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)

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 us 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-

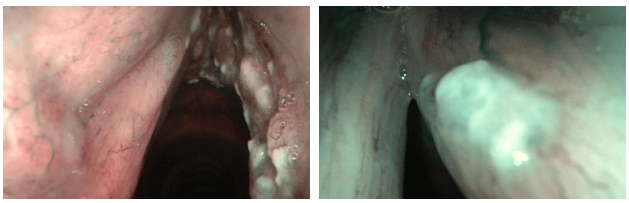


Figure 1: Sample images of vocal folds affected by leukoplakia, a pre-cancerous tissue alteration which causes vocal fold epithelium whitening and thickening.

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

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

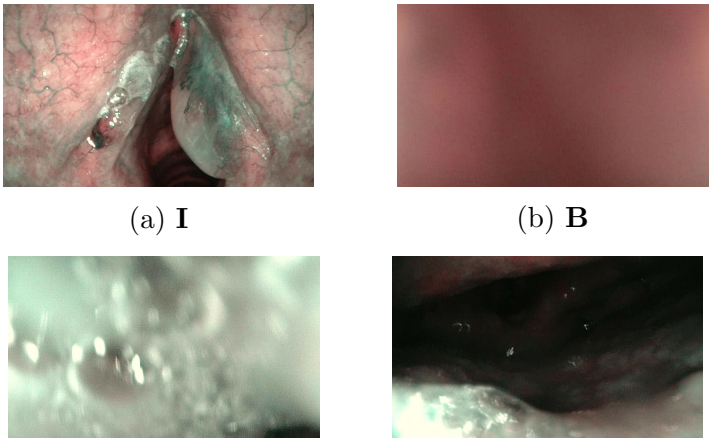
Developing a strategy to select informative frames has the potential to reduce the amount of data to review, lowering the surgeons' workload. The selection of informative frames can be beneficial also for computer-aided diagnosis algorithms. Preliminary efforts towards the automatic classification of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

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

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

A possible solution to lower the number of uninformative frames is exploiting preliminary visual image quality assessment through subjective evaluation. This operation is, however, prone to human error and usually too inconvenient and time-consuming [16].

Automatic selection of informative frames is a valuable alternative. In



(a) **I** (b) **B**

(c) **S** (d) **U**

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**).

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

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

1.1. Related work

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

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

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

1.1.1. Video clustering and keyframe extraction

The goal of this class of algorithm is to cluster video frames with similar informative content, exploiting similarity measures between features extracted from the images.

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

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

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

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

This class of algorithms potentially brings the advantage of summarizing the video content. Nonetheless, such algorithms do not make any assumptions about the presence of uninformative video portions, which can potentially represent a high percentage of the endoscopic video content without bringing any useful information for diagnosis.

1.1.2. Uninformative frame removal

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

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

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

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

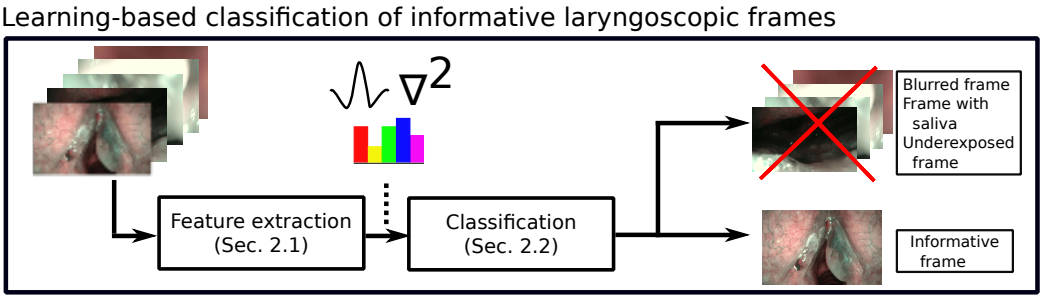


Figure 3: Workflow of the proposed approach to automatic learning-based classification of informative laryngoscopy frames.

