Uncertainty-Aware Organ Classification for Surgical Data Science Applications in Laparoscopy

Sara Moccia, Sebastian J. Wirkert, Hannes Kenngott, Anant S. Vemuri, Martin Apitz, Benjamin Mayer, Elena De Momi, *Senior Member, IEEE*, Leonardo S. Mattos, *Member, IEEE*, and Lena Maier-Hein

***Abstract*—Objective: Surgical data science is evolving into a research field that aims to observe everything occurring within and around the treatment process to provide situation-aware data-driven assistance. In the context of endoscopic video analysis, the accurate classification of organs in the field of view of the camera proffers a technical challenge. Herein, we propose a new approach to anatomical structure classification and image tagging that features an intrinsic measure of confidence to estimate its own performance with high reliability and which can be applied to both RGB and multispectral imaging (MI) data. Methods: Organ recognition is performed using a superpixel classification strategy based on textural and reflectance information. Classification confidence is estimated by analyzing the dispersion of class probabilities. Assessment of the proposed technology is performed through a comprehensive *in vivo* study with seven pigs. Results: When applied to image tagging, mean accuracy in our experiments increased from 65% (RGB) and 80% (MI) to 90% (RGB) and 96% (MI) with the confidence measure. Conclusion: Results showed that the confidence measure had a significant influence on the classification accuracy, and MI data are better suited for anatomical structure labeling than RGB data. Significance: This work significantly enhances the state of art in automatic labeling of endoscopic videos by introducing the use of the confidence metric, and by being the first study to use MI data for *in vivo* laparoscopic tissue classification. The data of our experiments will be released as the first *in vivo* MI dataset upon publication of this paper.**

***Index Terms*—Surgical data science, laparoscopy, multispectral imaging, image tagging, confidence estimation.**

I. Introduction

Surgical Data Science (SDS) has recently emerged as a new scientific field which aims to improve the quality of interventional healthcare [1]. SDS involves the observation of all elements occurring within and around the treatment process in order to provide the right assistance to the right person at the right time.

S. Moccia is with the Department of Electronics, Information and Bioengineering (DEIB), Politecnico di Milano, Milan, Italy, with the Department of Advanced Robotics (ADVR), Istituto Italiano di Tecnologia, Genoa, Italy, and with the Department of Computer Assisted Medical Interventions (CAMI), German Cancer Research Center (DKFZ), Heidelberg, Germany. e-mail: sara.moccia@polimi.it

S. J. Wirkert, A. S. Vemuri, and L. Maier-Hein are with the Department of Computer Assisted Medical Interventions (CAMI), German Cancer Research Center (DKFZ), Heidelberg, Germany.

H. Kenngott, M. Apitz, and B. Mayer are with the Department for General, Visceral, and Transplantation Surgery, Heidelberg University Hospital, Heidelberg, Germany.

E. De Momi is with the Department of Electronics, Information and Bioengineering (DEIB), Politecnico di Milano, Milan, Italy.

L. S. Mattos is with the Department of Advanced Robotics (ADVR), Istituto Italiano di Tecnologia, Genoa, Italy.

In laparoscopy, some of the major opportunities that SDS offers to improve surgical outcomes are surgical decision support [2] and context awareness [3]. Here, technical challenges include the detection and localization of anatomical structures and surgical instrumentation, intra-operative registration, and workflow modeling and recognition. To date, however, clinical translation of the developed methodology continues to be hampered by the poor robustness of the existing methods. In fact, a grand international initiative on SDS [1] concluded that the robustness and reliability of SDS methods are of crucial importance. With the same perspective, several researches in the case-base reasoning community (e.g. [4], [5], [6]) have pointed out the benefits of estimating method confidence level in assigning a result. The aim of this paper is to address this issue in the specific context of organ classification and image tagging in endoscopic video images.

Guided by the hypotheses that (**H1**) automatic confidence estimation can significantly increase the accuracy and robustness of automatic image labeling methods, and that (**H2**) multispectral imaging (MI) data are more suitable for *in vivo* anatomical structure labeling than RGB data, the contributions of this paper are summarized as follows:

1) Uncertainty-aware organ classification (Sec. II-A): Development of a new method for superpixel (Spx)-based anatomical structure classification, which features an intrinsic confidence measure for self-performance estimation and which can be generalized to MI data;
2) Automatic image tagging (Sec. II-B): Development of an approach to automatic image tagging, which relies on the classification method and corresponding confidence estimation to label endoscopic RGB/multispectral images with the organs present in that image;
3) *In vivo* validation (Sec. III): A comprehensive *in vivo* study is conducted using seven pigs to experimentally investigate hypotheses **H1** and **H2**.

It is worth noting that, when we mention image tagging, we refer to the action of identifying organs present in an image. Instead, when mentioning organ classification, we refer to the classification of the organ present in an Spx.

To the best of our knowledge, we are the first to use MI data for *in vivo* abdominal tissue classification. Furthermore, this is the first study to address the topic of classification uncertainty estimation. We will make our validation dataset fully available online.

