C}\sum_{x_j \epsilon C}\|(\Phi(x_i)-\Phi(x_j))\|^2} \quad (1)$$

The numerator represents the scatter of elements between classes, and the denominator captures scatter within-classes. N is the number of pairs that are composed of one member from either class, while M is the number of pairs that can be made of all members combined.

$\Phi(x)$ we define to be composed of a linear combination of various individual pixel statistics, as follows:

$$\Phi(x) = \alpha^T \phi(x) \quad (2)$$

where

$$\alpha = [\alpha_1, \alpha_2, ...\alpha_n]^T; \phi(x) = [\phi_1(x), \phi_2(x), ...\phi_n(x)]^T$$

Coefficients, or α, which define $\Phi(x)$ can be found by solving the generalized eigenvalue problem corresponding to equation (1), as in [4].

As in [5], we choose to use linear combinations of R, G and B color values as individual features, $\phi_i(x)$, at each pixel. We could, however, chose to evaluate a much broader range of image features, including texture cues or various nonlinear combinations of texture and color. For our task, however, simple image features suffice, and so we compose $\phi_i(x)$ as follows:

$$\phi_i(x) = \omega_1 \times R + \omega_2 \times G + \omega_3 \times B; \omega_* \epsilon [-1, 0, 1] \quad (3)$$

The final process by which trachea and vocal folds are segmented from input images can now be decomposed into a simple training and classification procedure. During training, three images at the begining of a sequence are selected and regions of interest are segmented by hand. Color-based pixel features are computed for each region, and optimal coefficients, α, that discriminate between regions are determined. In addition for each class, the mean and variance of embedded training samples ($\mu(A), \sigma(A)$ and $\mu(B), \sigma(B)$) are stored.

Classification is then performed of the remaining images in a given input sequence. For each input image, color features are determined and projected onto α. At every input pixel we determine the ratio of log probabilities:

$$log(P(\alpha^T \phi(x)|A)) - log(P(\alpha^T \phi(x)|B)) \quad (4)$$

If we assume a uniform distribution for $P(A)$ and $P(B)$, this corresponds to the Bayes decision criterion, whose decision boundary lies at zero. Values greater than zero indicate a pixel's membership with the those of class A, while values less than zero indicate probable membership in class B. $P(\alpha^T \phi(x)|A)$ and $P(\alpha^T \phi(x)|B)$ are defined by normal distributions with sufficient statistics ($\mu(A), \sigma(A)$ and $\mu(B), \sigma(B)$).

In practice, we threshold classified pixel values that lie under zero and scale remaining values so they lie between 0 and 255. This produces a grayscale image, the gradient of which is refined with active contours. The boundary surrounding pixels in a highlighted region forms an initial guess at a contour for that region. In our experiments, we create two boundaries to optimize for every image that define the vocal cord.

2.2. Refinement with Active Contours

Like our classification procedure, our active contour implementation is similarly driven by image energy as opposed to priors on shape. The basic algorithm used was developed by Stetten and Drezek[7] and recently added to the architecture of the Insight Toolkit by Galeotti.

Using dynamic programming, orthogonal offsets (or corrections) which maximize image energy are located with re-

spect to an input contour at evenly spaced intervals. To facilitate the search for these offsets, "swaths" around a given contour are formed by traversing the input path and interpolating image pixels orthogonal to the path at regularly spaced intervals.

In order to ensure that the normal of the input contour is always well defined, the initial path is Fourier smoothed before execution. Fourier smoothing eliminates high order coefficients in frequency space, thereby removing contour spikes [7]. Representing our curves in Fourier space also has the desirable property that the normals to a contour are easily formulated, which facilitates sub-pixel resolution. Smoothness constraints are additionally imposed by forbidding offsets of adjacent pixels that differ by more than one.

In our application, closure of the vocal cords is detected during the refinement process by exploring the image region bounded by an input contour. Vocal cords are only considered open if the count of interior pixels labeled "trachea" exceed a specified threshold.

3. EXPERIMENTS

We evaluate our system on three individuals' stroboscopic video taken during sustained phonation. Video was taken with a rigid endoscope at a rate 30 frames per second and converted to uncompressed digital images, each 320 by 240 pixels. Two sequences were 120 frames in length, while the third was 840, and contains periods of both phonation and arytenoid abduction.

For each video sequence, three frames (less than 1 percent of all data) were selected to train a discriminating color transformation. Vocal folds were segmented in these images and two classifiers for each sequence were trained. The first was trained to distinguish between the trachea and pixels within a 50 pixel radius of the trachea's center of mass. The second was trained to distinguish between the surface of vocal folds and pixels surrounding them, also within a 50 pixel radius.

All remaining frames were automatically classified to create pairs of grayscale images depicting classification weights, as in Figure 1. Positively classified regions were then segmented into connected components, and components sharing overlap with prior segmentations were followed through the entirety of each sequence. Finally, boudaries were refines with our active contour formulation to yield final results.

4. RESULTS

As can be seen in Figure 2, thresholding each classification image achieves a reasonable first approximation to regions of interest. Active contours are able to refine this first approximation and delete spurious pixels at borders. Figure 3 depicts example outputs of the contour refinement procedure in cases where initial segmentations based on thresholds were flawed. As can be seen, active contour refinement is often capable of achieving a more accurate segmentation.

