Automatic Workflow for Narrow-Band Laryngeal Video Stitching

Sara Moccia[⋆†‡], Veronica Penza[⋆†‡], Gabriele Omodeo Vanone[⋆‡], Elena De Momi[⋆], Leonardo S. Mattos[†]
[⋆]Department of Electronics, Informatics and Bioengineering, Politecnico di Milano, Milano, Italy
[†]Department of Advanced Robotics, Istituto Italiano di Tecnologia, Genoa, Italy
email: {sara.moccia, leonardo.demattos}@iit.it

Abstract— **In narrow band (NB) laryngeal endoscopy, the clinician usually positions the endoscope near the tissue for a correct inspection of possible vascular pattern alterations, indicative of laryngeal malignancies. The video is usually reviewed many times to refine the diagnosis, resulting in loss of time since the salient frames of the video are mixed with blurred, noisy, and redundant frames caused by the endoscope movements. The aim of this work is to provide to the clinician a unique larynx panorama, obtained through an automatic frame selection strategy to discard non-informative frames. Anisotropic diffusion filtering was exploited to lower the noise level while encouraging the selection of meaningful image features, and a feature-based stitching approach was carried out to generate the panorama. The frame selection strategy, tested on on six pathological NB endoscopic videos, was compared with standard strategies, as uniform and random sampling, showing higher performance of the subsequent stitching procedure, both visually, in terms of vascular structure preservation, and numerically, through a blur estimation metric.**

I. Introduction

Laryngeal early-stage malignancies are diagnosed inspecting the vessel pattern alterations on the mucosal surface. In recent clinical practice, Narrow Band (NB) endoscopy [1] has been introduced as an innovative diagnostic tool. NB employs a filtered spectrum illumination source containing wavelengths centered around blue and green frequencies only, which correspond to the hemoglobin absorption peaks, thus allowing to enhance superficial vessels as long as the procedure is performed nearby the tissue. The clinician is therefore constrained to a limited view of the larynx during the endoscopy, and usually reviews the video many times to refine the diagnosis, resulting in loss of time due to the presence of non-informative portions of the video.

The generation of a single panoramic image that synthesizes the inspection can help the clinician with an easier and unique visualization of the laryngeal tissue in the diagnostic phase. However, the composition of the larynx panorama is a challenging task under different points of view [2]. During the examination, the clinician moves the endoscope inside and outside the larynx, possibly rotating it. Moreover, the larynx does not remain still due to both vocal fold movements and patient swallowing. This directly results in sudden changes of the Field Of View (FOV), and in blurred frames. A second issue is related to the out-of-focus blurring, due to poor convergence of light from objects on the image sensor plane. The image quality is further compromised by the presence of high noise levels, caused by the camera sensor. Moreover, the video examination is made of several hundreds of consecutive single image frames, some of which containing redundant information.

In the literature, previous experiences of endoscopic video stitching applied to different body districts are reported [3]. To the best of the authors' knowledge, however, very little efforts have been invested in the laryngeal field. In [4], white light laryngeal video stitching was performed, using the general purpose software *AutoStitch* (http://matthewalunbrown.com/autostitch/autostitch.html). Results were encouraging, although there was no evidence of the endoscopic frame extraction strategy to achieve a fully automated process.

The goal of this research is to provide to the clinician a unique panorama of the laryngeal inspection video, obtained through a preliminary automatic frame selection performed to discard both blurred and redundant frames. Anisotropic diffusion filtering was carried out to lower the image noise level while enhancing edges. On the selected frames, a feature-based stitching algorithm was exploited to obtain the panorama. In order to maximize the amount and the quality of the information retained, the stitching was refined with error and false edge minimization strategies.

The paper is organized as follows: Sec. II describes the pipeline for the video frames selection and the stitching algorithm in detail; Sec. III concerns the stitching pipeline assessment, and describes materials on which the algorithm was tested; Sec. IV reports visual and numeric examples, which demonstrate the effectiveness of the proposed stitching workflow. In addition, comments and suggestions for potential improvements are offered in Sec. V.

II. Methods

The workflow of the proposed algorithm for laryngeal video stitching is shown in Fig. 1. The method consisted of three main steps: Frame selection (Sec. II-A), responsible for discarding non-informative video frames, Pre-processing (Sec. II-B), aiming at removing noise without blurring meaningful edges, and Stitching (Sec. II-C), performing all the operations needed to register and compose all of the selected frames as to obtain the full larynx panorama.

[‡] These authors contribute equally to this work.