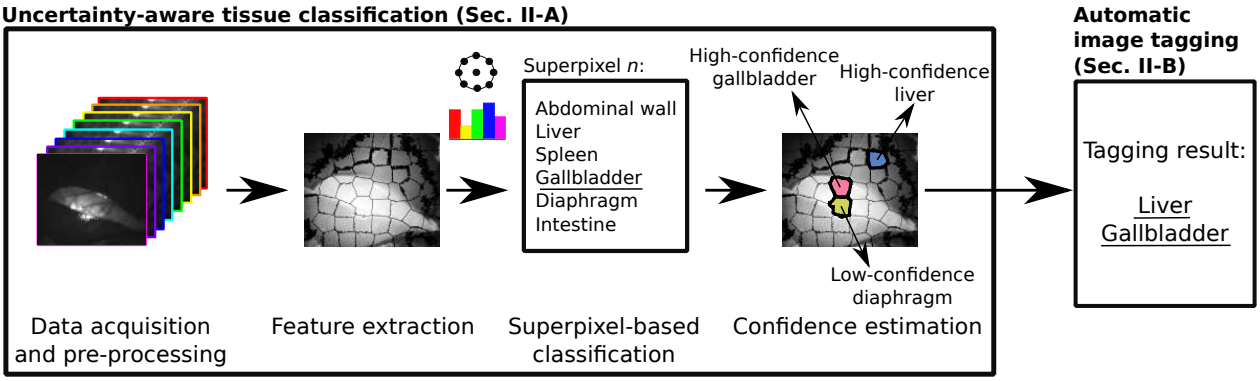


Fig. 1: Workflow of proposed approaches for uncertainty-aware organ classification and automatic image tagging.

A. *Related work*

First attempts at image-guided classification of tissues in RGB endoscopic images primarily used parameter-sensitive morphological operations and intensity-based thresholding techniques, which are not compatible with the high levels of inter-patient multi-organ variability (e.g. [7], [8]). The method for multiple-organ segmentation in laparoscopy reported in [9] relied on non-rigid registration and deformation of pre-operative tissue models on laparoscopic images using color cues. This deformation was achieved using statistical deformable models, which may not always represent the patient-specific tissue deformation, thus resulting in a lack of robustness in terms of inter-patient variability. Recently, machine learning based classification algorithms for tissue classification have been proposed to attenuate this issue. The method described in [10] exploited a machine learning approach to segment the uterus. Gabor filtering and intensity-based features were exploited to segment the uterus from background tissues with support vector machines (SVM) and morphology operators. However, this approach is limited to single organ segmentation and the performance is influenced by the position of the uterus. Similarly, the method presented in [11] was specifically designed for segmentation of fallopian tubes, as it exploits tube-specific geometrical features, such as orientation and width, and cannot be transferred to other anatomical targets.

In parallel to the development of new computer-assisted strategies to tissue classification, the biomedical imaging field is also evolving thanks to new technologies such as MI [12]. MI is an optical technique that enables us to capture both spatial and spectral information on structures. MI provides images that generally have dozens of channels, each corresponding to the reflection of light within a certain wavelength band. Multispectral bands are usually optimized to encode the informative content which is relevant for a specific application. Thus, MI can potentially reveal tissue-specific optical characteristics better than standard RGB imaging systems [12].

One of the first *in vivo* applications of MI was proposed by Afromowitz et al. [13], who developed a MI system to evaluate the depth of burns on the skin, showing that MI provides more accurate results than standard RGB imaging for such application. For abdominal tissue classification, Akbari et al. [14] and Triana et al. [15] exploited pixel-based reflectance features in open surgery and *ex vivo* tissue classification. The work that is most similar to the present study was recently presented by Zhang et al. [16]. It pointed out the advantages of combining both reflectance and textural features. However, the validation study for this focused on patch-based classification and was limited to *ex vivo* experiments in a controlled environment, including only 9 discrete endoscope poses to view the tissues, with only single organs in the image and without tissue motion and deformation. Furthermore, the challenges of confidence estimation were not addressed.

As for automatic laparoscopic image tagging, there is no previous work in the literature that has specifically addressed this challenging topic. However, it has been pointed out that there is a pressing need to develop methods for tagging images with semantic descriptors, e.g. for decision support or context awareness [17], [18]. For example, context-aware augmented reality (AR) in surgery is becoming a topic of interest. By knowing the surgical phase, it is possible to adapt the AR to the surgeon's needs. Contributions in the field include [19], [3]. The AR systems in [19], [3] provide context awareness by identifying surgical phases based on (i) surgical activity, (ii) instruments and (iii) anatomical structures in the image. This is something that is commonly assumed as standard [20]. However, a strategy for retrieving the anatomical structures present in the image was not proposed.