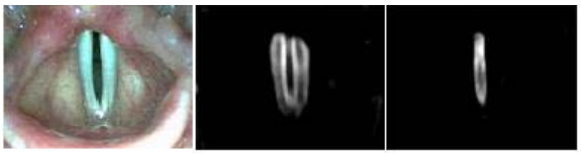

Fig. 1. Input image, classification weights for "folds" and "trachea".

Fig. 2. Input sequences and output segmentations based on simple thresholding. Rows alternate between input images and output segmentations.

As an additional form of evaluation, 30 frames from each sequence were selected at random from periods in videos corresponding to sustained phonation and a series of 6 points were placed on perceived boundaries of vocal folds. One point was on the anterior commissure of the vocal folds and another at the point maximally opposite it. Remaining points were placed approximately equidistant to these points on visual boundaries surrounding the vocal folds. We then compared the distance between manual points and corresponding automatically detected contours.

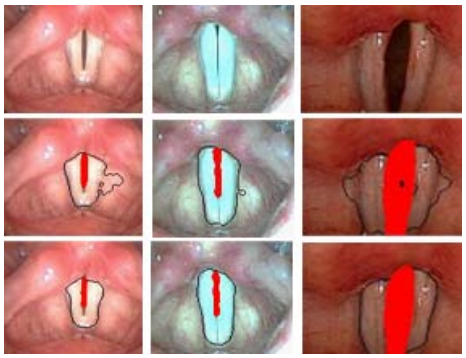

Fig. 3. Top row shows input images, second row shows thresholded images, and the bottom row depicts thresholded contours refined.

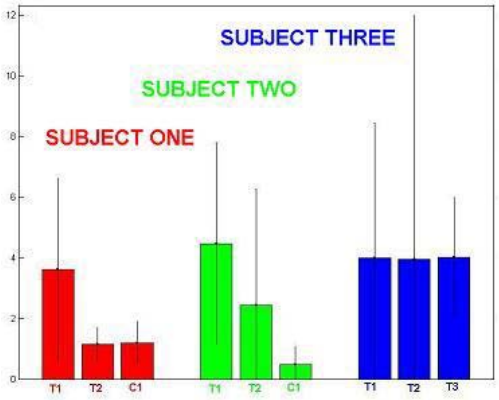

Fig. 4. Deviation between manually and automatically extracted points. T1 indicates points on the major internal axis, T2 indicates points on the minor internal axis and C1 on the minor external axis.

Figure 4 illustrates the results of the comparison. Mean deviations for points located on the minor axes of trachea and folds averaged 2.2 pixels across all subjects, with a standard deviation of 3.6. Deviation around the major axis of the trachea averaged 4.0 pixels across all subjects, with a standard deviation of 3.6. Note that pixels on the major axis of the boundary between vocal folds and trachea exhibit more error than those on the visual boundary between vocal folds and surrounding tissue. This is due to the fact that folds, when closed, exhibit no discernible opening to the trachea and thus accurate segmentation is difficult. In addition, errors for subject three are larger than those for subject one and two which, in large part, is a result of the relative scale of the regions of interest with respect to the image boundary.

5. DISCUSSION

Most techniques for the segmentation of vocal cords rely on prior shape terms to perform a visual segmentation vocal cords from stroboscopic endoscopic images. Here, we have demonstrated the capability to segment vocal cords without heavy reliance on a shape model that may not be appropriate in pathological cases. While this work tests our algorithm on muscular membraneous vocal fold oscillation, we expect it to extend to arytenoid motion. Eventually, we seek to use our segmentation tool for the evaluation of gross laryngeal movement disorders.

6. ACKNOWLEDGEMENTS

The authors would like to thank James Kobler, David Talliver and Chris Atkeson for feedback, as well as the Insight Consortium and everyone who has contributed to ITK.

7. REFERENCES

[1] B. Marendic, N. Galatsanos, and D. Bless, "A new active contour algorithm for tracking vocal folds," in *Proc. IEEE Int. Conf. on Image Processing*, 2001.

[2] C. Palm, T. Lehmann, Bredno, S. Neuschaefer-Rube, Klajman, and K. Spitzer, "Automated analysis of stroboscopic image sequences by vibration profile diagrams," in *5th International Conference on Advances in Quantitative Laryngoscopy, Voice and Speech Research*, 2001.

[3] P. Belheumer, J. Hespanha, and D. Kriegman, "Eigenfaces vs. fisherfaces:recognition using class specific linear projection," *Proceedings of the European Conference On Computer Vision*, 1996.

[4] O. Carmichael and M. Hebert, "Discriminant filters for object recognition," Carnegie Mellon University, CMU-RI-TR-02-09, Tech. Rep., 2002.

[5] R. Collins and Y. Liu, "On-line selection of discriminative tracking features," in *Proceedings of the 2003 International Conference of Computer Vision*, 2003.

[6] Kass, Witkin, and Terzopoulos, "Snakes: Acitive contour models," in *Int. Journal of Computer Vision 1*, 1987.

[7] G. Stetten and R. Drezek, "Active fourier contour applied to real time 3d ultrasound of the heart," in *International Journal of Image and Graphics*, vol. 1(4), 2001, pp. 647–658.