978-1-4577-0220-4/16/$31.00 ©2016 IEEE

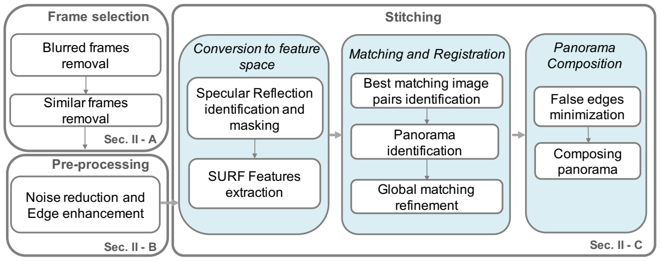


Fig. 1. Proposed automatic workflow for laryngeal video stitching.

A. Frame selection

The stitching quality strongly depends on the properties of the frames employed to obtain it. Therefore, only informative frames were retained and contributed to the laryngeal panorama generation, as described in this section.

The main sources of endoscopic frame degradation are the motion blur, due to the reciprocal movement of the camera and the larynx, and the presence of out-of-focus blur. These lower the image frequency content, decreasing the amount of useful information both for the human eye and for an automatic image processing tool.

The blur level in each frame was described by a single normalized numeric value, which ranges between 0 (in-focus) and 1 (blurred). This value was assigned according to the *Intentional Blurring Pixel Difference* (IBD) [5] computed on the image luminance channel. The luminance was chosen since it is well known that the image sharpness is encoded in its gray-level component. The idea behind IBD is to evaluate the differences between the analyzed image and its blurred version. The sharper is the original image, the higher will be this difference. Results on two frames are shown in Fig. 2. Receiver Operative Characteristic (ROC) curve analysis was performed to define the optimal IBD threshold, as a trade-off between sensitivity and specificity, based on the ground-truth provided by one subject on 150 video frames.

Once two consecutive selected frames were identified as sharp, a second issue was related to the redundant information contained in them, which increases the algorithm time-consumption without bringing useful contribution to the panorama. The two consecutive frames, converted in gray-scale, were compared using the Mean Structural SIMilarity (MSSIM) Index [6] in order to define their degree of similarity:

$$MSSIM(I_1, I_2) = \frac{1}{M}\sum_{j=1}^{M} SSIM(I_{1j}, I_{2j}) \qquad (1)$$

$$SSIM(\cdot) = [l(\cdot)]^{\alpha}[c(\cdot)]^{\beta}[s(\cdot)]^{\gamma} \qquad (2)$$

where I_{1j} and I_{2j} are the image contents at the j^{th} local window; l, c and s are comparison terms, respectively, for luminance, contrast and structure computed locally via convolution of the images with a circular symmetric Gaussian weighting function (standard deviation = 1.5) normalized to unit sum, M is the resulting number of local comparisons and $\alpha > 0$, $\beta > 0$ and $\gamma > 0$ are constants that adjust the contribution of each function.
This metric assigned to the frame pairs a single value comprised between 0 (low similarity) and 1 (high similarity). If a pair was classified as depicting the same scene, the sharper frame was retained as new comparison ground.

When performing the diagnostic endoscopy, the clinician often returns to previously inspected areas, generating multiple non-consecutive similar frames. A pairwise similarity comparison in straight temporal order was not able to take into account this possibility, thus resulting in retaining several times the same FOV. To further reduce these frames with redundant information, a global paired comparison was performed.

B. Pre-processing

In this work, pre-processing was based on non-linear anisotropic rotation invariant diffusion scheme based on Hybrid Diffusion with a Continuous Switch (HDCS) [7]. The main idea behind anisotropic diffusion filtering is to smooth homogeneous areas with an isotropic Gaussian-like kernel, and edge-like structures with an anisotropic kernel, elongated in the direction parallel to the edge itself. The diffusion tensor **D**, which drives the diffusion process, was built to have the same eigenvectors of the structure tensor **J** (Eq. 3). This constraint allows driving the diffusion according to the image distribution of gradient directions, since **J** describes the predominant directions of the image gradient in a specified neighborhood of a pixel.

$$\mathbf{J} = G_\rho \otimes \begin{vmatrix} (I_{\sigma_n} \otimes S_x)^2 & (I_n \otimes S_x)(I_{\sigma_n} \otimes S_y) \\ (I_{\sigma_n} \otimes S_y)(I_{\sigma_n} \otimes S_x) & (I_{\sigma_n} \otimes S_y)^2 \end{vmatrix} \qquad (3)$$

where G_ρ is a Gaussian kernel with standard deviation ρ, $\otimes$ is the convolution operator, I_{σ_n} was obtained convolving the image with a Gaussian of standard deviation σ_n, and S_x and S_y are the derivative Scharr kernels.
The amount of diffusion in the **J**-eigenvector directions was defined by the **D** eigenvalues, which were set as a combination of the eigenvalues defined in Coherence-Enhancing Diffusion (CED), which preserves plate-like structures, and Edge-Enhancing Diffusion (EED), which preserves tubular-like structures. An explicit finite-difference discretization of the anisotropic diffusion was employed, approximating the total diffusion time (*t*) as the number of iteration times the time step size (τ).

