

Learning-based classification of informative laryngoscopic frames

Sara Moccia^{a,b}, Gabriele O. Vanone^a, Elena De Momi^a, Andrea Laborai^c,
Luca Guastini^c, Giorgio Peretti^c, Leonardo S. Mattos^b

^a*Department of Electronics, Information, and Bioengineering, Politecnico di Milano, Milan, Italy*

^b*Department of Advanced Robotics, Istituto Italiano di Tecnologia, Genoa, Italy*

^c*Department of Otorhinolaryngology, Head and Neck Surgery, University of Genoa, Genoa, Italy*

Abstract

Background and Objective Early-stage diagnosis of laryngeal cancer is of primary importance to reduce patient morbidity. Narrow-band imaging (NBI) endoscopy is commonly used for screening purposes, reducing the risks linked to a biopsy but at the cost of some drawbacks, such as large amount of data to review to make the diagnosis. The purpose of this paper is to present a strategy to perform automatic selection of informative endoscopic video frames, which can reduce the amount of data to process and potentially increase diagnosis performance. **Methods** A new method to classify NBI endoscopic frames based on intensity, keypoint and image spatial content features is proposed. Support vector machines with the radial basis function and the one-versus-one scheme are used to classify frames as informative, blurred, with saliva or specular reflections, or underexposed. **Results** When

Email address: sara.moccia@iit.it, sara.moccia@polimi.it (Sara Moccia)

Preprint submitted to Computer Methods and Programs in Biomedicine February 9, 2018

tested on a balanced set of 720 images from 18 different laryngoscopic videos, a classification recall of 91% was achieved for informative frames, significantly overcoming three state of the art methods (Wilcoxon rank-signed test, significance level = 0.05). **Conclusions** Due to the high performance in identifying informative frames, the approach is a valuable tool to perform informative frame selection, which can be potentially applied in different fields, such as computer-assisted diagnosis and endoscopic view expansion.

Keywords: Larynx, endoscopy, frame selection, supervised classification.

1. Introduction

Laryngeal cancer is a malignancy of the laryngeal tract, which, in terms of histopathology, takes the form of squamous cell carcinomas (SCC) in the 95% to 98% of cases [1]. It has been widely demonstrated in the clinical literature that the early-stage diagnosis of laryngeal SCC is crucial to improve the survival rate and the quality of life of patients after surgery [2].

Histopathological examination of tissue samples extracted with biopsy is the gold-standard for diagnosis. However, the relevance of visual analysis of tissues for screening purposes has recently led to the development of new optical-biopsy techniques, such as narrow-band imaging (NBI) endoscopy [3]. With NBI endoscopy, the clinician can benefit from an enhanced view of superficial blood vessels with respect to classic white-light endoscopy. This is crucial since an altered vascular pattern is a clear sign of tumor onset [1]. Similarly, a pre-cancerous tissue alteration known as leukoplakia is more visi-

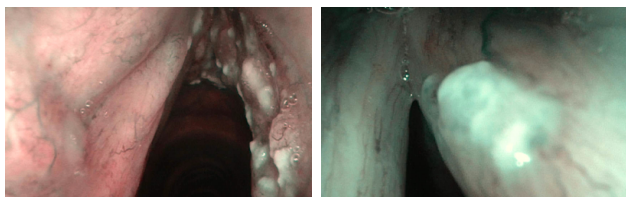


Figure 1: Sample images of vocal folds affected by leukoplakia, a pre-cancerous tissue alteration which causes vocal fold epithelium whitening and thickening.

15 ble with NBI with respect to standard white-light endoscopy [4]. Leukoplakia
16 implies thickening and whitening of the epithelial layer, as shown in Fig. 1,
17 and it is associated with an increased risk of cancer onset [4]. From the
18 patient’s side, the benefits of endoscopy examination with respect to tissue
19 biopsy are associated with reduced risk, trauma and shorter convalescence
20 time. Indeed, in case of biopsy, a tissue sample has to be extracted from the
21 patient, posing risks related to bleeding, pain, and infection [3].

22 Nevertheless, it is recognized that, from the clinician’s side, reviewing
23 an endoscopic video is a labour-intensive operation [5]. While focusing on
24 particular structures during the video examination, clinicians may miss im-
25 portant clues indicating suspicious conditions (e.g., early tumors). This pro-
26 cess could be further compromised by the presence of uninformative video
27 portions, which prolong the revision time of the endoscopic video.

28 Developing a strategy to select informative frames has the potential to
29 reduce the amount of data to review, lowering the surgeons’ workload. The
30 selection of informative frames can be beneficial also for computer-aided di-
31 agnosis algorithms. Preliminary efforts towards the automatic classification
32 of cancerous laryngeal tissues can be found in the literature (e.g. [6, 7]), but

33 they require manual frame selection so that the frames to be processed show
34 clearly the structures of interest. Frame selection strategies can benefit au-
35 tomatic diagnostic algorithms by (i) lowering the amount of computational
36 power required, and (ii) avoiding the processing of frames that do not show
37 structures of interest. Indeed, frames that do not show interesting structures
38 can dilute any further post-processing (such as classification and segmen-
39 tation) in computer-assisted diagnosis systems. This has the potential to
40 significantly enhance the performance of the diagnosis algorithms.

41 Informative frame selection also finds application in the context of image
42 stitching algorithms for endoscopic view expansion, for which the presence
43 of uninformative video frames is recognized to strongly affect the resulting
44 panoramic image quality [8, 9]. Researches in the field of endoscopic view
45 expansion include [10, 11, 12, 13, 14]. Unfortunately, the laryngeal field has
46 been underrepresented in these studies, with only one significant contribution
47 in [5] and our preliminary work in [15]. However, a clear limitation of the
48 system proposed in [5] is the lack of a robust and automatic strategy for the
49 selection of informative video frames, while the work in [15] is limited to the
50 removal of blurred frames using a sensitive threshold-based approach.

51 A possible solution to lower the number of uninformative frames is ex-
52 ploiting preliminary visual image quality assessment through subjective eval-
53 uation. This operation is, however, prone to human error and usually too
54 inconvenient and time-consuming [16].

55 Automatic selection of informative frames is a valuable alternative. In

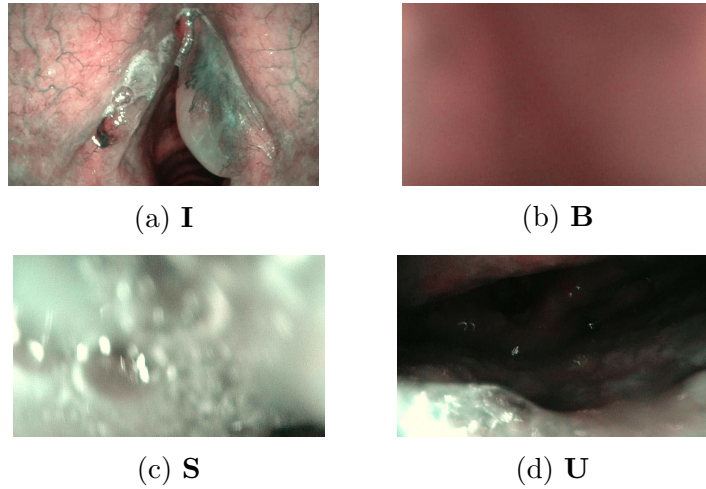


Figure 2: Visual examples of laryngeal video frames. (a) Informative frame (**I**); (b) Blurred frame (**B**); (c) Frame with saliva and specular reflections (**S**); (d) Underexposed frame (**U**).

56 the context of laryngoscopic video analysis, the identification of informative
 57 frames, such as the one in Fig. 2a, is not trivial [8]. Beside the well-known
 58 challenges associated to endoscopy, such as high camera-noise level in the
 59 images, major challenges typical of the laryngeal district include:

- 60 • Movement of swallowing muscles and vocal folds, as well as free and
 61 varying endoscope pose, which produces blurring in the images (Fig. 2b)
- 62 • Presence of specular reflections, due to the smooth and wet laryngeal
 63 surface, and saliva (Fig. 2c)
- 64 • Varying illumination conditions, resulting in underexposed video frames
 65 (Fig. 2d)

66 *1.1. Related work*

67 Several attempts to automatic frame selection can be found in the liter-
68 ature, even though none of them specifically focuses on laryngoscopic video
69 analysis, probably due to the lack of publicly available dataset for testing.