A possible reason for such a lack in the literature can be seen in the challenging nature of tagging images recorded during *in vivo* laparoscopy. Tissues may look very different across images and may be only partially visible. The high level of endoscopic image noise, the wide range of illumination and the variation of the endoscope pose with respect to the recorded tissues further increase the complexity of the problem. As a result, standard RGB systems may be not powerful enough to achieve the task, even when exploiting advanced machine learning approaches to process the images. With **H1** and **H2**, we aim at investigating if the use of MI and the introduction of a measure of classification confidence may face such complexity.

TABLE I: Table of symbols used in Sec. II.

Symbol	Description
N_c	Number of image channels
λ_i	Camera light-filter central wavelength for channel i
$I(\lambda_i)$	Row image for channel i
$Sr(\lambda_i)$	Spectral reflectance image for channel i
$D(\lambda_i)$	Reference dark image for channel i
$W(\lambda_i)$	Reference white image for channel i
N	Number of superpixels in the image
Spx_n	n^{th} superpixel $n \in [0, N)$
$LBP^{R,P}_{riu2}$	Uniform rotation-invariant local binary pattern
R	Radius used to compute $LBP^{R,P}_{riu2}$
P	Number of points used to compute $LBP^{R,P}_{riu2}$
$\{\mathbf{p}_p\}_{p\in(0,P-1)}$	Points used to compute $LBP^{R,P}_{riu2}$
$g_{\mathbf{c}}$	Intensity value of pixel $\mathbf{c}$
H_{LBP}	Histogram of $LBP^{R,P}_{riu2}$
AS_{Spx_n}	Average spectrum for Spx_n
M	Number of pixels in Spx_n
$l_{H_{LBP}}$	Length of H_{LBP} for Spx_n and channel i
l_{AS}	Length of AS for Spx_n and channel i
f	Support vector machine decision function
$\mathbf{x_k}$	k^{th} input feature vector
y_k	k^{th} output label
γ, C	Support vector machine hyperparameters
N_t	Number of training samples
J	Total number of considered abdominal tissues
$Pr(Spx = j)$	Probability for the n^{th} Spx to belond to the j^{th} organ
$E(Spx_n)$	Shannon entropy computed for Spx_n
$PPCI(Spx_n)$	Posterior probability certainty index computed for Spx_n
$GC(Spx_n)$	Gini coefficient computed for Spx_n
L	Lorentz curve

II. Methods

Figure 1 shows an overview of the workflow of the proposed methods for uncertainty-aware organ classification (Sec. II-A) and automatic image tagging (Sec. II-B). Table I lists the symbols used in Sec. II.

A. Uncertainty-aware tissue classification

The steps comprising the proposed approach to organ classification are presented in the following subsections.

1) Pre-processing: To remove the influence of the dark current and to obtain the spectral reflectance image $Sr(\lambda_i)$ for each MI channel $i \in [1, N_C]$), where N_C is the number of MI bands, the raw image $I(\lambda_i)$ was pre-processed by subtracting the reference dark image $D(\lambda_i)$ of the corresponding channel from the multispectral image. λ_i refers to the band central wavelength of the i^{th} channel. This result was then divided by the difference between the reference white image $W(\lambda_i)$ of the corresponding channel and $D(\lambda_i)$, as suggested in [21]:

$$Sr(\lambda_i) = \frac{I(\lambda_i) - D(\lambda_i)}{W(\lambda_i) - D(\lambda_i)} \tag{1}$$

Note that $W(\lambda_i)$ and $D(\lambda_i)$ had to be acquired only once for a given camera setup and wavelength. These images were obtained by placing a white reference board in the field of view and by closing the camera shutter, respectively. Each reflectance image was additionally processed with anisotropic diffusion filtering to remove noise while preserving the sharp edges [22]. The specular reflections were segmented by converting the RGB image into hue, saturation, value (HSV) color space and thresholding the V value. They were then masked from all channels [23].

2) Feature extraction: In the method proposed in this study, we extracted features from Spx. Spx were selected because, compared to regular patches, they are built to adhere to image boundaries better [24]. This characteristic is particularly useful considering the classification of multiple organs within one single image. To obtain the Spx segmentation, we applied linear spectral clustering (LSC) [24] to the RGB image and then used the obtained Spx segmentation for all multispectral channels.

Inspired by the recently published *ex vivo* study by Zhang et al. [16], we extracted both textural and spectral reflectance features from each multispectral channel. Indeed, as stated in Sec. I, the authors demonstrated that incorporating textural information improved the classification performance with respect to single pixel-based features in their controlled experimental setup. As laparoscopic images are captured from various viewpoints under various illumination conditions, the textural features should be robust to the pose of the endoscope as well as to the lighting conditions. Furthermore, their computational cost should be negligible to enable real-time computation with a view to future clinical applications.

The histogram (H_{LBP}) of the uniform rotation–invariant local binary pattern ($LBP^{R,P}_{riu2}$), which fully meets these requirements, was here used to describe the tissue texture of an Spx.