C. Stitching

Once the valid frames were automatically selected, a feature based stitching process [8] was carried out according to the following pipeline:

a) Conversion to feature space: In order to find correspondences between frames, and to register them on the same plane, salient features were identified on each retained frame using *Speeded Up Robust Features* (SURF) algorithm [9]. One major limitation is related to the identification of features in correspondence of specular reflections (SR).

SR are produced by the endoscopic light that reflects on the smooth and wet laryngeal surfaces, resulting in non-regular areas with a strong contrast with respect to the background. Since SR are characterized by low saturation and high brightness [10], a thresholding on these two image channels was employed to mask SR while extracting features. Threshold were selected as *0.155 * max(S)* and *0.710 * max(V)*, where *max(S)* and *max(V)* are the maximum values for saturation and brightness in the image converted to HSV space, respectively.

b) Matching and Registration: Matching between corresponding feature sets had to be established to perform the registration. Each feature set was coupled with the best matching one, i.e. the one which globally minimized the distances between matching features, using the Fast Library for Approximate Nearest Neighbors (FLANN) strategy. Only the biggest subset of matching frames was retained to create the panorama. Registration was performed with affine homographies, which were pairwise computed and made robust by the application of RANdom SAmple Consensus [11]. Bundle Adjustment [8] was subsequently used to globally correct the computed homographies.

c) Panorama Composition: The panorama was composed reprojecting each pixel of the frames on a plane at a unitary distance from the camera focal center, using each correspondent computed homography. A common practice in image stitching is to perform false edge minimization through blending operations. Therefore, *Multi-band blending* [12] was applied on the overlapping portions of the frames. However, this leads to ghost effects, which contribute to increasing the misregistration errors due to: (i) Imperfect registration; (ii) Differences in color; (iii) Exposition; (iv) Geometry of the larynx; (v) Muscle contraction; (vi) Loss of information after applying the homography. To minimize this issue, a *seam cut* approach [13] in the color gradient domain was used. This operation helped in preserving the blood vessel pattern in the output panorama, and was performed after *gain block* exposure compensation [14].

OpenCV 2.4.8 (http://opencv.org/) functions were used to implement the stitching algorithm.

III. Evaluation

The evaluation of the proposed pipeline aimed at verifying if the frame selection strategy improved the panorama composition. The algorithm was applied to 6 laryngeal endoscopic videos of patients with spinocellular carcinoma at different stages (*frame rate* = 25 fps, *frame size* = 768×576 and 1920×1072 px). The videos were provided by San Martino Hospital (University of Genoa, Italy). All patients gave the informed consent. From each video, the longest possible sequence of NB frames was selected, which lasted 5 minutes on average. This selection process was the only required human interaction.

The parameters used during the evaluation were: IBD threshold = 0.7, MSSIM threshold = 0.7 and, for anisotropic diffusion filtering t = 1, τ = 0.2, σ_n = 1, ρ = 4.

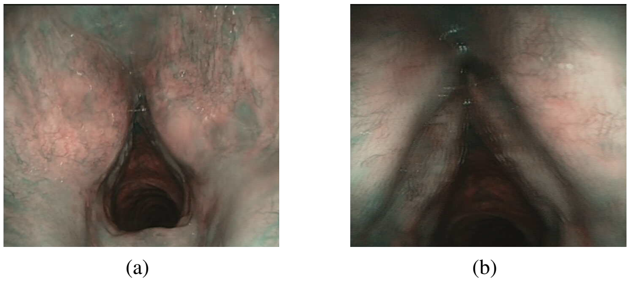

(a)
(b)

Fig. 2. Blur level estimation using Intentional Blurring Pixel Difference (IBD). (2a) In-focus frame (IBD value = 0.688). (2b) Blurred frame (IBD value = 0.833).