70 Many of the approaches exploit simple uniform frame sampling to reduce
71 the amount of data to process (e.g., [17, 13] for bladder). Uniform sam-
72 pling is fast in terms of run-time, but there is no guarantee that informative
73 frames are extracted from all semantically important video segments. At
74 the same time, for long segments with identical content a large number of
75 redundant keyframes are selected. Moreover, also uninformative frames can
76 be potentially elected as keyframes.

77 More advanced state of the art frame selection strategies applied to the
78 endoscopic medical field can be roughly divided into two branches:

79 *1.1.1. Video clustering and keyframe extraction*

80 The goal of this class of algorithm is to cluster video frames with simi-
81 lar informative content, exploiting similarity measures between features ex-
82 tracted from the images.

83 In [18, 19] keyframes are extracted using a keypoint-based approach. A
84 keyframe is extracted if the distance between consecutive frames in the key-
85 point space overcomes an user defined threshold.

86 In [20, 21, 22] features based on color, texture and motion displacement
87 are used to identify, with a threshold-sensitive approach, frames with redun-

88 dant informative content. Instead of using simple thresholding, in [23] linear
89 discriminant analysis is applied in the feature space.

90 In [24], color and edge features are clustered with k-means. From each
91 cluster, a representative frame is arbitrarily extracted as keyframe. In [25],
92 clusters are obtained with uniform sampling and non-negative matrix factor-
93 ization is used to extract keyframes from each cluster.

94 This class of algorithms potentially brings the advantage of summarizing
95 the video content. Nonetheless, such algorithms do not make any assump-
96 tions about the presence of uninformative video portions, which can poten-
97 tially represent a high percentage of the endoscopic video content without
98 bringing any useful information for diagnosis.

99 1.1.2. *Uninformative frame removal*

100 This class of algorithms aims at evaluating if the content of a frame is
101 of interest for a given application (e.g., its quality is sufficient to appreciate
102 structures of interest).

103 In [26, 27, 28], after uniform sampling, uninformative frames are removed
104 if the number of keypoints is lower than a threshold.

105 The work in [29, 15] uses an intensity-based similarity score to assess the
106 degree of images blur. Thresholding is applied to discard low quality images.
107 Shannon entropy is instead used in [30, 31].

108 In [32, 33], frames are clustered as informative and uninformative us-
109 ing features in the image frequency domain and k-means. Gray-level co-

Learning-based classification of informative laryngoscopic frames

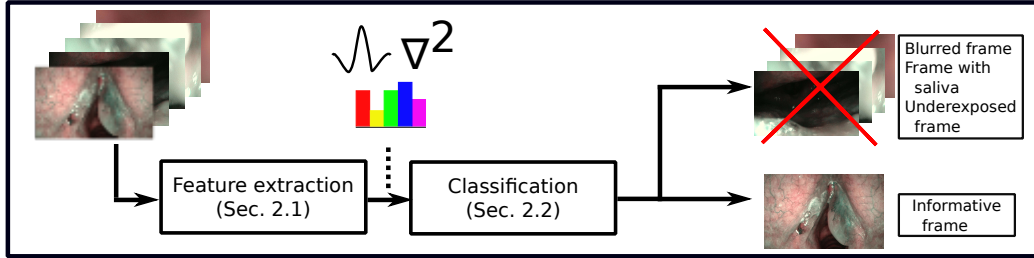


Figure 3: Workflow of the proposed approach to automatic learning-based classification of informative laryngoscopy frames.

110 occurrence (GLCM)-based features and Gaussian mixture model are used
 111 in [34].

112 A more advanced approach to uninformative frame removal, which is also
 113 the first attempt at using machine-learning for this aim, has been proposed in
 114 [35, 36]. The classification process exploits support vector machines (SVMs)
 115 trained on local color histogram features to discriminate between uninforma-
 116 tive frames with residual food and potentially informative frames.

117 *1.2. Aim of the work*

118 In this paper we specifically address the problem of robust and automatic
 119 classification of informative frames with applications in laryngoscopy. The
 120 proposed approach exploits the strong generalization power of machine learn-
 121 ing, overcoming issues related to the definition of threshold values to assess
 122 the quality of the image. Instead of focusing on the identification of just one
 123 class of uninformative frames, our approach is to extend the classification
 124 process to four classes (as to deal with all the typically encountered types of
 125 uninformative frames in NBI laryngeal endoscopic videos) namely:

- 126 • Underexposed frames (**U**)
- 127 • Frames with saliva or specular reflections (**S**)
- 128 • Blurred frames (**B**)
- 129 • Informative frames (**I**)

130 In addition to identify informative frames, being able to identify these
131 classes of uninformative frames may help in:

- 132 1. Processing initially excluded frames to increase frame quality
- 133 2. Informing the clinician on the quality of images he/she is acquiring in
134 real-time

135 In the first case, post-processing algorithms could be used to increase bright-
136 ness/contrast for underexposed frame or try to extract residual useful infor-
137 mation from frames with saliva or specular reflections. In the second one, the
138 clinician could perform corrective actions, e.g. increase the illumination level,
139 move the endoscope slower to minimize motion blur, rinse the endoscope.

140 The paper is organized as follows: the proposed approach to the learning-
141 based classification of informative laryngoscopy frames is explained in Sec. 2.
142 The materials used and the evaluation protocol are described in Sec. 3. Re-
143 sults are presented in Sec. 4 and discussed in Sec. 5. Major strengths, limi-
144 tations and future work are given in Sec. 6 to conclude this paper.

145 **2. Methods**

146 In this section, we present a detailed description of the proposed approach
147 to learning-based classification of NBI laryngoscopy video frames. The feature
148 extraction strategy is explained in Sec. 2.1, and the classification in Sec. 2.2.
149 The workflow of the proposed approach is shown in Fig. 3.

150 *2.1. Feature extraction*

151 The aim of the classification features is to encode the main distinctive
152 characteristics of the four frame classes. For each class, specific assumptions
153 on the image content can be made. For instance, underexposed frames can
154 be classified according to intensity-based features, since they contain high
155 percentage of dark pixels (Fig. 2d). Informative frames have higher spatial
156 frequencies content than blurred frames, due to the presence of sharp edges,
157 such as blood vessels, as can be seen by comparing Fig. 2a and Fig. 2b.
158 However, the presence of saliva or bubbles in the image creates similar com-
159 ponents in the spatial frequency domain, too. Frames with saliva or specular
160 reflections can be differentiated based on the color domain, as images with
161 such content present high components in the green and blue color channels
162 (Fig. 2c).

163 In addition to such assumptions, features should be computationally
164 cheap in order to minimize the effort with a view to real-time applications.

165 Part of the feature set used in this work was borrowed from state of the
166 art methods with application in medical imaging (such as [26], [29], [35]). In

Table 1: Tested feature vectors and corresponding number of features.

Descriptor	Length
Blind/referenceless image spatial quality evaluator (<i>BRISQUE</i>) [37]	1
Variance of the image Laplacian (Δ_{VAR})	3
Sobel-Tenengrad focus evaluation function score (<i>TEN</i>) [38]	3
Image entropy (<i>ENTROPY</i>)	3
Local variance of the luminance channel intensity (<i>VAR</i>)	3
Image intensity variance (<i>G-VAR</i>)	3
Image histogram (<i>H</i>)	3
Number of detected keypoints (<i>N-P</i>)	1
Total	20

167 addition, new features were included to make the classification robust to the
 168 laryngoscopy scenario.

169 The set of features, which is summarized in Table 1, consisted of:

- 170 • Blind/referenceless image spatial quality evaluator (*BRISQUE*):

171 The blind/referenceless image spatial quality evaluator (*BRISQUE*) [37]
 172 is a no-reference image quality assessment holistic metric that operates
 173 in the spatial domain. To obtain *BRISQUE*, we first computed the
 174 image normalized luminance coefficient, i.e. mean-subtracted contrast-
 175 normalized luminance pixel values. Such coefficient were approximated
 176 by the asymmetric generalized Gaussian distribution (AGGD):

$$f(x, \alpha, \sigma_1^2, \sigma_2^2) = \begin{cases} \frac{a}{(\beta_l + \beta_r)\gamma(1/\alpha)} \exp\left(-\left(\frac{-x}{\beta_l}\right)^\alpha\right) & x < 0 \\ \frac{a}{(\beta_l + \beta_r)\gamma(1/\alpha)} \exp\left(-\left(\frac{x}{\beta_r}\right)^\alpha\right) & x \geq 0 \end{cases} \quad (1)$$

177 where α is a shape parameter, σ_l, σ_r are scale parameters and γ, β_l, β_r

178 depend on $\alpha, \sigma_l, \sigma_r$, as explained in [37]. *BRISQUE* was computed
 179 by regression on the computed AGGD parameters, using a regressor
 180 trained on non-distorted natural images as in the original work [37].

