Deshaking Endoscopic Video for Kymography

David C. Schneider*, Anna Hilsmann, Peter Eisert
Fraunhofer Heinrich Hertz Institute, Berlin, Germany

The opening and closing of the vocal folds (*plica vocalis*) at high frequencies is a major source of sound in human speech. *Videokymography* [Svec and Schutte 1995] 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 high framerate to capture vocal fold vibration (see fig. 1 for example frames). The *kymogram* used for medical diagnosis is a time-slice image, i.e. an *X-t*-cut through the *X-Y-t* image cube of the endoscopic video (fig. 2). 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. Therefore, we propose an approach to stabilizing the motion of endoscopic video for kymography.

This motion compensation problem is challenging and different from deshaking handheld video (e. g. [Liu et al. 2009]) in several respects: Firstly, the camera motion to be eliminated may be significantly larger than a typical camera shake due to the short distance between camera and scene. Secondly, not only the camera and the vocal folds move but the entire scene may be highly nonrigid. Finally, the image quality of the input material can be challenging due to high noise levels, areas of saturated highlights, interlacing artifacts, etc.

The proposed algorithm deviates from the typical feature-based approaches to motion compensation, but is nevertheless parallelizable and realtime capable even on the CPU. We use an image-based inverse mesh warping approach similar to [Hilsmann et al. 2010] that can be stated as an optimization problem and solved efficiently in a robust Gauss-Newton framework. Our method is described in more detail in [Schneider et al. 2011].

Mesh-based warping is a standard approach to computing complex image deformations by deforming a control mesh in the image plane. The inverse problem, i. e. solving for a control mesh deformation given two images, can be stated as an energy minimization task: Define the residual for pixel P as

$$r_P = \mathcal{I}(\mathbf{x}_P) - \mathcal{K}\left(\mathbf{x}_P + \mathbf{b}_P^T \mathbf{D}_T\right), \quad \mathbf{D}_T = \begin{bmatrix} \Delta u_1 & \Delta v_1 \\ \Delta u_2 & \Delta v_2 \\ \Delta u_3 & \Delta v_3 \end{bmatrix}$$

where $\mathcal{I}, \mathcal{K}$ are images and $\mathbf{x}_P$ is a pixel coordinate. Vector $\mathbf{b}_P$ contains the barycentric coordinates of pixel P with respect to its surrounding triangle T in the control mesh. $\mathbf{D}_T$ is a matrix of T's vertex displacements Δu_i and Δv_i. Estimating $\mathbf{D}_T$ for all triangles amounts to solving $\arg\min \sum_P \rho(r_P) + \lambda \mathcal{S}$ where ρ is a robust norm-like function such as Huber's and $\mathcal{S}$ is a smoothness term based on the mesh Laplacian. This energy can be minimized by a robust Gauss-Newton scheme that differs only slightly from the standard least squares case.

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 which can be efficiently evaluated between the vertex locations. Thereby, an ROI, user annotated or the center region of the first frame, can be tracked troughout the sequence and a stabilizing image transformation, which is restricted to be rigid for the kymography application, can be computed.

*e-mail: david.schneider@hhi.fraunhofer.de

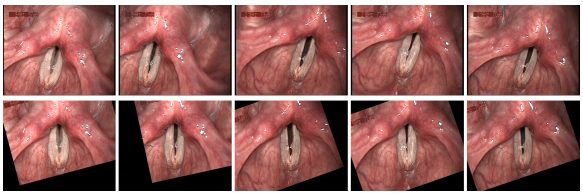

Figure 1: *Frames from an endoscopic video sequence of the vocal folds (top), motion compensated frames (bottom).*

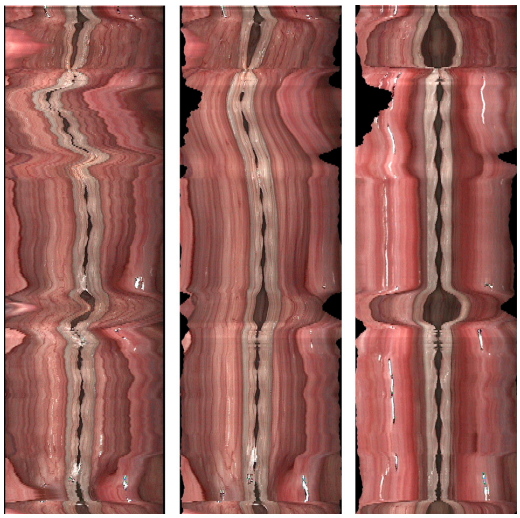

Figure 2: *Vocal fold kymograms from two endoscopic sequences. Left: no motion compensation. Center: Deshaker compensation. Right: proposed method.*

The image-based approach to motion estimation has several advantages for our application: (1) It is highly robust on images with high noise level and significant artifacts if used with a robust error metric. (2) For computing the transformation, no explicit handling of outliers (e.g. RANSAC) is required. This is an advantage over feature-based approaches. (3) As a global optimization scheme, the approach benefits from the "filling in" effect of the smoothness term that propagates information into image regions with little gradient information. (4) 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.

References

HILSMANN, A., SCHNEIDER, D. C., AND EISERT, P. 2010. Realistic cloth augmentation in single view video under occlusions. *Comput. Graph. 34* (October), 567–574.

LIU, F., GLEICHER, M., JIN, H., AND AGARWALA, A. 2009. Content-preserving warps for 3d video stabilization. *ACM Trans. Graphics 28* (July), 44:1–44:9.

SCHNEIDER, D. C., HILSMANN, A., AND EISERT, P. 2011. Warp-based Motion Compensation for Endoscopic Kymography. In *Eurographics 2011, Llandudno*.

SVEC, J. G., AND SCHUTTE, H. K. 1995. Videokymography: High-speed line scanning of vocal fold vibration. *Journal of Voice 10 / 2*, 201–205.