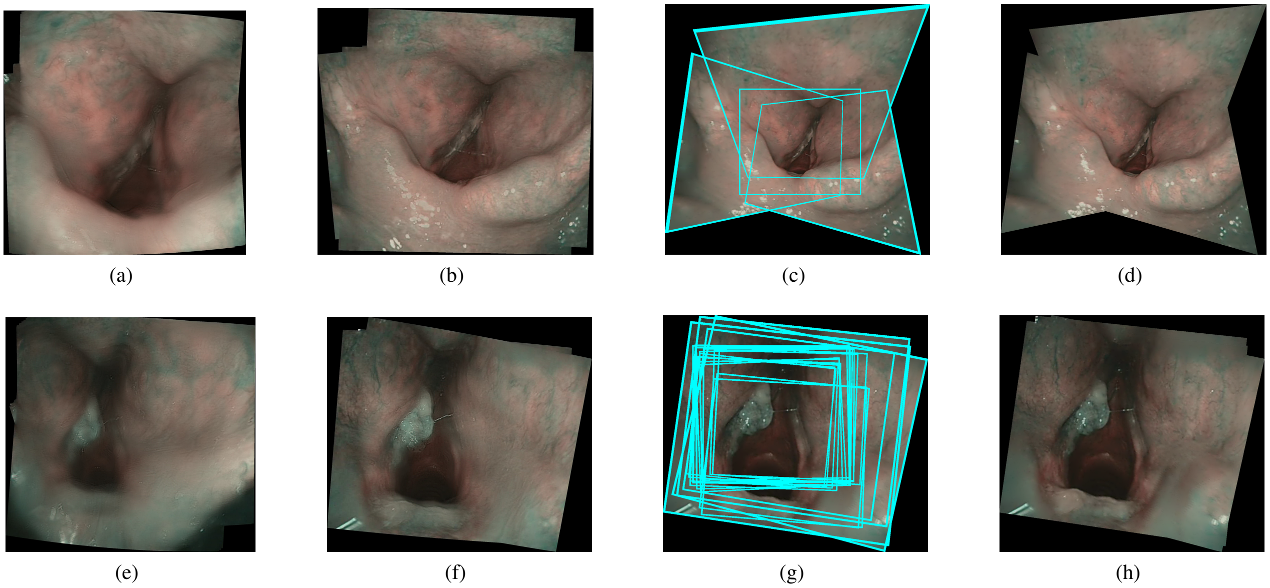
TABLE I
IBD values to estimate the blur level of panoramas obtained with different frame selection strategies

Video	1	2	3	4	5	6
AutoStitch	0.700	0.668	0.545	0.696	0.736	0.715
AutoStitch & FS	0.714	0.717	0.672	0.714	0.717	0.661
Proposed & RS	0.775	0.637	0.722	0.680	-	0.598
Proposed & US	-	0.656	0.772	0.664	-	0.598
Proposed & FS	0.708	0.625	0.658	0.600	0.497	0.619

The Frame Selection strategy (FS) described in Sec. II-A was compared with standard strategies, as uniform (US) and random (RS) sampling [2]. In these two latter cases, for a fair comparison, the number of extracted frames was kept equal to the number of frames extracted with FS. A qualitative evaluation was performed comparing the panorama obtained using the totality of frames with the one obtained only with the frames selected with FS, US, and RS; (i) Panorama blurring level, (ii) Vessel visibility, and (iii) Optic distortion were considered. In order to quantitatively estimate the panorama blurring level, IBD values were computed for all the obtained panoramas. A panorama was also generated using *AutoStitch* in order to perform a comparison with a state-of-the-art algorithm, exploiting both the totality of frames and the frames selected with FS.

IV. Results

The achieved results for 2 videos are shown in Fig. 3 for visual comparison. The first and second columns refer to the panoramas obtained with *AutoStitch* considering all the frames and only the frames selected according to FS, respectively. The third column shows the original frame locations in the proposed panoramas, which are depicted in the fourth column. The vascular information was clearly missed when no frame-selection was performed. In Table I, the IBD values are presented for the panoramas obtained with: *AutoStitch* with and without FS; the proposed stitching pipeline with RS and US; the full stitching pipeline. The lowest median IBD values was achieved with the proposed method (0.622), with a low inter-quartile range (0.045), attesting the smallest blurring level. Additionally, uniform and random sampling did not succeed in extracting valid frames for the stitching process in 2 videos.


(a) (b) (c) (d)

(e) (f) (g) (h)

Fig. 3. Panorama composition for video 1 (first row) and video 5 (second row). (a), (e) Panorama obtained with *AutoStitch*. (b), (f) Panorama obtained with *AutoStitch* and the frame selection strategy. (c), (g) Video frames superimposed on the panorama obtained with the proposed method. (d), (h) Panorama obtained with the proposed method.