181 • Variance of the image Laplacian (Δ_{VAR}):

182 Since a high percentage of informative content (i.e., sharp edges, such
 183 as blood vessels and vocal fold borders) is encoded in high frequen-
 184 cies in the spatial frequency domain, a measure (Δ_{VAR}) based on the
 185 Laplacian (L) of the image I was used as feature, as suggested in [38]
 186 for autofocusing in light microscopy videos. Given I of size $M \times N$,
 187 Δ_{VAR} is computed as:

$$\Delta_{VAR} = \sum_m^M \sum_n^N (L(m, n) - \bar{L})^2 \quad (2)$$

188 where \bar{L} is:

$$\bar{L} = \frac{1}{MN} \sum_m^M \sum_n^N |L(m, n)| \quad (3)$$

189 and L is obtained by convolving I with the Laplacian kernel (K_L):

$$K_L = \frac{1}{6} \begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix} \quad (4)$$

190 • Sobel-Tenengrad focus evaluation function (TEN):

191 The Sobel-Tenengrad focus evaluation function (TEN) [38] is another
 192 measure typically used in microscopy autofocusing [39], based on the
 193 image gradient magnitude value. Being G_x , G_y the image gradient
 194 along the x and y direction, respectively, TEN is defined as:

$$TEN = \sum_m^M \sum_n^N [S(m, n)]^2, \quad \text{for } S(m, n) > T \quad (5)$$

195 where T is a threshold and:

$$S(m, n) = \sqrt{[G_x(m, n)]^2 + [G_y(m, n)]^2} \quad (6)$$

196 As to obtain G_x and G_y , I was convolved with the Sobel's kernel (K_S)
 197 and its transpose, respectively, where:

$$K_S = \begin{bmatrix} 1 & 0 & -1 \\ 2 & 0 & -2 \\ 1 & 0 & -1 \end{bmatrix} \quad (7)$$

198 • Image entropy ($ENTROPY$):

199 Image entropy ($ENTROPY$) is an effective measure of the amount of
 200 information in an image [40]. Here, it was used as feature, as suggested
 201 in [41] for quality assessment of natural images:

$$ENTROPY = - \sum_i h_i \log_2(h_i) \quad (8)$$

202 where h_i refers to the I histogram counts of the $i \in [0, 255]$ bin.

203 • Local variance of the luminance channel intensity (VAR):

204 The frame edge content can be roughly assessed also from variations in
 205 the local intensity variance (VAR) [38]. VAR is defined as:

$$VAR = \frac{1}{MN} \sum_m^M \sum_n^N [lv(m, n) - \bar{lv}]^2 \quad (9)$$

206 where:

$$lv(m, n) = \frac{1}{w_x w_y} \sum_i^{w_x} \sum_j^{w_y} [I(m+i, n+j) - \bar{I}_w]^2 \quad (10)$$

$$\bar{lv} = \frac{1}{MN} \sum_m^M \sum_n^N lv(m, n) \quad (11)$$

207 and \bar{I}_w is the mean intensity value on the window.

208 • Image intensity variance (G_VAR):

209 In addition to the VAR local focus measure, a global intensity variance
 210 (G_VAR) focus measure was computed. G_VAR (Eq. 12) was used in
 211 addition to VAR to improve the feature robustness against noise.

$$G_VAR = \frac{1}{MN} \sum_m^M \sum_n^N [I(m, n) - \bar{I}]^2 \quad (12)$$

212 where \bar{I} is the mean intensity of all pixel in I .

213 • Image histogram (H):

214 Inspired by [35] and to include intensity-related features, the first quar-
215 tile, median and third quartile of the image histogram were added to
216 the feature vector.

217 • Number of keypoints (N_P):

218 The number of keypoints in a frame is a trivial measure of information,
219 as suggested in [26]. Here, oriented fast and rotated brief (ORB) [42]
220 was used to detect keypoints in the gray-scale version of I . ORB is a
221 fast binary descriptor, rotation invariant and robust to noise.

222 Δ_{VAR} , TEN , $ENTROPY$, VAR , G_VAR and H were computed for
223 each I color channel in the RGB space.

224 Prior to feature extraction, anisotropic diffusion filtering [43] was used to
225 lower noise while preserving sharp edges in NBI images [15].

226 2.2. Classification

227 To perform tissue classification, a support vector machine (SVM) was
228 used [44]. The SVM *kernel-trick* prevents parameter proliferation, lower-
229 ing computational complexity and limiting over-fitting. Moreover, the SVM
230 decisions are only determined by the support vectors, which makes SVM
231 robust to noise in training data. Here, the SVM with Gaussian kernel (Ψ)
232 was used. For a binary classification problem, given a training set of N data
233 $\{y_k, \mathbf{x}_k\}_{k=1}^N$, where \mathbf{x}_k is the k^{th} input feature vector and y_k is the k^{th} output
234 label, the SVM decision function (f) takes the form of:

$$f(\mathbf{x}) = \text{sign} \left[\sum_{k=1}^N a_k^* y_k \Psi(\mathbf{x}, \mathbf{x}_k) + b \right] \quad (13)$$

235 where:

$$\Psi(\mathbf{x}, \mathbf{x}_k) = \exp\{-\gamma \|\mathbf{x} - \mathbf{x}_k\|_2^2 / \sigma^2\}, \quad \gamma > 0 \quad (14)$$

236 b is a real constant and a_k^* is computed as follow:

$$a_k^* = \max \left\{ -\frac{1}{2} \sum_{k,l=1}^N y_k y_l \Psi(\mathbf{x}_k, \mathbf{x}_l) a_k a_l + \sum_{k=1}^N a_k \right\} \quad (15)$$

237 with:

$$\sum_{k=1}^N a_k y_k = 0, \quad 0 \leq a_k \leq C, \quad k = 1, \dots, N \quad (16)$$

238 In this paper, the SVM parameters γ and C were computed with grid
 239 search and cross-validation, as explained in Sec. 3. To implement multi-class
 240 SVM classification, the *one-vs-one* scheme was used, assigning ambiguous
 241 test points to the nearest decision boundary. With the *one-vs-one* scheme,
 242 one binary SVM classifier was constructed for pairs of frame classes. For each
 243 binary learner, one class was considered positive, another was negative, and
 244 the rest were ignored. This design exhausted all combinations of class pair
 245 assignments. At prediction time, the class which received the most votes was
 246 selected.

247 Prior to classification, the feature matrices were standardized.

Table 2: Evaluation dataset. For each video (video ID), and for each class (**I**, **B**, **S**, **U**), the number of frames that contributed to build the dataset are reported. The dataset is split in 3 folds to perform robust estimation of the classification performance. The folds are balanced both at patient- and class-level. **I**: informative frame; **B**: blurred frame; **S**: frame with saliva or specular reflections; **U**: underexposed frame.

	video ID	I	B	S	U
Fold 1	1	10	10	20	11
	2	10	0	6	9
	3	10	0	0	2
	4	10	40	23	20
	5	10	10	11	3
	6	10	0	0	15
	total	60	60	60	60
Fold 2	7	10	28	19	0
	8	10	8	21	5
	9	10	3	10	10
	10	10	21	10	16
	11	10	0	0	14
	12	10	0	0	15
	total	60	60	60	60
Fold 3	13	10	17	0	11
	14	10	21	34	22
	15	10	0	11	10
	16	10	0	9	5
	17	10	12	0	2
	18	10	10	6	11
	total	60	60	60	60

248 3. Evaluation

249 In this study, 18 NBI endoscopic videos, referring to 18 different patients
 250 affected by SCC, were retrospectively analyzed (average video length: 39s).
 251 Videos were acquired with a NBI endoscopic system (Olympus Visera Elite
 252 S190 video processor and an ENF-VH rhino-laryngo videoscope) with frame
 253 rate of $25fps$ and image size of 1920×1072 pixels.