The $LBP^{R,P}_{riu2}$ formulation requires to define, for a pixel $\mathbf{c} = (c_x, c_y)$, a spatial circular neighborhood of radius R with P equally-spaced neighbor points ($\{\mathbf{p}_p\}_{p\in(0,P-1)}$):

$$LBP^{R,P}_{riu2}(\mathbf{c}) = \begin{cases} \sum_{p=0}^{P-1} s(g_{\mathbf{p}_p} - g_{\mathbf{c}}), & \text{if } U(LBP^{R,P}) \leq 2 \\ P+1, & \text{otherwise} \end{cases} \tag{2}$$

where $g_{\mathbf{c}}$ and $g_{\mathbf{p}_p}$ denote the gray values of the pixel $\mathbf{c}$ and of its p^{th} neighbor $\mathbf{p}_p$, respectively. $s(g_{\mathbf{p}_p} - g_{\mathbf{c}})$ is defined as:

$$s(g_{\mathbf{p}_p} - g_{\mathbf{c}}) = \begin{cases} 1, & g_{\mathbf{p}_p} \geq g_{\mathbf{c}} \\ 0, & g_{\mathbf{p}_p} < g_{\mathbf{c}} \end{cases} \tag{3}$$

and $U(LBP^{R,P})$ is defined as:

$$\begin{aligned} U(LBP^{R,P}) = &|s(g_{\mathbf{p}_{P-1}} - g_{\mathbf{c}}) - s(g_{\mathbf{p}_0} - g_{\mathbf{c}})| + \\ &\sum_{p=1}^{P-1} |s(g_{\mathbf{p}_p} - g_{\mathbf{c}}) - s(g_{\mathbf{p}_{p-1}} - g_{\mathbf{c}})| \end{aligned} \tag{4}$$

The H_{LBP}, which counts the occurrences of $LBP^{R,P}_{riu2}$, was normalized to the unit length to account for the different pixel numbers in an Spx.

Spectral reflectance information was encoded in the average spectrum (AS), which is the average spectral reflectance value in an Spx. The AS for the i^{th} channel and the n^{th} Spx (Spx_n), with $n \in (1, N)$ and N the total number of Spx, is defined as:

$$AS_{Spx_n}(\lambda_i) = \frac{1}{M} \sum_{\mathbf{p}\in Spx_n} Sr_p(\lambda_i) \tag{5}$$

where M is the number of pixels in Spx_n and $Sr_p(\lambda_i)$

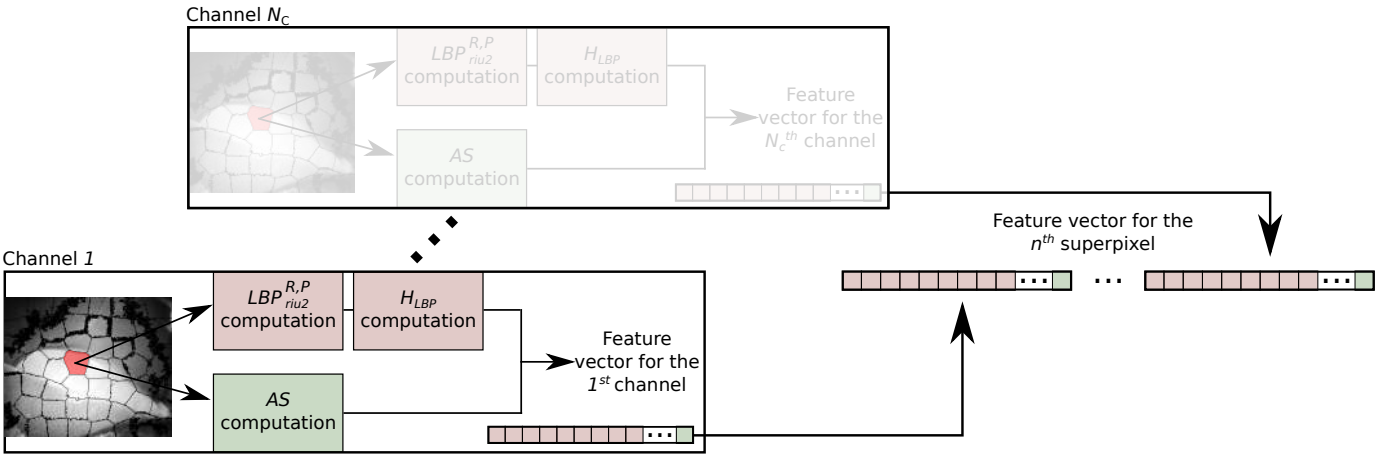


Fig. 2: A feature vector is extracted from each $n \in N$ superpixel (Spx_n), where N is the number of superpixels in the image. The feature vector for Spx_n is obtained by concatenating the histogram (H_{LBP}) of uniform rotation–invariant local binary pattern ($LBP^{R,P}_{riu2}$) and the average spectrum (AS), for each $i \in N_C$ image channel, where N_C is the number of channels in the image.

is the reflectance value of the p^{th} pixel of Spx_n in the i^{th} channel.

The L2-norm was applied to the AS in order to accommodate lighting differences. AS was exploited instead of the simple spectral reflectance at one pixel to improve the feature robustness against noise, although this is detrimental to spatial resolution.