V. Conclusion and Future work

This work aimed at providing to the clinician a single panoramic laryngeal image that summarizes the diagnostic examination, proposing a frame selection strategy to automate the stitching process. Results demonstrated that the proposed frame selection is crucial to achieve good stitching quality. Moreover, the inclusion of a denoising step and of stitching refinement strategies allowed to keep meaningful diagnostic structures, such as vessels, encoded in the final panorama. The quantitative evaluation based on IBD showed that the frame selection strategy improved the sharpness of the final panorama, which is an encouraging, thus still preliminary, step for the completely automatic generation of documentation reference panoramas from endoscopic videos.

A more accurate quantitative evaluation of the resulting panorama, employing synthetic data for which a ground-truth is available, is the next step of this research. In addition, a larger video dataset is also to be tested. This work could be further enriched by highlighting in the obtained laryngeal panorama, possible altered vascular pattern [15].

References

[1] A. Watanabe, M. Taniguchi, H. Tsujie, M. Hosokawa, M. Fujita, and S. Sasaki, “The value of narrow band imaging for early detection of laryngeal cancer,” *European Archives of Oto-Rhino-Laryngology*, vol. 266, no. 7, pp. 1017–1023, 2009.

[2] K. Schoeffmann, M. Del Fabro, T. Szkaliczki, L. Böszörmenyi, and J. Keckstein, “Keyframe extraction in endoscopic video,” *Multimedia Tools and Applications*, vol. 74, no. 24, pp. 11 187–11 206, 2015.

[3] T. Bergen and T. M. Wittenberg, “Stitching and surface reconstruction from endoscopic image sequences: a review of applications and methods,” *IEEE journal of biomedical and health informatics*, vol. 20, no. 1, pp. 304–321, 2014.

[4] M. Schuster, T. Bergen, M. Reiter, C. Münzenmayer, S. Friedl, and T. Wittenberg, “Laryngoscopic image stitching for view enhancement and documentation–first experiences,” *Biomedical Engineering*, vol. 57, no. SI-1 Track-H, pp. 704–707, 2012.

[5] F. Crete, T. Dolmiere, P. Ladret, and M. Nicolas, “The blur effect: perception and estimation with a new no-reference perceptual blur metric,” in *Electronic Imaging*. International Society for Optics and Photonics, 2007.

[6] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, “Image quality assessment: from error visibility to structural similarity,” *Image Processing, IEEE Transactions on*, vol. 13, no. 4, pp. 600–612, 2004.

[7] A. M. Mendrik, E.-J. Vonken, A. Rutten, M. A. Viergever, and B. Van Ginneken, “Noise reduction in computed tomography scans using 3-d anisotropic hybrid diffusion with continuous switch,” *Medical Imaging, IEEE Transactions on*, vol. 28, no. 10, pp. 1585–1594, 2009.

[8] M. Brown and D. G. Lowe, “Automatic panoramic image stitching using invariant features,” *International journal of computer vision*, vol. 74, no. 1, pp. 59–73, 2007.

[9] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool, “Speeded-up robust features (surf),” *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346 – 359, 2008.

[10] T. M. Lehmann and C. Palm, “Color line search for illuminant estimation in real-world scenes,” *Journal of the Optical Society*, vol. 18, no. 11, pp. 2679–2691, Nov.

[11] M. A. Fischler and R. C. Bolles, “Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography,” *Commun. ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[12] P. J. Burt and E. H. Adelson, “A multiresolution spline with application to image mosaics,” *ACM Transactions on Graphics (TOG)*, vol. 2, no. 4, pp. 217–236, 1983.

[13] V. Kwatra, A. Schödl, I. Essa, G. Turk, and A. Bobick, “Graphcut textures: image and video synthesis using graph cuts,” in *ACM Transactions on Graphics (ToG)*, vol. 22, no. 3, 2003, pp. 277–286.

[14] M. Uyttendaele, A. Eden, and R. Skeliski, “Eliminating ghosting and exposure artifacts in image mosaics,” in *Computer Vision and Pattern Recognition, Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 2, 2001.

[15] S. Moccia, E. De Momi, A. Ghilardi, A. Lad, and L. De Mattos, “Supervised hessian-based vessels segmentation in narrow-band laryngeal images,” in *Computer Assisted Radiology and Surgery, Proceedings of the 2016 International Congress and Exhibition on*. In press.