254 A total of 720 video frames, 180 for each of the four classes (**I**, **B**, **S**,
 255 **U**) was extracted and labeled from the 18 videos, see Table 2. For each
 256 video, video frames were randomly extracted and presented to two human

257 evaluators first. Then, the two evaluators were asked to label the frames. In
258 case the two evaluators did not agree on the class, a third evaluator was asked
259 to choose the ultimate class among the two proposed by the two evaluators.
260 This process was repeated until all the 720 frames were extracted from the
261 videos. For the manual labeling process, the following set of rules was defined:
262 **I** frames should have an adequate exposure and clearly visible blood vessels;
263 they may also present micro-blur and small portions of specular reflections
264 (up to 10% of the image area). **B** frames should show a homogeneous and
265 widespread blur. **S** frames should present bright white/light-green bubbles
266 or blobs, overlapping with at least half of the image area. Finally, **U** frames
267 should present a high percentage of dark pixels, even though small image
268 portions (up to 10% of the image area) with over- or normal exposure are
269 allowed.

270 In addition, one of the videos was fully labeled (length = 17.64s; number
271 of frames: **I** = 341, **B** = 7, **S** = 9, **U** = 84).

272 All the frames underwent the pre-processing step described in Sec. 2.
273 The anisotropic diffusion filtering parameters were set as in [43].

274 From each frame, the features described in Sec. 2.1 were obtained using
275 the following parameters:

- 276 • BRISQUE: *BRISQUE* code was downloaded from the *Laboratory*
277 *for Image & Video Engineering* website¹ and the parameters were set

¹[<http://live.ece.utexas.edu/research/quality/index.htm>]

278 as in [37]

279 • TEN: the threshold T of Eq. 5 was set to 0, as suggested in [39] to
280 include all pixels in the computation

281 • VAR: to compute VAR , the local window size was 5×5 pixels

282 • N_P: the parameters of ORB were set as in the original paper [42]

283 As for performing the classification presented in Sec. 2.2, the SVM hyper-
284 parameters (γ , C) were retrieved via grid-search and 10 fold cross-validation
285 on the training set. The grid-search space for γ and C was set to $[10^{-7}, 10^{-1}]$
286 and $[10^{-3}, 10^3]$, respectively, with seven values spaced evenly on \log_{10} scale
287 in both cases.

288 The feature computation was implemented using OpenCV ². The classi-
289 fication was implemented with scikit-learn ³.

290 3.1. Experimental setup

291 To obtain a robust estimation of the classification performance of the
292 frames reported in Table 2, 3-fold cross-validation was performed, separating
293 data at patient level. We separated data at patient level to ensure that
294 frames from the same class were classified due to features that are peculiar to
295 that class, and not due to features linked to the patient itself (e.g. vocal fold
296 anatomy). When the classification of the frames in fold 3 was performed, folds

²[\[http://docs.opencv.org/3.1.0/index.html\]](http://docs.opencv.org/3.1.0/index.html)

³[\[http://scikit-learn.org/stable/index.html\]](http://scikit-learn.org/stable/index.html)

297 1 and 2 were used to train the SVM. To retrieve the SVM parameters during
 298 the training phase, 10 fold cross-validation and grid-search were performed
 299 on the training set (i.e. using images from folds 1 and 2), as explained in
 300 Sec. 3. We did the same for testing the classification of frames in fold 1 and 2,
 301 using fold 2 and 3, and fold 1 and 3 for hyper-parameter tuning, respectively.

302 We built a balanced dataset both at patient level and frame class level,
 303 as shown in Table 2. It can be noticed from Table 2 that, for some videos,
 304 selecting an equal number of frames for the four classes was not always pos-
 305 sible, especially for the uninformative ones. The reason is that either the
 306 videos could contribute only with ambiguous frames (i.e. frames with mixed
 307 characteristics among the four classes) or a sufficient number of frames was
 308 not available for all the classes. When this was the case, the other videos in
 309 the fold contributed to balance the number of frames. The approach followed
 310 to balance the dataset is common for studies with limited amount of data.
 311 A similar approach was followed, for example, in [35].

312 In order to evaluate the classification performance, the class-specific recall
 313 ($\mathbf{Rec}_{\mathbf{class}} = \{Rec_{class_j}\}_{j \in [1, J=4]}$), the precision ($\mathbf{Prec}_{\mathbf{class}} = \{Prec_{class_j}\}_{j \in [1, J=4]}$),
 314 and the F1 score ($\mathbf{F1}_{\mathbf{class}} \{F1_{class_j}\}_{j \in [1, J=4]}$), were computed, where:

$$Rec_{class_j} = \frac{TP_j}{TP_j + FN_j} \quad (17)$$

$$Prec_{class_j} = \frac{TP_j}{TP_j + FP_j} \quad (18)$$

$$F1_{class_j} = 2 \frac{Prec_{class_j} \times Rec_{class_j}}{Prec_{class_j} + Rec_{class_j}} \quad (19)$$

315 being TP_j the true positive of the j^{th} class, FN_j the false negative of the j^{th}
 316 class, and FP_j the false positive of the j^{th} class.

317 The area (AUC) under the receiver operating characteristic (ROC) curve
 318 was also computed. Since our task is a multi-class classification problem and
 319 the dataset was balanced, the macro-average ROC curve was computed.

320 The computational time required to extract and classify the proposed
 321 features was computed, as well. Experiments were performed on a CPU
 322 Intel® Core™2 Duo @ 2.26GHz with 8GB of available RAM; Linux oper-
 323 ative system, kernel 4.4.0-98-generic (x86_64) Ubuntu 16.04.3 LTS distribu-
 324 tion.

325 We also investigated the use of feature selection. We applied principal
 326 component analysis (PCA) [45] to our feature set to retrieve a relevant set
 327 of features. We then performed the classification explained in Sec. 2.2. For
 328 the PCA implementation, principal components were retrieved as to explain
 329 the 99% of the variance encoded in the features.

330 For the sake of completeness, the performances of random forest (RF) [46]

331 in classifying the proposed feature set were also investigated and compared
332 with those obtained with SVM. The number of trees in the forest for RF was
333 found with grid-search and cross-validation with a grid-search space set to
334 [40,100] with six values spaced evenly.

335 The SVM performance were compared also with those obtained using
336 the features commonly exploited in the state of the art. As explained in
337 Sec. 1.1.2, commonly exploited features are (i) image keypoints, (ii) intensity-
338 based similarity score, (iii) color features and (iv) textural information. There-
339 fore, we decided to compare our method with one research per feature cat-
340 egory. We considered [26, 29, 35, 34], which use ORB keypoints, intensity-
341 based similarity score, color histogram and GLCM, respectively. The pa-
342 rameters for the state of the art methods implemented for comparison were
343 set as reported in their reference papers: ORB parameters for [26], thresh-
344 olding values for the intensity-based similarity score for [29], histogram bin
345 number for [35] and orientation and radius for GLCM computation for [34].
346 As stated in Sec. 1, such methods rely on thresholding instead of machine
347 learning-based methods. However, the features from the state of the art were
348 classified with SVM, for fair comparison.

349 The Wilcoxon signed-rank test (significance level $\alpha = 0.05$) for paired
350 samples was used to assess whether the classification achieved with the pro-
351 posed feature vector (reported in Table 1) significantly differs from the ones
352 achieved with the state of the art feature sets and with the proposed fea-
353 ture set using PCA. When significant differences were not found, the time

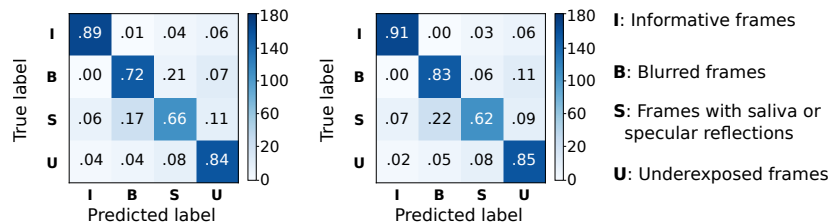


Figure 4: Confusion matrices with (left) and without (right) applying principal component analysis to the proposed feature set. Frame classification was obtained with support vector machines. Matrices refer to the classification of the balanced dataset of 720 narrow-band imaging laryngoscopic video frames. The colorbar indicates the number of images.

354 required to extract the features was computed. We also used the Wilcoxon
 355 signed-rank test to assess if the performance of SVM and RF in classifying
 356 the proposed feature set was significantly different. In all cases, we performed
 357 the Wilcoxon signed-rank test by comparing the \mathbf{Rec}_{class} vectors.