The steps for obtaining the feature vector are shown in Fig. 2.

3) Superpixel-based classification: To classify the Spx-based features, we used SVM with the radial basis function. For a binary classification problem, given a training set of N_t data $\{y_k, \mathbf{x_k}\}_{k=1}^{N_t}$, 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:

$$f(\mathbf{x}) = sign\Big[\sum_{k=1}^{N_t} a_k^* y_k \Psi(\mathbf{x}, \mathbf{x_k}) + b\Big] \tag{6}$$

where:

$$\Psi(\mathbf{x}, \mathbf{x_k}) = exp\{-\gamma||\mathbf{x} - \mathbf{x_k}||_2^2/\sigma^2\}, \qquad \gamma > 0 \tag{7}$$

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

$$a_k^* = \max\Big\{-\frac{1}{2}\sum_{k,l=1}^{N_t} y_k y_l \Psi(\mathbf{x_k}, \mathbf{x_l}) a_k a_l + \sum_{k=1}^{N_t} a_k\Big\} \tag{8}$$

with:

$$\sum_{k=1}^{N_t} a_k y_k = 0, \qquad 0 \le a_k \le C, \qquad k = 1, ..., N_t \tag{9}$$

In this paper, γ and C were computed with grid search, as explained in Sec. III.

Since our classification task is a multiclass classification problem, we implemented SVM with the *one-against-one* scheme. Specifically, six organ classes were involved in the SVM training process, as described in Sec. III. Prior to classification, we standardized the feature matrix within each feature dimension.

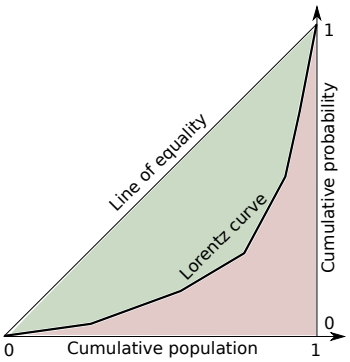


Fig. 3: The Gini coefficient (GC) is computed as twice the area (green area) between the line of equality and the Lorentz curve. The Lorentz curve represents the cumulative classification probability among the outcome classification states rank-ordered according to the decreasing values of their individual probabilities. A uniform discrete probability distribution has $GC = 0$, as the Lorentz curve overlays the line of equality, while for a state with probability 100% and the others at 0%, $GC = 1$.

As a prerequisite for our confidence estimation, we retrieved the probability $Pr(Spx_n = j)$ for the n^{th} Spx, to belong to the j^{th} organ ($j \in [1, J]$), J is the number of considered organs. In particular, $Pr(Spx_n = j)$ was obtained, according to the pairwise comparison method proposed in [25] (which is an extension of [26] for the binary classification case), by solving:

$$Pr(Spx_n = j) = \sum_{i=1, i\neq j}^{J} \frac{Pr(Spx_n = j) + Pr(Spx_n = i)}{J - 1} r_{ji}, \forall j \tag{10}$$

subject to:

$$\sum_{j=1}^{J} Pr(Spx_n = j) = 1, \quad Pr(Spx_n = j) \geq 0, \quad \forall j \quad (11)$$

where r_{ij} is the estimates of $Pr(Spx_n = j | Spx_n \in \{i, j\})$ with $r_{j,i} + r_{i,j} = 1, \forall j \neq i$. The estimator $r_{j,i}$ was obtained according to [26], mapping the SVM output to probabilities by training the parameters of a sigmoid function.

4) Confidence estimation: To estimate the SVM classification performance, we evaluated two intrinsic measures of confidence: (i) a measure based on the normalized Shannon entropy (E), called posterior probability certainty index ($PPCI$), and (ii) the Gini coefficient (GC) [27].

For the n^{th} Spx, $PPCI(Spx_n)$ is defined as:

$$PPCI(Spx_n) = 1 - E(Spx_n) \quad (12)$$

where E is:

$$E(Spx_n) = -\frac{\sum_{j=1}^{J} Pr(Spx_n = j) log(Pr(Spx_n = j))}{log(J)} \quad (13)$$

and:

$$log(Pr(Spx_n = j)) = \begin{cases} log(Pr(Spx_n = j)), & \text{if } Pr(Spx_n = j) > 0 \\ 0, & \text{if } Pr(Spx_n = j) = 0 \end{cases} \quad (14)$$

For the n^{th} Spx, $GC(Spx_n)$ is defined as:

$$GC(Spx_n) = 1 - 2 \int_0^1 L(x)\, \mathrm{d}x. \quad (15)$$

where L is the Lorentz curve, which is the cumulative probability among the J outcome states rank-ordered according to the decreasing values of their individual probabilities ($Pr(Spx_n = 1), ..., Pr(Spx_n = J)$). As can be seen from Fig. 3, in case of uniform discrete probability distribution (complete uncertainty), L corresponds to the line of equality. Thus, the integral in Eq. 15 (red area in Fig. 3) has values 0.5 and $GC = 0$. On the contrary, for the case of a single state at 100% with the others at 0% (complete certainty), the integral value is 0 and $GC = 1$. The GC computation can be also seen as twice the area (green area in Fig. 3) between the line of equality and the Lorentz curve.