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

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

1.2. Aim of the work

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

- Underexposed frames (**U**)
- Frames with saliva or specular reflections (**S**)
- Blurred frames (**B**)
- Informative frames (**I**)

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

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

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

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

2. Methods

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

2.1. Feature extraction

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

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

Part of the feature set used in this work was borrowed from state of the 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 (*BRISQUE*) [37]	1
Variance of the image Laplacian (Δ_{VAR})	3
Sobel-Tenengrad focus evaluation function score (*TEN*) [38]	3
Image entropy (*ENTROPY*)	3
Local variance of the luminance channel intensity (*VAR*)	3
Image intensity variance (*G_VAR*)	3
Image histogram (*H*)	3
Number of detected keypoints (*N_P*)	1
Total	20

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

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

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

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

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

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

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

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

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

$$\Delta_{VAR} = \sum_{m}^{M} \sum_{n}^{N} (L(m,n) - \bar{L})^2 \tag{2}$$

  where $\bar{L}$ is:

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

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

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

- Sobel-Tenengrad focus evaluation function (*TEN*):

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

$$TEN = \sum_{m}^{M} \sum_{n}^{N} [S(m,n)]^2, \qquad \text{for } S(m,n) > T \tag{5}$$

where T is a threshold and:

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

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

$$K_S = \begin{bmatrix} 1 & 0 & -1 \\ 2 & 0 & -2 \\ 1 & 0 & -1 \end{bmatrix} \tag{7}$$

- Image entropy ($ENTROPY$):

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

$$ENTROPY = -\sum_{i} h_i \log_2(h_i) \tag{8}$$

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

- Local variance of the luminance channel intensity (VAR):

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

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

where:

$$lv(m,n) = \frac{1}{w_x w_y} \sum_{i}^{w_x} \sum_{j}^{w_y} [I(m+i, n+j) - \bar{I_w}]^2 \tag{10}$$

$$\bar{lv} = \frac{1}{MN} \sum_{m}^{M} \sum_{n}^{N} lv(m,n) \tag{11}$$

and $\bar{I_w}$ is the mean intensity value on the window.

- Image intensity variance (G_VAR):

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

$$G\_VAR = \frac{1}{MN} \sum_{m}^{M} \sum_{n}^{N} \left[I(m,n) - \bar{I} \right]^2 \tag{12}$$

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

- Image histogram (H):

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

- Number of keypoints (N_P):

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

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

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

2.2. Classification

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

$$f(\mathbf{x}) = sign\Big[\sum_{k=1}^{N} a_k^* y_k \Psi(\mathbf{x}, \mathbf{x_k}) + b\Big] \tag{13}$$

where:

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

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

$$a_k^* = \max\Big\{ -\frac{1}{2} \sum_{k,l=1}^{N} y_k y_l \Psi(\mathbf{x_k}, \mathbf{x_l}) a_k a_l + \sum_{k=1}^{N} a_k \Big\} \tag{15}$$

with:

$$\sum_{k=1}^{N} a_k y_k = 0, \qquad 0 \leq a_k \leq C, \qquad k = 1, ..., N \tag{16}$$

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

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

3. Evaluation

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

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

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

In addition, one of the videos was fully labeled (length = 17.64s; number of frames: $\mathbf{I} = 341$, $\mathbf{B} = 7$, $\mathbf{S} = 9$, $\mathbf{U} = 84$).

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

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

- *BRISQUE*: *BRISQUE* code was downloaded from the *Laboratory for Image & Video Engineering* website[1] and the parameters were set

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

as in [37]

- $\underline{TEN}$: the threshold T of Eq. 5 was set to 0, as suggested in [39] to include all pixels in the computation
- $\underline{VAR}$: to compute VAR, the local window size was 5×5 pixels
- $\underline{N\_P}$: the parameters of ORB were set as in the original paper [42]

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

The feature computation was implemented using OpenCV [2]. The classification was implemented with scikit-learn [3].