358 As for classifying the completely labeled video sequence, when training
 359 the SVM, all the frames from the three folds (excluding the ones relative to
 360 the specific analyzed video) were used, for a total of 689 training frames.

361 4. Results

362 With the proposed feature set and SVM classification, a median \mathbf{Rec}_{class}
 363 = 84% with inter-quartile range (IQR) = 9% was obtained (Table 3 bottom).
 364 It is worth noting that misclassification occurred mainly when classifying
 365 uninformative frames, while informative frames were classified with a recall
 366 of 91%. The relative confusion matrix is reported in Fig. 4 (right). From
 367 the ROC curve analysis (Fig. 5 left), a mean AUC of 91% was achieved.

Table 3: Classification performance of the proposed approach. Results are relative to support vector machines (SVM) and random forest (RF) classification on the proposed feature set. SVM results with principal component analysis (PCA) performed on the feature set are reported, too. Class-specific recall ($\mathbf{Rec}_{\text{class}}$), precision ($\mathbf{Prec}_{\text{class}}$), and F1 score ($\mathbf{F1}_{\text{class}}$) are reported for the four different frame classes. **I**: informative frame; **B**: blurred frame; **S**: frame with saliva or specular reflections; **U**: underexposed frame. Median and inter-quartile range (IQR) of the metrics are reported, too.

Proposed feature set and PCA + SVM						
	I	B	S	U	median	IQR
$\mathbf{Prec}_{\text{class}}$	0.90	0.77	0.66	0.78	0.78	0.13
$\mathbf{Rec}_{\text{class}}$	0.89	0.72	0.66	0.84	0.78	0.18
$\mathbf{F1}_{\text{class}}$	0.90	0.74	0.66	0.81	0.78	0.16
Proposed feature set + RF						
	I	B	S	U	median	IQR
$\mathbf{Prec}_{\text{class}}$	0.89	0.76	0.61	0.72	0.74	0.16
$\mathbf{Rec}_{\text{class}}$	0.78	0.73	0.59	0.86	0.76	0.16
$\mathbf{F1}_{\text{class}}$	0.83	0.75	0.60	0.78	0.77	0.13
Proposed feature set + SVM						
	I	B	S	U	median	IQR
$\mathbf{Prec}_{\text{class}}$	0.91	0.76	0.78	0.76	0.77	0.09
$\mathbf{Rec}_{\text{class}}$	0.91	0.83	0.62	0.85	0.84	0.16
$\mathbf{F1}_{\text{class}}$	0.91	0.79	0.69	0.80	0.80	0.12

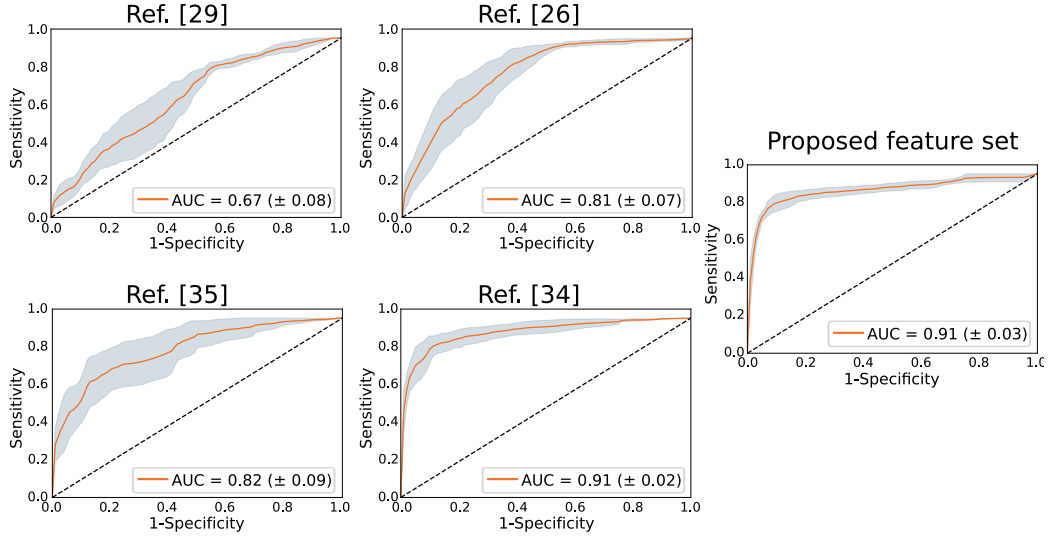


Figure 5: Macro-averaging receiver operating characteristic (ROC) curve analysis. ROC were obtained using support vector machines. No principal component analysis for feature reduction was performed. The mean (\pm standard deviation) curves obtained from the 3 cross-validation folds are reported by the orange solid lines (gray area). The mean (\pm standard deviation) area under the ROC curve is reported in the legend.

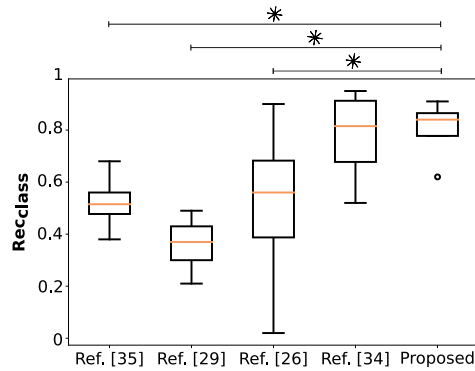


Figure 6: Comparison of the proposed feature set performance with the state of the art feature performance. Classification was performed using support vector machines. No principal component analysis for feature reduction was performed. Boxplots of class-specific recall ($\mathbf{Rec}_{\text{class}}$) are reported. Stars indicate significant differences (Wilcoxon signed-rank test (significance level $\alpha = 0.05$) for paired samples).

Table 4: Classification performance of the state of the art features using support vector machines. Class-specific recall ($\mathbf{Rec}_{\text{class}}$), precision ($\mathbf{Prec}_{\text{class}}$), and F1 score ($\mathbf{F1}_{\text{class}}$) are reported for the four different frame classes. **I**: informative frame; **B**: blurred frame; **S**: frame with saliva or specular reflections; **U**: underexposed frame. Median and interquartile range (IQR) of the metrics are reported, too. The parameters for the state of the art methods implemented for comparison were set as reported in their reference papers: oriented fast and rotated brief (ORB) parameters for [26], thresholding values for the intensity-based similarity score for [29], histogram bin number for [35] and orientation and radius for GLCM computation for [34].

Ref. [29]						
	I	B	S	U	median	IQR
$\mathbf{Prec}_{\text{class}}$	0.31	0.61	0.30	0.33	0.32	0.09
$\mathbf{Rec}_{\text{class}}$	0.49	0.41	0.33	0.21	0.37	0.13
$\mathbf{F1}_{\text{class}}$	0.38	0.49	0.31	0.26	0.35	0.11
Ref. [26]						
	I	B	S	U	median	IQR
$\mathbf{Prec}_{\text{class}}$	0.62	0.52	0.16	0.44	0.48	0.18
$\mathbf{Rec}_{\text{class}}$	0.61	0.90	0.02	0.51	0.56	0.30
$\mathbf{F1}_{\text{class}}$	0.62	0.66	0.04	0.47	0.55	0.27
Ref. [35]						
	I	B	S	U	median	IQR
$\mathbf{Prec}_{\text{class}}$	0.53	0.59	0.51	0.55	0.54	0.04
$\mathbf{Rec}_{\text{class}}$	0.38	0.68	0.52	0.61	0.57	0.14
$\mathbf{F1}_{\text{class}}$	0.45	0.63	0.52	0.58	0.55	0.09
Ref. [34]						
	I	B	S	U	median	IQR
$\mathbf{Prec}_{\text{class}}$	0.95	0.60	0.66	0.90	0.78	0.29
$\mathbf{Rec}_{\text{class}}$	0.95	0.73	0.52	0.90	0.81	0.30
$\mathbf{F1}_{\text{class}}$	0.95	0.66	0.58	0.90	0.78	0.31

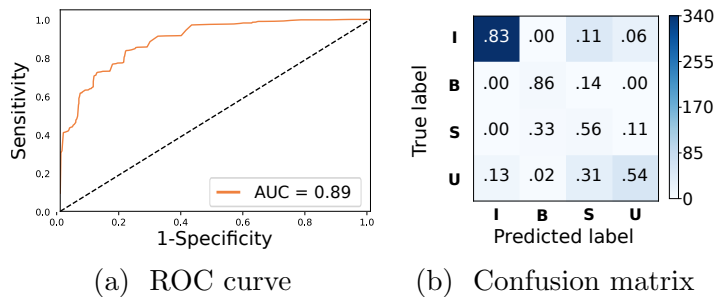


Figure 7: Classification performance of the proposed method for a complete video sequence. (a) Macro-averaging receiver operating characteristic (ROC) curve analysis. The area under the ROC curve is reported in the legend. (b) Confusion matrix for a complete video sequence. The colorbar indicates the number of frames. The number of frame for each class is 341 (**I**, informative frame); 7 (**B**, blurred frame); 9 (**S**, frame with saliva or specular reflections); 84 (**U**, underexposed frame).

