Warp-based motion compensation for endoscopic kymography

David C. Schneider[1], Anna Hilsmann[1] and Peter Eisert[1,2]

[1]Fraunhofer Heinrich Hertz Institute, Berlin, Germany
[2]Humboldt Universität zu Berlin, Germany

Abstract

Endoscopic videokymography is a method for visualizing the motion of the plica vocalis *(vocal folds) for medical diagnosis. The diagnostic interpretability of a kymogram deteriorates if camera motion interferes with vocal fold motion, which is hard to avoid in practice. We propose an algorithm for compensating strong camera motion for videokymography. The approach is based on an image-based inverse warping scheme that can be stated as an optimization problem. The algorithm is parallelizable and real-time capable on the CPU. We discuss advantages of the image-based approach and address its use for approximate structure visualization of the endoscopic scene.*

Categories and Subject Descriptors (according to ACM CCS): I.3.8 [Computer Graphics]: Application— I.4.8 [Image Processing and Computer Vision]: Scene Analysis—Motion

1. Introduction and related work

The opening and closing of the vocal folds (*plica vocalis*) at high frequencies is a major source of sound in human speech. *Videokymography* [SS95] is a technique for visualizing the motion of the vocal folds for medical diagnosis: The vibrating folds are filmed with an endoscopic camera pointed into the larynx. The camera records at a very high framerate to capture vocal fold vibration. Alternatively, a low framerate and stroboscopic lighting at a frequency synchronized with the vibratory frequency of the vocal folds is used to obtain a temporal subsampling of the motion (see figure 2 for example frames). The *kymogram* used for medical diagnosis is essentially a time-slice image, i.e. an *X*-*t*-cut through the *X*-*Y*-*t* image cube of the endoscopic video (figure 3). The quality and diagnostic interpretability of a kymogram deteriorates significantly if the camera moves relative to the scene as this motion interferes with the vibratory motion of the vocal fold in the kymogram. Scene-to-camera motion caused by the patient or the operator of the endoscope is hard to avoid in medical practice. In this paper, we propose an approach to stabilizing the motion of endoscopic video for kymography.

This motion compensation problem is challenging and different from motion compensation of handheld video (e. g. [LGJA09, LCCO09, GL08]) in several respects: Firstly, the camera motion to be eliminated may be significantly larger than a typical camera shake due to the short disctance between camera and scene. Secondly, not only the camera and the vocal folds move but the entire scene may be highly nonrigid, for example when the *ariepiglottic fold* and the *cuneiform cartilage* move when the patient takes breath. Therefore, a 3D camera estimation approach [LGJA09] is not possible throughout the entire endoscopic sequence. Finally, the image quality of the input material can be challenging. Depending on the endoscopic system, the algorithm has to cope with high noise levels, large areas of saturated highlights, interlacing artifacts, depth of field blur, false colors, etc. (see figure 1).

We therefore propose an algorithm that deviates from the typical feature-based approaches to motion compensation, but is neverthelss parallelizable and realtime capable even on the CPU. Our method uses an image-based inverse mesh warping approach similar to [HSE10] that can be stated as an optimization problem and solved efficiently in a robust Gauss-Newton framework (section 2). The inverse warping yields a piecewise affine deformation field between two successive frames. Using the motion field, a rigid image transformation can be computed to compensate for the camera motion. We discuss advantages of the image-based approach and show how the warp can be used to visualize the approximate 3D structure of the endoscopic scene from near-rigid parts of the sequence (section 3).

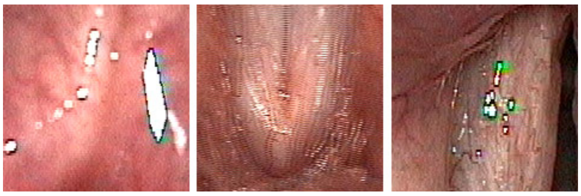

Figure 1: *Typical image artifacts the algorithm has to cope with: Saturated highlights, noise, interlacing and color artifacts.*

2. Inverse mesh warping

Mesh-based warping is a standard approach to computing complex image deformations by deforming a control mesh in the image plane. In the following we describe an image-based approach to solving the inverse problem, i. e. solving for a control mesh deformation given two images.