3.1. Experimental setup

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

[2] [http://docs.opencv.org /3.1.0/index.html]
[3] [http://scikit-learn.org/stable/ index.html]

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

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

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

$$Rec_{class_j} = \frac{TP_j}{TP_j + FN_j} \tag{17}$$

$$Prec_{class_j} = \frac{TP_j}{TP_j + FP_j} \tag{18}$$

$$F1_{class_j} = 2\ \frac{Prec_{class_j} \times Rec_{class_j}}{Prec_{class_j} + Rec_{class_j}} \tag{19}$$

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

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

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

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

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

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

The SVM performance were compared also with those obtained using the features commonly exploited in the state of the art. As explained in Sec. 1.1.2, commonly exploited features are (i) image keypoints, (ii) intensity-based similarity score, (iii) color features and (iv) textural information. Therefore, we decided to compare our method with one research per feature category. We considered [26, 29, 35, 34], which use ORB keypoints, intensity-based similarity score, color histogram and GLCM, respectively. The parameters for the state of the art methods implemented for comparison were set as reported in their reference papers: 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]. As stated in Sec. 1, such methods rely on thresholding instead of machine learning-based methods. However, the features from the state of the art were classified with SVM, for fair comparison.

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

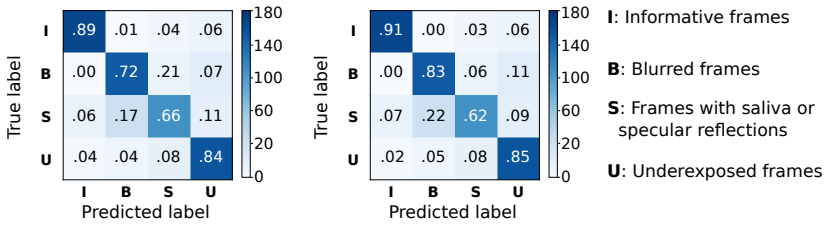


Figure 4: Confusion matrices with (left) and without (right) applying principal component analysis to the proposed feature set. Frame classification was obtained with support vector machines. Matrices refer to the classification of the balanced dataset of 720 narrow-band imaging laryngoscopic video frames.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

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

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

4. Results

With the proposed feature set and SVM classification, a median $\mathbf{Rec_{class}}$ = 84% with inter-quartile range (IQR) = 9% was obtained (Table 3 bottom). It is worth noting that misclassification occurred mainly when classifying uninformative frames, while informative frames were classified with a recall of 91%. The relative confusion matrix is reported in Fig. 4 (right). From 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 ($\mathbf{Rec_{class}}$), precision ($\mathbf{Prec_{class}}$), and F1 score ($\mathbf{F1_{class}}$) 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
$\mathbf{Prec_{class}}$	0.90	0.77	0.66	0.78	0.78	0.13
$\mathbf{Rec_{class}}$	0.89	0.72	0.66	0.84	0.78	0.18
$\mathbf{F1_{class}}$	0.90	0.74	0.66	0.81	0.78	0.16
Proposed feature set + RF						
	I	**B**	**S**	**U**	median	IQR
$\mathbf{Prec_{class}}$	0.89	0.76	0.61	0.72	0.74	0.16
$\mathbf{Rec_{class}}$	0.78	0.73	0.59	0.86	0.76	0.16
$\mathbf{F1_{class}}$	0.83	0.75	0.60	0.78	0.77	0.13
Proposed feature set + SVM						
	I	**B**	**S**	**U**	median	IQR
$\mathbf{Prec_{class}}$	0.91	0.76	0.78	0.76	0.77	0.09
$\mathbf{Rec_{class}}$	0.91	0.83	0.62	0.85	0.84	0.16
$\mathbf{F1_{class}}$	0.91	0.79	0.69	0.80	0.80	0.12