Although both metrics are suitable to evaluate the dispersion of the classification probability, GC is faster to compute, as it does not require the logarithm computation. Moreover, GC is more sensitive than $PPCI$ at higher values [27].

B. Automatic image tagging

Automatic image tagging uses the SVM Spx-based classification and the corresponding confidence estimation. Specifically, test images were tagged considering Spx labels with high confidence values only. The value of $GC(Spx_n)$ was thresholded to obtain binary confidence information. An Spx was considered to have an acceptable confidence level if $GC(Spx_n) > \tau$, for the threshold τ. The same procedure was performed using $PPCI$ instead of GC.

III. In vivo validation

Seven pigs were used to examine the **H1** and **H2** introduced in Sec. I. Raw multispectral images (I) were acquired using a custom-built MI laparoscope. In this study, the multispectral laparoscope was comprised of a Richard Wolf (Knittlingen, Germany) laparoscope and a 5–MP Pixelteq Spectrocam (Largo, FL, USA) multispectral camera. The λ_i for each i^{th} band index and the corresponding full widths at half maximum (FWHM) are reported in Table II. The filters were chosen according to the band selection strategy for endoscopic spectral imaging presented in [28]. The method makes use of the Sheffield index [29], which is an information theory based band selection method originally proposed by the remote sensing community. The 700, 560 and 470 nm channels were chosen to simulate RGB images as the camera did not provide RGB images directly. The image size was $1228 \times 1029 \times 8$ for MI and $1228 \times 1029 \times 3$ for RGB.

The physical size of the multispectral camera was 136 x 124 x 105 mm, with a weight of 908 g. The acquisition time of one multispectral image stack took 400 ms.

From the seven pigs, three pigs were used for training (29 images) and four for testing (28 images). The number of images used to test the SVM performance on RGB and MI data was the same, as RGB data were directly obtained from MI data by selecting 3 of the 8 MI channels. The total number of Spx in the training and testing dataset, for both MI and RGB data, was 1382 and 1559, respectively.

We considered six porcine organ tissues typically encountered during hepatic laparoscopic surgery: the liver, gallbladder, spleen, diaphragm, intestine, and abdominal wall. These tissues were recorded during *in vivo* laparoscopy. Challenges associated with the *in vivo* dataset include:

- Wide range of illumination
- Variation of the endoscope pose
- Presence of specular reflections
- Presence of multiple organs in one image
- Organ movement

Visual samples of the dataset challenges are shown in Fig. 4.

The multispectral images were pre-processed as described in Sec. II-A. The Spx segmentation with LSC was achieved using an average Spx size of 150^2 pixels and an Spx compactness factor of 0.1. Accordingly, 55 Spx on average were obtained for each image. The $LBP^{R,P}_{riu2}$ were computed considering the following (R, P) combinations: (1, 8), (2, 16), and (3, 24). The feature vector for an Spx was obtained by concatenating the H_{LBP} with the AS value for all 8 multispectral channels (for MI) and for $\lambda_i = 700$, 560 and 470 nm (for RGB). The feature vector size for an Spx was:

$$(l_{H_{LBP}} + l_{AS}) \times N_C \quad (16)$$

where $l_{H_{LBP}}$ is the length of H_{LBP}, equal to 54, l_{AS} is the length of AS, equal to 1, and N_C is the number of channels, 3 for RGB and 8 for multispectral data.

The SVM kernel parameters ($C = 10^4$ and $\gamma = 10^{-5}$) were retrieved during the training phase via grid-search and 10-fold cross-validation on the training set. The grid-search spaces for γ and C were set to [10^{-8}, 10^1] and [10^1, 10^{10}],

TABLE II: Camera light-filter central wavelengths and full width at half maximum (FWHM) for each $i(= 1–8)$ band in multispectral imaging (MI) and RGB.

i	1	2	3	4	5	6	7	8
MI	470 nm	480 nm	511 nm	560 nm	580 nm	600 nm	660 nm	700 nm
RGB	470 nm	-	-	560 nm	-	-	-	700 nm
FWHM	20 nm	25 nm	20 nm	20 nm	20 nm	20 nm	20 nm	20 nm