We denote the set of mesh vertices, identified by integer indices, as $\mathcal{V} = \{1 \dots K\}$, the set of image pixels as $\mathcal{P} = \{1 \dots N\}$, the vertex coordinates of the undeformed mesh as $[u_V \, v_V]^T$, $V \in \mathcal{V}$, and the pixel coordinates prior to the deformation as $[x_P \, y_P]^T$, $P \in \mathcal{P}$. Our mesh topology is a rectangular grid with diagonal edges through the cells. Denote by $T_P \in \mathcal{V}^3$ the surrounding triangle of pixel $P \in \mathcal{P}$ and be $c_P^{(1)}, c_P^{(2)}, c_P^{(3)}$ the barycentric coordinates of P with respect to T_P in the *undeformed* control mesh. For each pixel that lies under the mesh be $\mathbf{b}_P^T = \left[b_P^{(1)} \dots b_P^{(K)} \right]$ a sparse row vector of length K with

$$b_P^{(V)} = \begin{cases} c_P^{(i)} & \text{if } V \text{ is the } i\text{th vertex of triangle } T_P \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

With $\mathbf{B} = [\mathbf{b}_1 \dots \mathbf{b}_N]^T$, the transformation induced by a mesh deformation can written as

$$\mathbf{X}^* = \mathbf{X} + \mathbf{BD} \tag{2}$$

where $\mathbf{X}$ is the matrix of all pixel coordinates and $\mathbf{D}$ is a matrix of vertex displacements, i.e. a control mesh deformation:

$$\mathbf{X} = \begin{bmatrix} x_1 & y_1 \\ \vdots & \vdots \\ x_N & y_N \end{bmatrix} \quad \mathbf{D} = \begin{bmatrix} \Delta u_1 & \Delta v_1 \\ \vdots & \vdots \\ \Delta u_K & \Delta v_K \end{bmatrix}$$

Due to the use of barycentric coordinates, this amounts to a piecewise affine deformation of the image. Other interpolation schemes are possible.

The inverse problem—i. e. solving for the vertex displacement $\mathbf{D}$ given two images and an undeformed control mesh—can be stated as an optimization problem as follows. Assuming that $\mathcal{I}, \mathcal{K} : \mathbb{R}^2 \to \mathbb{R}$ are single channel images, we define the pixel-wise residual as

$$r_{P \in \mathcal{P}} = \mathcal{I}([x_P \, y_P]) - \mathcal{K}\left([x_P \, y_P] + \mathbf{b}_P^T \mathbf{D}\right). \tag{3}$$

Estimating $\mathbf{D}$ amounts to finding

$$\arg\min_{\mathbf{D}} \sum_{i=1}^{N} \rho(r_i) + \lambda \mathcal{R}(\mathbf{D}) \tag{4}$$

where ρ is a robust norm-like function such as Huber's. $\mathcal{R}(\mathbf{D})$ is a smoothness term which is addressed below, and λ is its weight relative to the data term.

For arbitrary norm-like functions, this energy can be minimized by a robust Gauss-Newton scheme that differs only slightly from the standard least squares case (e.g. [MN98]). This requires the Jacobian of the energy function, whose rows are gradients of the residual:

$$\nabla r_P = -\nabla \mathcal{K}^T \begin{bmatrix} \mathbf{b}_P^T \\ \mathbf{b}_P^T \end{bmatrix}. \tag{5}$$

Note that the second factor on the right hand side is the Jacobian of equation (2) with respect to pixel P.

For the smoothness term $\mathcal{R}(\mathbf{D})$ in equation (4) we use an orientation-separated Laplacian approach. Denoting by $\mathbf{L}$ the uniform weight Laplace matrix of the mesh, a classic smoothness term is $\mathcal{R}(\mathbf{D}) = \|\mathbf{LD}\|_F^2$ where $\|\cdot\|_F$ is the Frobenius norm. However, this term is known to suffer from a "shrinking bias" [GP10] that ultimately drives the vertices from the image border to the center in the case of 2D meshes. We exploit the simple structure of our deformation meshes to formulate a computationally cheap unbiased smoothness term. We extract two subgraphs from the mesh, the first subgraph containing only the horizontal links and the second only the vertical links; diagonal links are ignored. For both graphs, we construct Laplace matrices $\mathbf{L}_\mathrm{H}$ and $\mathbf{L}_V$ respectively that only affect vertices with valence two in order to avoid shrinking. We then use the following smoothness term:

$$\mathcal{R}(\mathbf{D}) = \|\mathbf{L}_\mathrm{H}\mathbf{D}\|_F^2 + \|\mathbf{L}_\mathrm{V}\mathbf{D}\|_F^2 \tag{6}$$