368 The computational time for our feature set computation from one image was
 369 $\sim 0.03s$. The classification process took $\sim 10^{-5}s$.

370 When applying PCA (Table 3 top, Fig. 4 left), no significant differences
 371 were found with respect to using non-reduced features (p-value > 0.05). Since
 372 SVM performances with and without PCA were comparable, we decided to
 373 exclude PCA from our analysis.

374 When using RF (Table 3 middle) to classify our feature set, no significant
 375 difference (p-value > 0.05) were found with respect to SVM classification.

376 When applying the algorithm presented in [29] to our dataset, a mean
 377 AUC = 67% was obtained (Fig. 5). Worse performance with respect to the
 378 proposed approach was achieved also by the method in [26] (mean AUC =
 379 81%) and in [35] (mean AUC = 82%), while [34] achieved a comparable value
 380 of AUC = 91%. The **Rec_{class}**, **Prec_{class}**, **F1_{class}** values for [29] [26], [35] and
 381 [34] are reported in Table 4. The complete statistics of **Rec_{class}** relative to

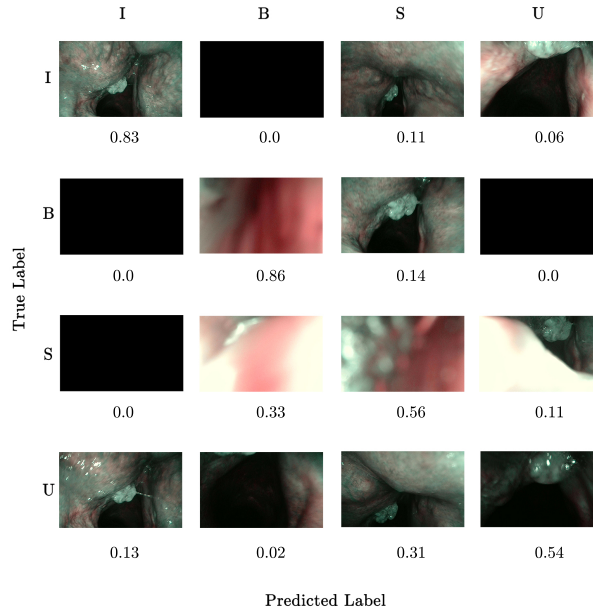


Figure 8: Visual confusion matrix for an entire laryngoscopic video sequence. Black boxes indicate the absence of misclassification between the true and predicted label. Numbers indicate the percentage of classified frames. **I**: informative frame; **B**: blurred frame; **S**: frame with saliva or specular reflections; **U**: underexposed frame.

382 the comparison of the proposed method with the state of the art is reported
 383 in Fig. 6. The proposed approach significantly outperformed [26, 35, 29]
 384 (p-value < 0.05). Comparable performances (p-value > 0.05) were instead
 385 achieved using GLCM as in [34]. The execution time to extract GLCM-based
 386 features from a single image using the scikit-image implementation [47] on
 387 the machine described in Sec. 3.1 was $\sim 0.71s$.

388 Results relative to the automatic classification of the complete video se-
 389 quence with the associated gold standard classification are reported in Fig. 7.
 390 The ROC curve is reported (AUC = 0.89), as well as the confusion matrix.

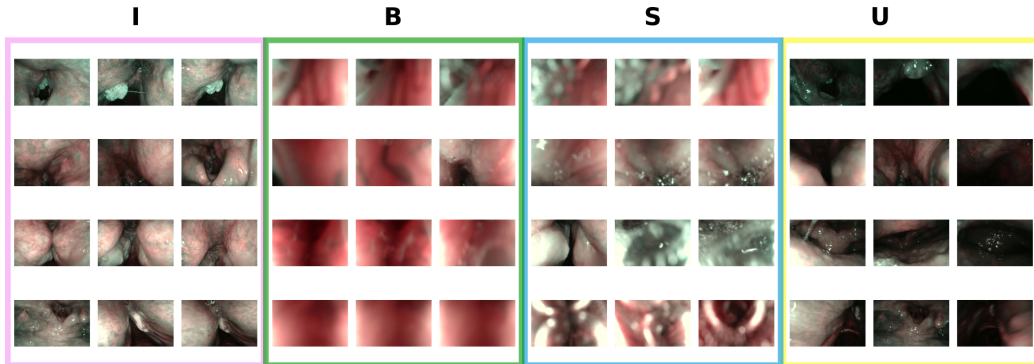


Figure 9: Qualitative analysis of the classification outcomes. Each row shows examples of frames from a video sequence. **I**: informative frame; **B**: blurred frame; **S**: frame with saliva or specular reflections; **U**: underexposed frame.

391 The 83% of the **I** frames were correctly classified. To qualitative appreciate
 392 the classification results, a visual confusion matrix is shown Fig. 8.

393 Visual samples of the classification performance for four videos are shown
 394 in Fig. 9.

395 5. Discussion

396 From the comparison with three state of the art methods, the proposed
 397 strategy proved to be a reliable and much better strategy for frame selection
 398 with respect to [26, 35, 29], with statistical evidence. Significant differences
 399 were not found when comparing the proposed performance with GLCM-
 400 based features [34]. However, it is worth noting that the GLCM computation
 401 time ($\sim 0.71s$) for one image was 1 order of magnitude higher than the
 402 computation time required to compute the proposed feature set ($\sim 0.03s$).
 403 This makes our feature set more suitable for the task of informative frame
 404 selection with respect to GLCM-based features considering that time is a

405 constraint with a view to real-time application. This is especially true if one
406 considers that the computational time of computer-aided diagnostic systems
407 must be eventually added. Moreover, compared with the state of the art,
408 the proposed learning-based method is simpler, as it eliminates the issue of
409 setting thresholding values (which is required by [15, 26, 29, 34]).

410 The SVM performance did not vary when applying feature selection with
411 PCA, suggesting that PCA can be avoided to avoid increasing computa-
412 tional time. Significant differences between RF and SVM were not found, as
413 expected considering results reported in the literature [48, 49, 50].

414 When testing the proposed approach on a complete labeled video se-
415 quence, the misclassifications occurred mainly for challenging frames, which
416 were not trivial to classify also for humans (Fig. 8). **I** frames were never
417 misclassified as **B**, while misclassification occurred with respect to **U** frames,
418 especially when the area between vocal folds covered a large portion of the
419 image, and to **S** frames, due to the presence in the frame of portions of
420 leukoplakia, which is visually similar to specular reflections. It is worth not-
421 ing that frames with leukoplakia were misclassified as uninformative only
422 when in presence of specular reflections and saliva. An example of a frame
423 that depicts a tissue with leukoplakia and that was correctly classified as
424 informative is shown in Fig. 8 (top-left). Nonetheless, we recognize that
425 the misclassification of informative frames is critical, as it could affect the
426 judgment of diagnosis. Training on a larger set, which would include a wider
427 range of laryngeal tissue conditions, should attenuate this issue. Moreover,

428 with a larger dataset, more advanced tools may also be investigated, such as
429 convolutional neural networks, following the current trends in non-medical
430 fields where large labeled dataset are available (e.g. [51, 52]). A further
431 solution could be exploring classification confidence estimation, as recently
432 investigated by the case-based reasoning community [53].

433 A limit of the proposed approach could be seen in the dimension of the
434 evaluation dataset. The number of frames of the balanced dataset used to test
435 the proposed approach was limited to ~ 700 images, to which ~ 500 frames
436 from the fully labeled video were added. Despite such number being much
437 smaller than those available for the methods used for performance comparison,
438 namely [26] (~ 3000 images) and [35] (~ 22000 images), it has the same
439 order of magnitude of other methods in the literature, such as [29] (~ 300
440 images). Moreover, it is worth noting that our dataset grants a more complete
441 overview on the inter-patient variability, presenting a higher number of
442 patients (18) compared to the 3 of [35], to the 2 of [26], and to the 6 of our
443 own previous study [15], which exploited the state of the art method in [29].
444 Therefore, to contribute to global research on laryngoscopic video analysis,
445 we decided to make our dataset fully available online.