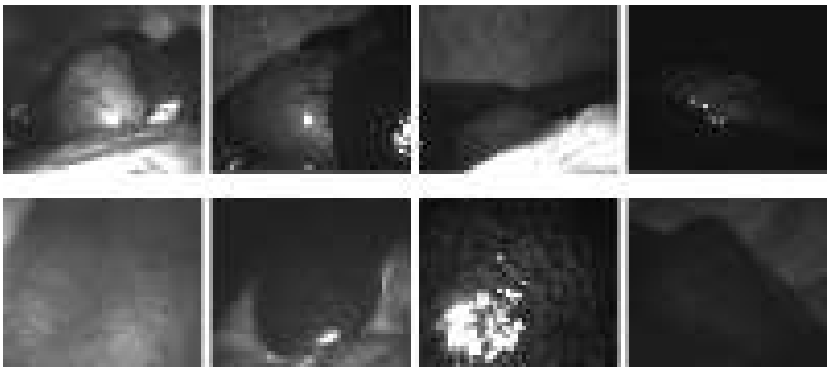


Fig. 4: Challenges of the evaluation dataset. Four samples of images showing the gallbladder (first row) and spleen (second row) are reported. Images were recorded with varying endoscope pose and illumination level. Specular reflections are present in the images due to the smooth and wet organ surfaces. Multiple organs can be present in a single image. All images refer to the same multispectral channel.

TABLE III: Median superpixel-based accuracy rate (Acc_{Spx}) and inter-quartile range (in brackets) for RGB and multispectral imaging (MI) using different features for the *Base* case (i.e., without confidence inclusion). H_{LBP}: Histogram of local binary patterns; AS: Average spectrum.

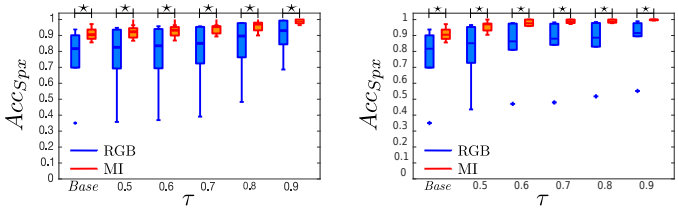
	H_{LBP}		AS		$H_{LBP} + AS$	
	RGB	MI	RGB	MI	RGB	MI
Acc_{Spx}	63% (17%)	77% (13%)	76% (39%)	88% (18%)	81% (20%)	90% (6%)

respectively, with 10 values spaced evenly on the log_{10} scale in both cases. The determined values for the hyperparameters were subsequently used in the testing phase.

The feature extraction was implemented using OpenCV [1]. The classification was implemented using scikit-learn [30] [2].

1) Investigation of H1: To investigate whether the inclusion of a confidence measure increases Spx-based organ classification accuracy (Acc_{Spx}), we evaluated the Acc_{Spx} dependence on $\tau \in [0.5 : 0.1 : 1)$ applied to both GC and $PPCI$. Acc_{Spx} is defined as the ratio of correctly classified confident Spx to all confident samples in the testing set. We evaluated whether differences existed between Acc_{Spx} obtained applying GC and $PPCI$ on the SVM output probabilities using the Wilcoxon signed-rank test for paired samples (significance level = 0.05). We also investigated the SVM performance with the inclusion of confidence when leaving one organ out of the training set. Specifically, we trained six SVMs, leaving each time one organ out. We computed, for each of the six cases, the percentage ($^{\%}LC_{Spx}$) of low-confidence Spx (considering $\tau = 0.9$). We did this both for the organ that was excluded (Ex) from the training set and for the included organs (In). For image tagging, we computed the tagging accuracy (Acc_{Tag}) for different τ, where Acc_{Tag} is the ratio of correctly classified organs in the image to all organs in the testing image.

2) Investigation of H2: To investigate whether MI data are more suitable for anatomic structure classification than conventional RGB video data, we performed the same analysis for RGB and compared the results with those from the MI. To complete our evaluation, we also evaluated the performance



(a) GC-based confidence estimation

(b) $PPCI$-based confidence estimation

Fig. 5: Effect of confidence threshold (τ) on the superpixel-based organ classification accuracy rate (Acc_{Spx}) for RGB and multispectral imaging (MI). *Base* refers to classification without confidence estimation. The stars indicate significant differences. The confidence is computed with: (a) the Gini coefficient (GC), (b) the posterior probability certainty index ($PPCI$).

[1] http://opencv.org/

[2] http://scikit-learn.org/

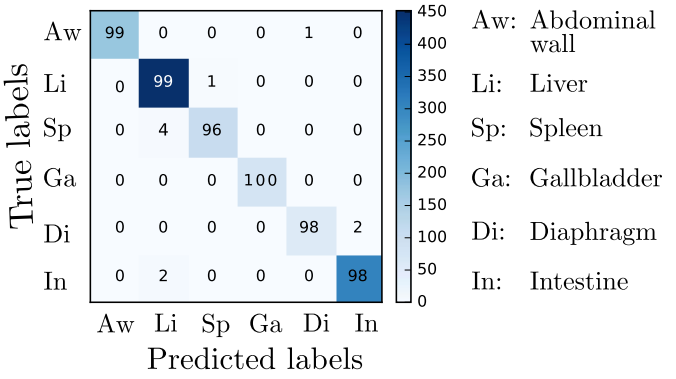


Fig. 6: Confusion matrix for confidence threshold $\tau = 0.9$ on the Gini coefficient and multispectral imaging. The values are in percentages and the colorbar indicates the number of superpixels.