3. Stabilization and structure visualization

In the following we describe how we use inverse mesh warping to stabilize an endoscopic image sequence. As the endoscopic scene can be highly nonrigid, a region of interest (ROI) is required with respect to which the motion is to be stabilized. For fully automatic stabilization, the ROI can be defined, for example, as an area of a certain radius around the image center of the first frame. In the kymography application scenario the ROI is typically annotated by the physician by drawing a line along the vocal folds in the first frame. For the algorithm, the ROI is represented as a set of M points $\mathbf{p}_1 \dots \mathbf{p}_M$ in the image plane. Therefore, arbitrary interactive tools and shapes can be used for ROI definition. The mesh warp is computed independently for each image pair of the sequence. This step is computationally the most expensive part of the algorithm but it can be trivially parallelized to several cores due to the independence of the frame pairs.

The warp yields a piecewise affine deformation field between each frame pair $(i-1, i)$ which is represented as a

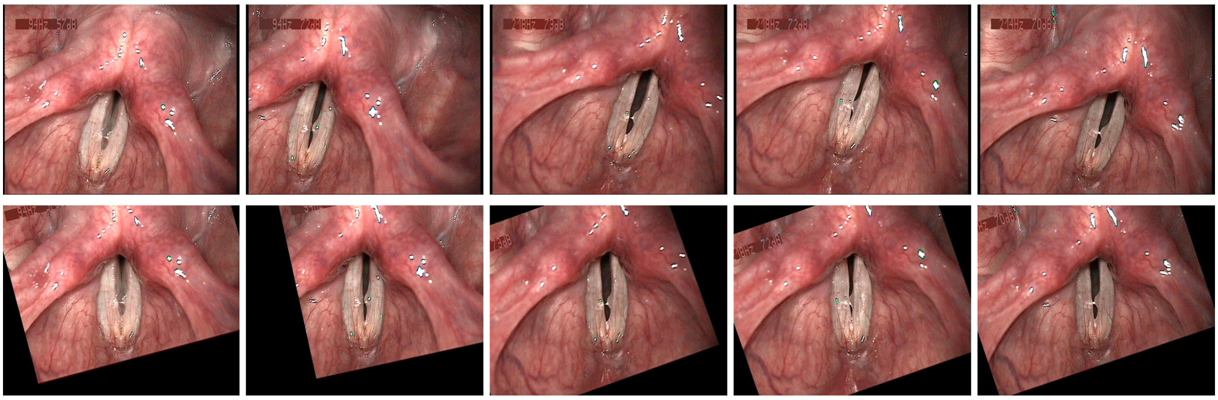

Figure 2: *Frames from an endoscopic video sequence of the vocal folds (top), motion compensated frames (bottom).*

matrix $\mathbf{D}_i$ of vertex displacements. It can be efficiently evaluated between the vertex locations using barycentric coordinates as in section 2 or a different interpolation scheme. Denote by $\mathcal{D}_i(\mathbf{p})$ the displacement at $\mathbf{p}$ between frames $(i-1,i)$. The tracked ROI point in frame i is given recursively by $\mathbf{p}_m^{(i)} = \mathbf{p}_m^{(i-1)} + \mathcal{D}_i\left(\mathbf{p}_m^{(i-1)}\right)$. A stabilizing transformation $\mathcal{T}$ for the ith frame is given by

$$\arg\min_{\mathcal{T}} \sum_{m=1}^{M} \left\| \mathbf{p}_m - \mathcal{T}\left(\mathbf{p}_m^{(i)}\right) \right\|^2 . \tag{7}$$

For the kymography application, $\mathcal{T}$ is constrained to be a rigid transformation and equation (7) is solved by an iterative descent scheme. Additional to the stabilizing transformation, the vocal fold is centered and aligned with the Y axis before kyogram computation.