446 Our evaluation protocol was focused on laryngeal video endoscopy, but
447 we expect similar results for other anatomical districts, such as the gastroin-
448 testinal and abdominal ones. We believe that the proposed methodology can
449 be easily and successfully integrated as pre-processing step for several ap-
450 plications, e.g. to provide informative sets of images for video stitching [8],

451 computer-aided diagnosis [54], tissue classification [55] and segmentation ap-
452 plications [56].

453 **6. Conclusion**

454 In this paper, we addressed the challenging topic of informative frame
455 classification in laryngoscopic videos. The method was retrospectively ap-
456 plied to ~ 1200 frames from 18 videos of 18 different subjects recorded during
457 the clinical practice. With our experimental protocol, an overall median clas-
458 sification recall of 84% among four frame classes (i.e. blurred, underexposed,
459 with saliva or specular reflections, and informative frames) was achieved.
460 Misclassification mainly occurred between classes of uninformative frames
461 and informative video frames were classified with a recall of 91%. Such perfor-
462 mances are significantly higher than those achieved applying other methods
463 in the literature to our evaluation dataset. Moreover, the proposed approach
464 is more robust, faster and simpler to implement since no parameter tuning is
465 required. It is recognized that future work is required to further ameliorate
466 the algorithm performance. However, the results obtained here are expected
467 to provide major contribution towards lowering the degree of manual in-
468 tervention required by computer-assisted systems intended to analyze and
469 summarize the endoscopic video content and increasing their performance.

470 **Disclosures**

471 The authors have no conflict of interest to disclose.

- 472 [1] P. Schultz, Vocal fold cancer, *European Annals of Otorhinolaryngology,*
473 *Head and Neck Diseases* 128 (2011) 301–308.
- 474 [2] J. Unger, J. Lohscheller, M. Reiter, K. Eder, C. S. Betz, M. Schuster, A
475 noninvasive procedure for early-stage discrimination of malignant and
476 precancerous vocal fold lesions based on laryngeal dynamics analysis,
477 *Cancer Research* 75 (2015) 31–39.
- 478 [3] C. Piazza, F. Del Bon, G. Peretti, P. Nicolai, Narrow band imaging in
479 endoscopic evaluation of the larynx, *Current Opinion in Otolaryngology*
480 *& Head and Neck Surgery* 20 (2012) 472–476.
- 481 [4] J. S. Isenberg, D. L. Crozier, S. H. Dailey, Institutional and compre-
482 hensive review of laryngeal leukoplakia, *Annals of Otology, Rhinology*
483 *& Laryngology* 117 (2008) 74–79.
- 484 [5] M. Schuster, T. Bergen, M. Reiter, C. Münzenmayer, S. Friedl, T. Wit-
485 tenberg, Laryngoscopic image stitching for view enhancement and
486 documentation—first experiences, *Biomedical Engineering* 57 (2012) 704–
487 707.
- 488 [6] C. Barbalata, L. S. Mattos, Laryngeal tumor detection and classification
489 in endoscopic video, *IEEE Journal of Biomedical and Health Informatics*
490 20 (2016) 322–332.
- 491 [7] S. Moccia, E. De Momi, M. Guarnaschelli, M. Savazzi, A. Laborai,
492 L. Guastini, G. Peretti, L. S. Mattos, Confident texture-based laryngeal

- 493 tissue classification for early stage diagnosis support, *Journal of Medical*
494 *Imaging* 4 (2017) 034502.
- 495 [8] T. Bergen, T. Wittenberg, Stitching and surface reconstruction from en-
496 doscopic image sequences: a review of applications and methods, *IEEE*
497 *Journal of Biomedical and Health Informatics* 20 (2016) 304–321.
- 498 [9] B. T. Truong, S. Venkatesh, Video abstraction: A systematic review
499 and classification, *ACM Transactions on Multimedia Computing, Com-*
500 *munications, and Applications* 3 (2007) 1–37.
- 501 [10] T. Vercauteren, Image registration and mosaicing for dynamic in vivo
502 fibered confocal microscopy, Ph.D. thesis, École Nationale Supérieure
503 des Mines de Paris, 2008.
- 504 [11] P. C. Cattin, H. Bay, L. Van Gool, G. Székely, Retina mosaicing using
505 local features, in: *International Conference on Medical Image Comput-*
506 *ing and Computer-Assisted Intervention*, Springer, pp. 185–192.
- 507 [12] A. Behrens, T. Stehle, S. Gross, T. Aach, Local and global panoramic
508 imaging for fluorescence bladder endoscopy, in: *Annual International*
509 *Conference of the IEEE Engineering in Medicine and Biology Society*,
510 *IEEE*, pp. 6990–6993.
- 511 [13] A. Behrens, Creating panoramic images for bladder fluorescence en-
512 doscopy, *Acta Polytechnica* 48 (2008).

- 513 [14] A. Behrens, M. Bommers, T. Stehle, S. Gross, S. Leonhardt, T. Aach,
514 Real-time image composition of bladder mosaics in fluorescence en-
515 doscopy, *Computer Science-Research and Development* 26 (2011) 51–64.
- 516 [15] S. Moccia, V. Penza, G. O. Vanone, E. De Momi, L. S. Mattos, Auto-
517 matic workflow for narrow-band laryngeal video stitching, in: *Annual*
518 *International Conference of the Engineering in Medicine and Biology*
519 *Society*, IEEE, pp. 1188–1191.
- 520 [16] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, Image quality
521 assessment: from error visibility to structural similarity, *IEEE Transac-*
522 *tions on Image Processing* 13 (2004) 600–612.
- 523 [17] T. Weibel, C. Daul, D. Wolf, R. Rösch, F. Guillemin, Graph based
524 construction of textured large field of view mosaics for bladder cancer
525 diagnosis, *Pattern Recognition* 45 (2012) 4138–4150.
- 526 [18] K. Schoeffmann, M. Del Fabro, T. Szkaliczki, L. Böszörményi, J. Keck-
527 stein, Keyframe extraction in endoscopic video, *Multimedia Tools and*
528 *Applications* 74 (2015) 11187–11206.
- 529 [19] U. von Öshen, et al., Key frame selection for robust pose estimation in
530 laparoscopic videos, in: *Proceedings of SPIE*, volume 8316, p. 83160Y.
- 531 [20] V. Hai, T. Echigo, R. Sagawa, K. Yagi, M. Shiba, K. Higuchi,
532 T. Arakawa, Y. Yagi, Adaptive control of video display for diagnostic

- 533 assistance by analysis of capsule endoscopic images, in: International
534 Conference on Pattern Recognition, volume 3, IEEE, pp. 980–983.
- 535 [21] I. Mehmood, M. Sajjad, S. W. Baik, Video summarization based tele-
536 endoscopy: a service to efficiently manage visual data generated during
537 wireless capsule endoscopy procedure, *Journal of Medical Systems* 38
538 (2014) 1–9.
- 539 [22] Y. Yuan, M. Q.-H. Meng, Hierarchical key frames extraction for WCE
540 video, in: *Mechatronics and Automation (ICMA), 2013 IEEE Interna-*
541 *tional Conference on, IEEE*, pp. 225–229.
- 542 [23] Q. Zhao, M. Q.-H. Meng, WCE video abstracting based on novel color
543 and texture features, in: *International Conference on Robotics and*
544 *Biomimetics, IEEE*, pp. 455–459.
- 545 [24] M. Lux, O. Marques, K. Schöffmann, L. Böszörményi, G. Lajtai, A
546 novel tool for summarization of arthroscopic videos, *Multimedia Tools*
547 *and Applications* 46 (2010) 521–544.
- 548 [25] D. K. Iakovidis, S. Tsevas, A. Polydorou, Reduction of capsule en-
549 doscopy reading times by unsupervised image mining, *Computerized*
550 *Medical Imaging and Graphics* 34 (2010) 471–478.
- 551 [26] T. Bergen, P. Hastreiter, C. Münzenmayer, M. Buchfelder, T. Witten-
552 berg, Image stitching of sphenoid sinuses from monocular endoscopic