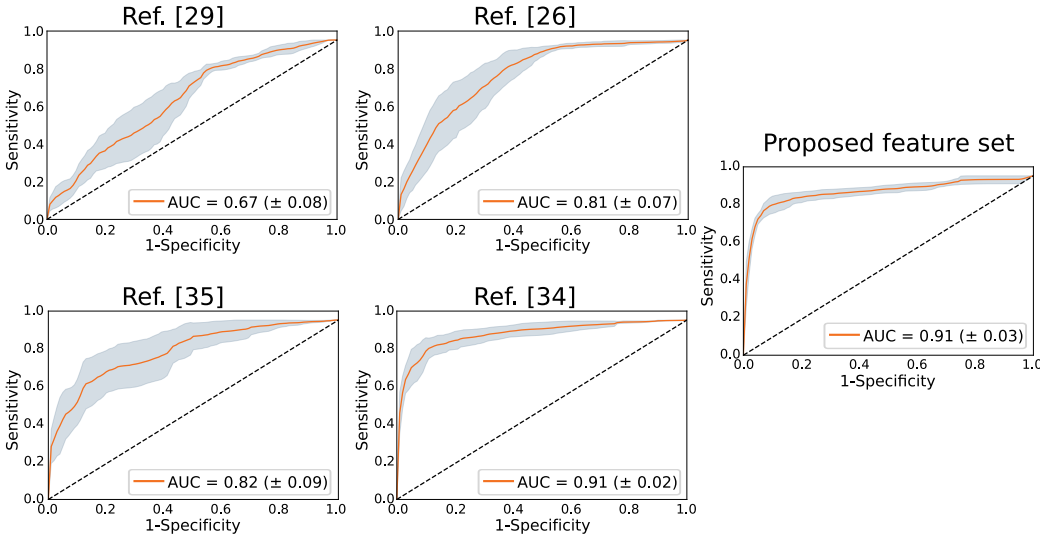


Figure 5: Macro-averaging receiver operating characteristic (ROC) curve analysis. ROC were obtained using support vector machines. No principal component analysis for feature reduction was performed. The mean (± standard deviation) curves obtained from the 3 cross-validation folds are reported by the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

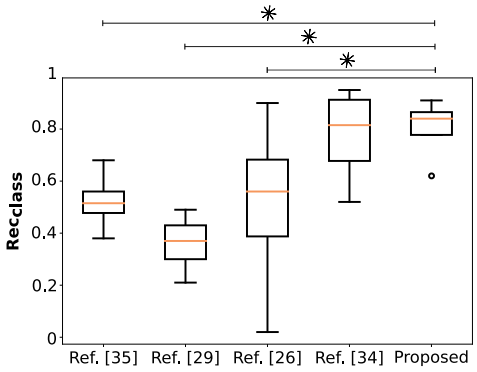
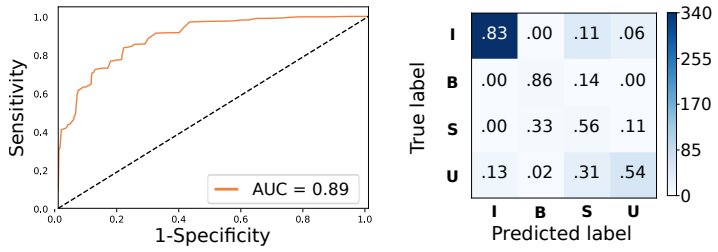


Figure 6: Comparison of the proposed feature set performance with the state of the art feature performance. Classification was performed using support vector machines. No principal component analysis for feature reduction was performed. Boxplots of class-specific recall ($\mathbf{Rec_{class}}$) are reported. 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 ($\mathbf{Rec_{class}}$), precision ($\mathbf{Prec_{class}}$), and F1 score ($\mathbf{F1_{class}}$) 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
$\mathbf{Prec_{class}}$	0.31	0.61	0.30	0.33	0.32	0.09
$\mathbf{Rec_{class}}$	0.49	0.41	0.33	0.21	0.37	0.13
$\mathbf{F1_{class}}$	0.38	0.49	0.31	0.26	0.35	0.11
Ref. [26]						
	I	**B**	**S**	**U**	median	IQR
$\mathbf{Prec_{class}}$	0.62	0.52	0.16	0.44	0.48	0.18
$\mathbf{Rec_{class}}$	0.61	0.90	0.02	0.51	0.56	0.30
$\mathbf{F1_{class}}$	0.62	0.66	0.04	0.47	0.55	0.27
Ref. [35]						
	I	**B**	**S**	**U**	median	IQR
$\mathbf{Prec_{class}}$	0.53	0.59	0.51	0.55	0.54	0.04
$\mathbf{Rec_{class}}$	0.38	0.68	0.52	0.61	0.57	0.14
$\mathbf{F1_{class}}$	0.45	0.63	0.52	0.58	0.55	0.09
Ref. [34]						
	I	**B**	**S**	**U**	median	IQR
$\mathbf{Prec_{class}}$	0.95	0.60	0.66	0.90	0.78	0.29
$\mathbf{Rec_{class}}$	0.95	0.73	0.52	0.90	0.81	0.30
$\mathbf{F1_{class}}$	0.95	0.66	0.58	0.90	0.78	0.31



(a) ROC curve (b) Confusion matrix

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).