The image-based approach to motion estimation has several key advantages for our application:

- We found it to be highly robust on images with high noise level and significant artifacts (see figure 1) if used with a robust error metric in equation (4). All results shown in the paper were generated with the Huber function.
- For computing the transformation, no explicit handling of outliers by RANSAC or similar methods is required. This is an significant advantage over feature-based approaches.
- As a global optimization scheme, the appraoch benefits from the "filling in" effect of the smoothness term that propagates information into image regions with little gradient information. It is therefore robust against uneven distribution of trackable image content.
- The choice of mesh granularity and weight of the regularization term allow for fine-grained control over the degree of deformation the warp is allowed to follow. It can be adjusted to track camera motion and large-scale scene deformation but to ignore the small-scale motion of the vocal folds which must not be "tracked away". High quality kymograms can be extracted even over large-scale deformations that occur when the patient takes breath.

The warp's displacement fields can also be used to visualize the approximate threedimensional structure of the endoscopic scene. Note that this part of the paper is work in progress. For structure computation, we manually identify a part of the image sequence that displays relatively little large-scale nonrigid motion. We then use the frame-to-frame displacement fields to track the vertices of a dense mesh in the image plane in the same way as the ROI points described above. Note that this mesh is different from the meshes used for warp estimation, which are not temporally consistent over all frames in order to allow for parallel computation of the warps. From the vertex trajectories, affine structure from motion is computed with the factorization algorithm of [TK92]. The endoscopic scenes generally violate the rigidity assumption on which the factorization algorithm is based even when the "most rigid" part of a sequence is chosen. Consequently, structure results are not precise reconstructions but rather visualizations of the approximate threedimensional shape of the scene. Examples are shown in figure 4.

4. Results and Conclusion

The algorithm processes about 25 frames per second on average on a six core machine with parallel warp estimation. The GPU is currently not used as capable GPUs still are rarely available in medical computing environments. Figure 2 shows frames from an endoscopic video sequence and their motion-compensated counterparts. Figure 3 shows kymograms computed from two sequences. For comparison, the kymograms were computed from the uncompensated video sequence, from the sequence compensated with the widely used "Deshaker" tool [Des] and from the sequence compensated with our approach. While the Deshaker kymograms show less jitter than the uncompensated ones, large-scale camera motion is not properly corrected. Especially the first row in figure 3 shows that the kymograms compensated with our method convey significantly more information than the comparison. Regarding differences in the kymograms,

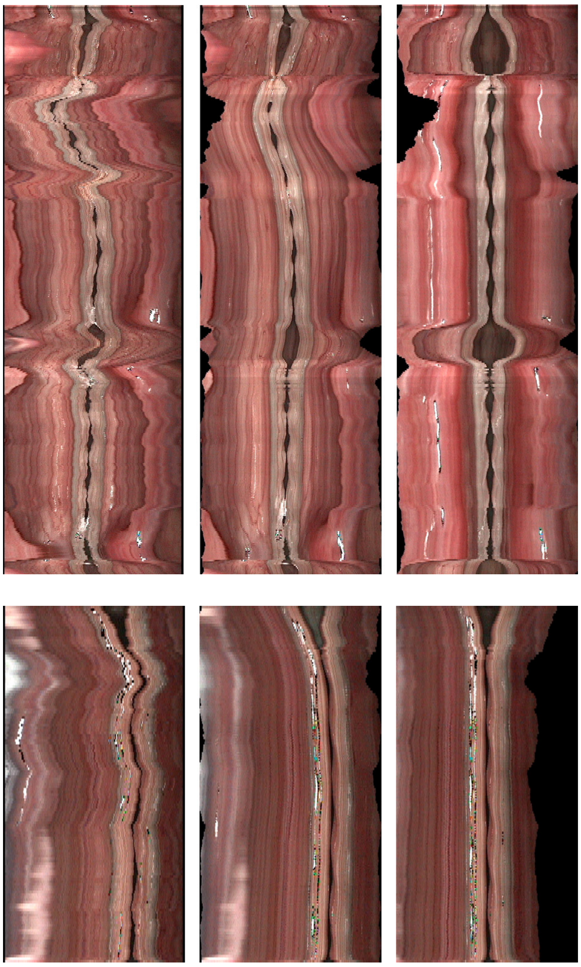

Figure 3: *Vocal fold kymograms from two endoscopic sequences. Left column: no motion compensation. Center column: Deshaker [Des] compensation. Right column: proposed method. At the wide openings in the first row the patient takes breath during the recording.*