- 553 views, in: College and University Retiree Associations of Canada, pp.
554 226–229.
- 555 [27] A. Ishijima, R. A. Schwarz, D. Shin, S. Mondrik, N. Vigneswaran, A. M.
556 Gillenwater, S. Anandasabapathy, R. Richards-Kortum, Automated
557 frame selection process for high-resolution microendoscopy, *Journal of*
558 *Biomedical Optics* 20 (2015) 046014–046014.
- 559 [28] M. A. Armin, G. Chetty, F. Jurgen, H. De Visser, C. Dumas, A. Fa-
560 zlollahi, F. Grimpen, O. Salvado, Uninformative frame detection in
561 colonoscopy through motion, edge and color features, in: *International*
562 *Workshop on Computer-Assisted and Robotic Endoscopy*, Springer, pp.
563 153–162.
- 564 [29] F. Crete, T. Dolmiere, P. Ladret, M. Nicolas, The blur effect: percep-
565 tion and estimation with a new no-reference perceptual blur metric, in:
566 *Human Vision and Electronic Imaging*, volume 12, pp. EI–6492.
- 567 [30] S. Y. Park, D. Sargent, I. Spofford, K. G. Vosburgh, A. Yousif, et al., A
568 colon video analysis framework for polyp detection, *IEEE Transactions*
569 *on Biomedical Engineering* 59 (2012) 1408–1418.
- 570 [31] C.-F. J. Kuo, Y.-H. Chu, P.-C. Wang, C.-Y. Lai, W.-L. Chu, Y.-S. Leu,
571 H.-W. Wang, Using image processing technology and mathematical al-
572 gorithm in the automatic selection of vocal cord opening and closing

- 573 images from the larynx endoscopy video, *Computer Methods and Pro-*
574 *grams in Biomedicine* 112 (2013) 455–465.
- 575 [32] S. Atasoy, D. Mateus, A. Meining, G.-Z. Yang, N. Navab, Endoscopic
576 video manifolds for targeted optical biopsy, *IEEE Transactions on Med-*
577 *ical Imaging* 31 (2012) 637–653.
- 578 [33] O. H. Maghsoudi, A. Talebpour, H. Soltanian-Zadeh, M. Alizadeh, H. A.
579 Soleimani, Informative and uninformative regions detection in WCE
580 frames, *Journal of Advanced Computing* 3 (2014) 12–34.
- 581 [34] A. Perperidis, A. Akram, Y. Altmann, P. McCool, J. Westerfeld, D. Wil-
582 son, K. Dhaliwal, S. McLaughlin, Automated detection of uninforma-
583 tive frames in pulmonary optical endomicroscopy, *IEEE Transactions*
584 *on Biomedical Engineering* 64 (2017) 87–98.
- 585 [35] M. K. Bashar, K. Mori, Y. Suenaga, T. Kitasaka, Y. Mekada, Detecting
586 informative frames from wireless capsule endoscopic video using color
587 and texture features, in: *International Conference on Medical Image*
588 *Computing and Computer-Assisted Intervention*, Springer, pp. 603–610.
- 589 [36] M. K. Bashar, T. Kitasaka, Y. Suenaga, Y. Mekada, K. Mori, Automatic
590 detection of informative frames from wireless capsule endoscopy images,
591 *Medical Image Analysis* 14 (2010) 449–470.
- 592 [37] A. Mittal, A. K. Moorthy, A. C. Bovik, No-reference image quality as-

- 593 assessment in the spatial domain, *IEEE Transactions on Image Processing*
594 21 (2012) 4695–4708.
- 595 [38] J. L. Pech-Pacheco, G. Cristóbal, J. Chamorro-Martinez, J. Fernández-
596 Valdivia, Diatom autofocusing in brightfield microscopy: a comparative
597 study, in: *International Conference on Pattern Recognition*, volume 3,
598 IEEE, pp. 314–317.
- 599 [39] J. M. Mateos-Pérez, R. Redondo, R. Nava, J. C. Valdiviezo,
600 G. Cristóbal, B. Escalante-Ramírez, M. J. Ruiz-Serrano, J. Pascau,
601 M. Desco, Comparative evaluation of autofocus algorithms for a real-
602 time system for automatic detection of mycobacterium tuberculosis, *Cy-
603 tometry* 81 (2012) 213–221.
- 604 [40] A. Rényi, On measures of entropy and information, in: *Proceedings of*
605 *the Fourth Berkeley Symposium on Mathematical Statistics and Prob-
606 ability*, volume 1, pp. 547–561.
- 607 [41] L. Liu, B. Liu, H. Huang, A. C. Bovik, No-reference image quality
608 assessment based on spatial and spectral entropies, *Signal Processing:
609 Image Communication* 29 (2014) 856–863.
- 610 [42] E. Rublee, V. Rabaud, K. Konolige, G. Bradski, ORB: An efficient
611 alternative to SIFT or SURF, in: *IEEE International Conference on*
612 *Computer Vision*, IEEE, pp. 2564–2571.

- 613 [43] A. M. Mendrik, E.-J. Vonken, A. Rutten, M. A. Viergever, B. van
614 Ginneken, Noise reduction in computed tomography scans using 3-D
615 anisotropic hybrid diffusion with continuous switch, *IEEE Transactions*
616 *on Medical Imaging* 28 (2009) 1585–1594.
- 617 [44] C. J. Burges, A tutorial on support vector machines for pattern recog-
618 nition, *Data Mining and Knowledge Discovery* 2 (1998) 121–167.
- 619 [45] S. Wold, K. Esbensen, P. Geladi, Principal component analysis, *Chemo-*
620 *metrics and Intelligent Laboratory Systems* 2 (1987) 37–52.
- 621 [46] L. Breiman, Random forests, *Machine learning* 45 (2001) 5–32.
- 622 [47] S. Van der Walt, J. L. Schönberger, J. Nunez-Iglesias, F. Boulogne, J. D.
623 Warner, N. Yager, E. Gouillart, T. Yu, scikit-image: image processing
624 in python, *PeerJ* 2 (2014) e453.
- 625 [48] D. C. Duro, S. E. Franklin, M. G. Dubé, A comparison of pixel-based
626 and object-based image analysis with selected machine learning algo-
627 rithms for the classification of agricultural landscapes using SPOT-5
628 HRG imagery, *Remote Sensing of Environment* 118 (2012) 259–272.
- 629 [49] A. Bosch, A. Zisserman, X. Munoz, Image classification using random
630 forests and ferns, in: *International Conference on Computer Vision*,
631 *IEEE*, pp. 1–8.
- 632 [50] G. Csurka, C. Dance, L. Fan, J. Willamowski, C. Bray, Visual catego-

- 633 rization with bags of keypoints, in: Workshop on Statistical Learning
634 in Computer Vision, volume 1, Prague, pp. 1–2.
- 635 [51] L. Kang, P. Ye, Y. Li, D. Doermann, Convolutional neural networks
636 for no-reference image quality assessment, in: Proceedings of the IEEE
637 Conference on Computer Vision and Pattern Recognition, IEEE, 2014,
638 pp. 1733–1740.
- 639 [52] J. Li, L. Zou, J. Yan, D. Deng, T. Qu, G. Xie, No-reference image qual-
640 ity assessment using Prewitt magnitude based on convolutional neural
641 networks, *Signal, Image and Video Processing* 10 (2016) 609–616.
- 642 [53] J. Kolodner, *Case-based reasoning*, Morgan Kaufmann, 2014.
- 643 [54] M. Liedlgruber, A. Uhl, Computer-aided decision support systems for
644 endoscopy in the gastrointestinal tract: a review, *IEEE Reviews in*
645 *Biomedical Engineering* 4 (2011) 73–88.
- 646 [55] S. Moccia, S. J. Wirkert, H. Kenngott, A. S. Vemuri, M. Apitz, B. Mayer,
647 E. De Momi, L. Mattos, L. Maier-Hein, Uncertainty-aware organ clas-
648 sification for surgical data science applications in laparoscopy, *arXiv*
649 *preprint arXiv:1706.07002* (2017).
- 650 [56] M. S. Nosrati, J.-M. Peyrat, J. Abinshed, O. Al-Alao, A. Al-Ansari,
651 R. Abugharbieh, G. Hamarneh, Efficient multi-organ segmentation in
652 multi-view endoscopic videos using pre-operative priors, in: Interna-

653 tional Conference on Medical Image Computing and Computer-Assisted
654 Intervention, Springer, pp. 324–331.