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

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

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

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

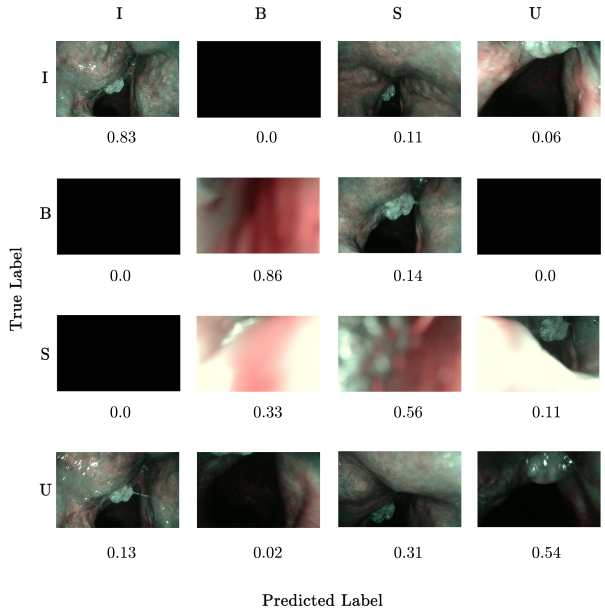


Figure 8: Visual confusion matrix for an entire laryngoscopic video sequence. Black boxes indicate the absence of misclassification between the true and predicted label. Numbers indicate the percentage of classified frames. **I**: informative frame; **B**: blurred frame; **S**: frame with saliva or specular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

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

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

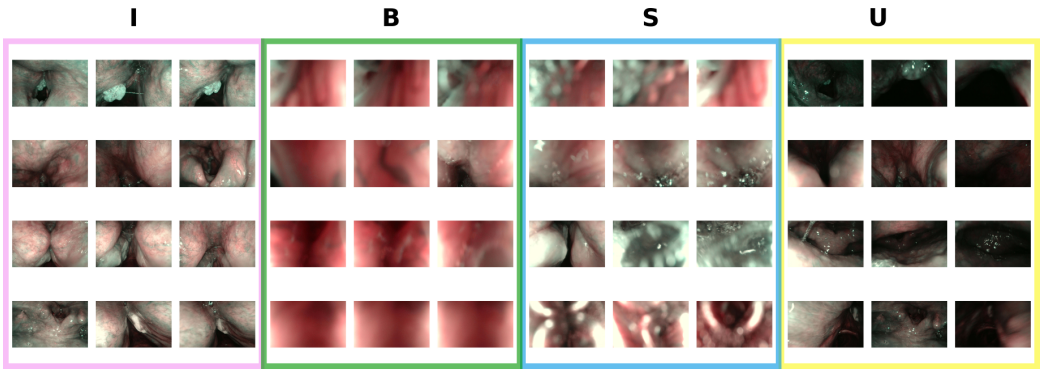


Figure 9: Qualitative analysis of the classification outcomes. Each row shows examples of frames from a video sequence. **I**: informative frame; **B**: blurred frame; **S**: frame with saliva or specular reflections; **U**: underexposed frame.

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

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

5. Discussion

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

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

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

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

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

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

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

computer-aided diagnosis [54], tissue classification [55] and segmentation applications [56].

6. Conclusion

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

Disclosures

The authors have no conflict of interest to disclose.

[1] P. Schultz, Vocal fold cancer, European Annals of Otorhinolaryngology, Head and Neck Diseases 128 (2011) 301–308.