also note that the scene moves relative to the scanline with respect to which the kymogram is computed if the sequence is not properly compensated. Figure 4 shows first results of our structure visualization approach.

In conclusion, we proposed an image-based warp estimation approach to compensate camera motion in endoscopic video. Our method deals with large motion, largely non-rigid scenes as well as poor image quality and image artifacts. Moreover, we showed first results on using our motion data for visualizing the approximate 3D shape of the endoscopic scene. Identifying the parts of the sequence suitable for structure computation and improving structure results will be the principal topic of further research.

References

[Des] http://www.guthspot.se/video/deshaker.htm. 3, 4

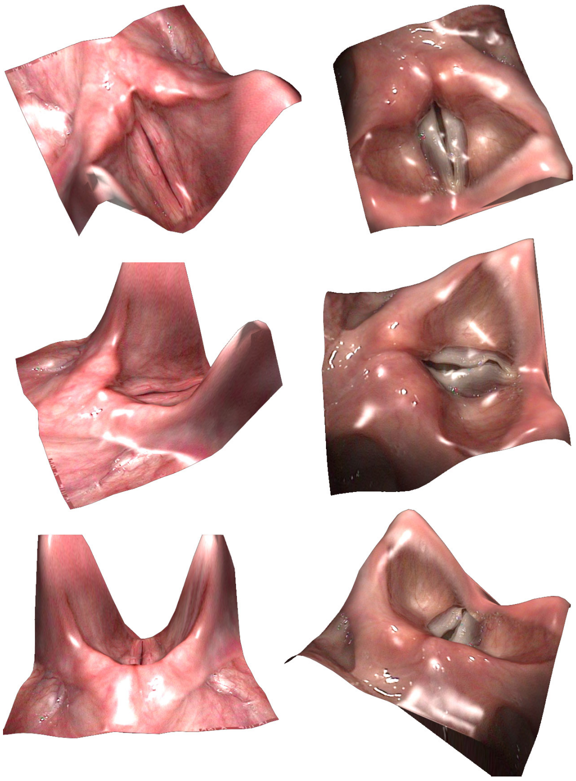

Figure 4: *Structure computation results for two sequences. Artificial lighting was added to emphasize the 3D shape.*

[GL08] Gleicher M. L., Liu F.: Re-cinematography: Improving the camera dynamics of casual video. *ACM Trans. on Multimedia 5* (2008). 1

[GP10] Grady L., Polimeni J.: *Discrete Calculus*. Springer, 2010. 2

[HSE10] Hilsmann A., Schneider D. C., Eisert P.: Realistic cloth augmentation in single view video under occlusions. *Comput. Graph. 34* (October 2010), 567–574. 1

[LCCO09] Lee K., Chuang Y., Chen B., Ouhyoung M.: Video stabilization using robust feature trajectories. In *Proc. Int. Conference on Computer Vision* (Kyoto, japan, 2009), pp. 1397–1404. 1

[LGJA09] Liu F., Gleicher M., Jin H., Agarwala A.: Content-preserving warps for 3d video stabilization. *ACM Trans. Graphics 28* (July 2009), 44:1–44:9. 1

[MN98] McCullagh P., Nelder J.: *Generalized Linear Models*. Chapman & Hall, 1998. 2

[SS95] Svec J. G., Schutte H. K.: Videokymography: High-speed line scanning of vocal fold vibration. *Journal of Voice 10 / 2* (1995), 201–205. 1

[TK92] Tomasi C., Kanade T.: Shape and motion from image streams under orthography: A factorization method. *International Journal of Computer Vision 9* (1992), 137–154. 3