[2] J. Unger, J. Lohscheller, M. Reiter, K. Eder, C. S. Betz, M. Schuster, A noninvasive procedure for early-stage discrimination of malignant and precancerous vocal fold lesions based on laryngeal dynamics analysis, Cancer Research 75 (2015) 31–39.

[3] C. Piazza, F. Del Bon, G. Peretti, P. Nicolai, Narrow band imaging in endoscopic evaluation of the larynx, Current Opinion in Otolaryngology & Head and Neck Surgery 20 (2012) 472–476.

[4] J. S. Isenberg, D. L. Crozier, S. H. Dailey, Institutional and comprehensive review of laryngeal leukoplakia, Annals of Otology, Rhinology & Laryngology 117 (2008) 74–79.

[5] M. Schuster, T. Bergen, M. Reiter, C. Münzenmayer, S. Friedl, T. Wittenberg, Laryngoscopic image stitching for view enhancement and documentation–first experiences, Biomedical Engineering 57 (2012) 704–707.

[6] C. Barbalata, L. S. Mattos, Laryngeal tumor detection and classification in endoscopic video, IEEE Journal of Biomedical and Health Informatics 20 (2016) 322–332.

[7] S. Moccia, E. De Momi, M. Guarnaschelli, M. Savazzi, A. Laborai, L. Guastini, G. Peretti, L. S. Mattos, Confident texture-based laryngeal

tissue classification for early stage diagnosis support, Journal of Medical Imaging 4 (2017) 034502.

[8] T. Bergen, T. Wittenberg, Stitching and surface reconstruction from endoscopic image sequences: a review of applications and methods, IEEE Journal of Biomedical and Health Informatics 20 (2016) 304–321.

[9] B. T. Truong, S. Venkatesh, Video abstraction: A systematic review and classification, ACM Transactions on Multimedia Computing, Communications, and Applications 3 (2007) 1–37.

[10] T. Vercauteren, Image registration and mosaicing for dynamic in vivo fibered confocal microscopy, Ph.D. thesis, École Nationale Supérieure des Mines de Paris, 2008.

[11] P. C. Cattin, H. Bay, L. Van Gool, G. Székely, Retina mosaicing using local features, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, pp. 185–192.

[12] A. Behrens, T. Stehle, S. Gross, T. Aach, Local and global panoramic imaging for fluorescence bladder endoscopy, in: Annual International Conference of the IEEE Engineering in Medicine and Biology Society, IEEE, pp. 6990–6993.

[13] A. Behrens, Creating panoramic images for bladder fluorescence endoscopy, Acta Polytechnica 48 (2008).

[14] A. Behrens, M. Bommes, T. Stehle, S. Gross, S. Leonhardt, T. Aach, Real-time image composition of bladder mosaics in fluorescence endoscopy, Computer Science-Research and Development 26 (2011) 51–64.

[15] S. Moccia, V. Penza, G. O. Vanone, E. De Momi, L. S. Mattos, Automatic workflow for narrow-band laryngeal video stitching, in: Annual International Conference of the Engineering in Medicine and Biology Society, IEEE, pp. 1188–1191.

[16] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Transactions on Image Processing 13 (2004) 600–612.

[17] T. Weibel, C. Daul, D. Wolf, R. Rösch, F. Guillemin, Graph based construction of textured large field of view mosaics for bladder cancer diagnosis, Pattern Recognition 45 (2012) 4138–4150.

[18] K. Schoeffmann, M. Del Fabro, T. Szkaliczki, L. Böszörmenyi, J. Keckstein, Keyframe extraction in endoscopic video, Multimedia Tools and Applications 74 (2015) 11187–11206.

[19] U. von Öshen, et al., Key frame selection for robust pose estimation in laparoscopic videos, in: Proceedings of SPIE, volume 8316, p. 83160Y.

[20] V. Hai, T. Echigo, R. Sagawa, K. Yagi, M. Shiba, K. Higuchi, T. Arakawa, Y. Yagi, Adaptive control of video display for diagnostic

assistance by analysis of capsule endoscopic images, in: International Conference on Pattern Recognition, volume 3, IEEE, pp. 980–983.

[21] I. Mehmood, M. Sajjad, S. W. Baik, Video summarization based tele-endoscopy: a service to efficiently manage visual data generated during wireless capsule endoscopy procedure, Journal of Medical Systems 38 (2014) 1–9.

[22] Y. Yuan, M. Q.-H. Meng, Hierarchical key frames extraction for WCE video, in: Mechatronics and Automation (ICMA), 2013 IEEE International Conference on, IEEE, pp. 225–229.

[23] Q. Zhao, M. Q.-H. Meng, WCE video abstracting based on novel color and texture features, in: International Conference on Robotics and Biomimetics, IEEE, pp. 455–459.

[24] M. Lux, O. Marques, K. Schöffmann, L. Böszörmenyi, G. Lajtai, A novel tool for summarization of arthroscopic videos, Multimedia Tools and Applications 46 (2010) 521–544.

[25] D. K. Iakovidis, S. Tsevas, A. Polydorou, Reduction of capsule endoscopy reading times by unsupervised image mining, Computerized Medical Imaging and Graphics 34 (2010) 471–478.

[26] T. Bergen, P. Hastreiter, C. Münzenmayer, M. Buchfelder, T. Wittenberg, Image stitching of sphenoid sinuses from monocular endoscopic

views, in: College and University Retiree Associations of Canada, pp. 226–229.

[27] A. Ishijima, R. A. Schwarz, D. Shin, S. Mondrik, N. Vigneswaran, A. M. Gillenwater, S. Anandasabapathy, R. Richards-Kortum, Automated frame selection process for high-resolution microendoscopy, Journal of Biomedical Optics 20 (2015) 046014–046014.

[28] M. A. Armin, G. Chetty, F. Jurgen, H. De Visser, C. Dumas, A. Fazlollahi, F. Grimpen, O. Salvado, Uninformative frame detection in colonoscopy through motion, edge and color features, in: International Workshop on Computer-Assisted and Robotic Endoscopy, Springer, pp. 153–162.

[29] F. Crete, T. Dolmiere, P. Ladret, M. Nicolas, The blur effect: perception and estimation with a new no-reference perceptual blur metric, in: Human Vision and Electronic Imaging, volume 12, pp. EI–6492.

[30] S. Y. Park, D. Sargent, I. Spofford, K. G. Vosburgh, A. Yousif, et al., A colon video analysis framework for polyp detection, IEEE Transactions on Biomedical Engineering 59 (2012) 1408–1418.

[31] C.-F. J. Kuo, Y.-H. Chu, P.-C. Wang, C.-Y. Lai, W.-L. Chu, Y.-S. Leu, H.-W. Wang, Using image processing technology and mathematical algorithm in the automatic selection of vocal cord opening and closing

images from the larynx endoscopy video, Computer Methods and Programs in Biomedicine 112 (2013) 455–465.

[32] S. Atasoy, D. Mateus, A. Meining, G.-Z. Yang, N. Navab, Endoscopic video manifolds for targeted optical biopsy, IEEE Transactions on Medical Imaging 31 (2012) 637–653.

[33] O. H. Maghsoudi, A. Talebpour, H. Soltanian-Zadeh, M. Alizadeh, H. A. Soleimani, Informative and uninformative regions detection in WCE frames, Journal of Advanced Computing 3 (2014) 12–34.

[34] A. Perperidis, A. Akram, Y. Altmann, P. McCool, J. Westerfeld, D. Wilson, K. Dhaliwal, S. McLaughlin, Automated detection of uninformative frames in pulmonary optical endomicroscopy, IEEE Transactions on Biomedical Engineering 64 (2017) 87–98.

[35] M. K. Bashar, K. Mori, Y. Suenaga, T. Kitasaka, Y. Mekada, Detecting informative frames from wireless capsule endoscopic video using color and texture features, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, pp. 603–610.

[36] M. K. Bashar, T. Kitasaka, Y. Suenaga, Y. Mekada, K. Mori, Automatic detection of informative frames from wireless capsule endoscopy images, Medical Image Analysis 14 (2010) 449–470.

[37] A. Mittal, A. K. Moorthy, A. C. Bovik, No-reference image quality as-

sessment in the spatial domain, IEEE Transactions on Image Processing 21 (2012) 4695–4708.

[38] J. L. Pech-Pacheco, G. Cristóbal, J. Chamorro-Martinez, J. Fernández-Valdivia, Diatom autofocusing in brightfield microscopy: a comparative study, in: International Conference on Pattern Recognition, volume 3, IEEE, pp. 314–317.

[39] J. M. Mateos-Pérez, R. Redondo, R. Nava, J. C. Valdiviezo, G. Cristóbal, B. Escalante-Ramírez, M. J. Ruiz-Serrano, J. Pascau, M. Desco, Comparative evaluation of autofocus algorithms for a real-time system for automatic detection of mycobacterium tuberculosis, Cytometry 81 (2012) 213–221.

[40] A. Rényi, On measures of entropy and information, in: Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, volume 1, pp. 547–561.

[41] L. Liu, B. Liu, H. Huang, A. C. Bovik, No-reference image quality assessment based on spatial and spectral entropies, Signal Processing: Image Communication 29 (2014) 856–863.

[42] E. Rublee, V. Rabaud, K. Konolige, G. Bradski, ORB: An efficient alternative to SIFT or SURF, in: IEEE International Conference on Computer Vision, IEEE, pp. 2564–2571.

[43] A. M. Mendrik, E.-J. Vonken, A. Rutten, M. A. Viergever, B. van Ginneken, Noise reduction in computed tomography scans using 3-D anisotropic hybrid diffusion with continuous switch, IEEE Transactions on Medical Imaging 28 (2009) 1585–1594.

[44] C. J. Burges, A tutorial on support vector machines for pattern recognition, Data Mining and Knowledge Discovery 2 (1998) 121–167.

[45] S. Wold, K. Esbensen, P. Geladi, Principal component analysis, Chemometrics and Intelligent Laboratory Systems 2 (1987) 37–52.

[46] L. Breiman, Random forests, Machine learning 45 (2001) 5–32.

[47] S. Van der Walt, J. L. Schönberger, J. Nunez-Iglesias, F. Boulogne, J. D. Warner, N. Yager, E. Gouillart, T. Yu, scikit-image: image processing in python, PeerJ 2 (2014) e453.

[48] D. C. Duro, S. E. Franklin, M. G. Dubé, A comparison of pixel-based and object-based image analysis with selected machine learning algorithms for the classification of agricultural landscapes using SPOT-5 HRG imagery, Remote Sensing of Environment 118 (2012) 259–272.

[49] A. Bosch, A. Zisserman, X. Munoz, Image classification using random forests and ferns, in: International Conference on Computer Vision, IEEE, pp. 1–8.

[50] G. Csurka, C. Dance, L. Fan, J. Willamowski, C. Bray, Visual catego-

rization with bags of keypoints, in: Workshop on Statistical Learning in Computer Vision, volume 1, Prague, pp. 1–2.

[51] L. Kang, P. Ye, Y. Li, D. Doermann, Convolutional neural networks for no-reference image quality assessment, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2014, pp. 1733–1740.

[52] J. Li, L. Zou, J. Yan, D. Deng, T. Qu, G. Xie, No-reference image quality assessment using Prewitt magnitude based on convolutional neural networks, Signal, Image and Video Processing 10 (2016) 609–616.

[53] J. Kolodner, Case-based reasoning, Morgan Kaufmann, 2014.

[54] M. Liedlgruber, A. Uhl, Computer-aided decision support systems for endoscopy in the gastrointestinal tract: a review, IEEE Reviews in Biomedical Engineering 4 (2011) 73–88.

[55] S. Moccia, S. J. Wirkert, H. Kenngott, A. S. Vemuri, M. Apitz, B. Mayer, E. De Momi, L. Mattos, L. Maier-Hein, Uncertainty-aware organ classification for surgical data science applications in laparoscopy, arXiv preprint arXiv:1706.07002 (2017).

[56] M. S. Nosrati, J.-M. Peyrat, J. Abinahed, O. Al-Alao, A. Al-Ansari, R. Abugharbieh, G. Hamarneh, Efficient multi-organ segmentation in 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.