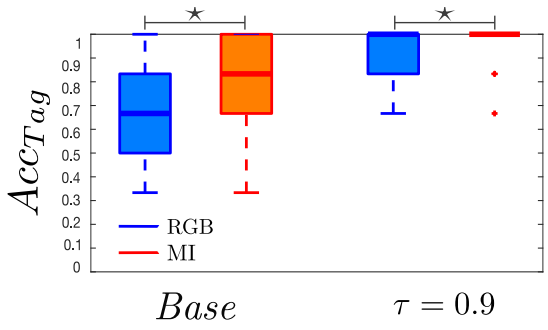


Fig. 7: Image tagging accuracy (Acc_{Tag}) for RGB and multispectral imaging (MI) for *Base* case and following introduction of confidence measure ($\tau = 0.9$ on the Gini coefficient). The stars indicate significant differences.

of H_{LBP} alone and AS alone for $\tau = 0$, which corresponds to the *Base* case, i.e., SVM classification without a confidence computation. Since the analyzed populations were not normal, we used the Wilcoxon signed-rank test for paired samples to assess whether differences existed between the mean ranks of the RGB and MI results (significance level $= 0.05$).

IV. Results

The descriptive statistics of Acc_{Spx} for the analyzed features are reported in Table III. For the *Base* case, the highest Acc_{Spx} (median $= 90\%$, inter-quartile range $= 6\%$) was obtained with $H_{LBP} + AS$ and MI. The other results all differ significantly (p-value < 0.05) from those obtained with $H_{LBP} + AS$ and MI.

When τ applied to GC (Fig. 5a) and $PPCI$ (Fig. 5b) was varied in $[0.5 : 0.1 : 1)$, the median Acc_{Spx} for the MI data increased monotonously to 99% ($\tau = 0.9$), when using both GC and $PPCI$. The same trend was observed for the RGB data, with an overall improvement of the median from 81% to 93% (using GC) and 91% (using $PPCI$). For both the *Base* case and after introduction of the confidence measures, the MI outperformed the RGB (p-value < 0.05). No significant differences were found when comparing the classification performance obtained with GC and $PPCI$. Therefore, as GC computation is more sensitive to high values and faster to compute than $PPCI$, we decided to use GC.

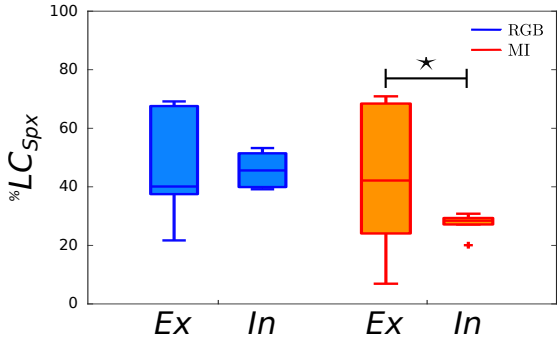


Fig. 8: Effect of previously unseen target structures on the uncertainty estimation. Percentage ($^{\%}LC_{Spx}$) of low-confidence Spx ($\tau = 0.9$) for organs that were seen (In) and not seen (Ex) in the training process.

Figure 6 shows the confusion matrix for MI and $\tau = 0.9$ on GC. Note that, in the case yielding the least accurate result, which corresponds to spleen classification, the accuracy rate still achieved 96%, whereas for RGB the lowest accuracy rate was 69%.

The $^{\%}LC_{Spx}$ boxplots relative to the leave-one-organ out experiment are shown in Fig. 8. The $^{\%}LC_{Spx}$ is significantly higher for organs that were not seen in the training phase (MI: 42% (Ex) vs. 23% (In); RGB: 36% (Ex) vs. 40% (In)).

When applied to endoscopic image tagging, the mean Acc_{Tag} values in our experiments were increased from 65% (RGB) and 80% (MI) to 90% (RGB) and 96% (MI) with the incorporation of the confidence measure (using GC). The descriptive statistics are reported in Fig. 7. In this instance, the MI also outperformed the RGB both in the *Base* case and with the confidence measure (p-value < 0.05). Figure 9 shows the influence of low-confidence Spx exclusion on the image tagging: after low-confidence Spx exclusion, all Spx in the image were classified correctly.

Sample results for the SVM classification and the corresponding confidence map (using GC) are shown in Fig. 10. For low-confidence Spx, the probable cause of uncertainty is also reported. The main sources of uncertainty are specular reflections, camera sensor noise at the image corner, and the partial organ effect, i.e., when two or more organs correspond to one Spx.

V. Discussion

The emerging field of surgical data science [1] aims at observing the entire patient workflow in order to provide the right assistance at the right time. One important prerequisite for context-aware assistance during surgical treatment is to correctly classify the phase within an intervention. While a great amount of effort has been put into automatic instrument detection (e.g. [31], [18], [32]), the problem of automatic organ classification has received extremely little extension. We attribute this to the fact that the task is extremely challenging. In fact, the related problem of organ boundary detection

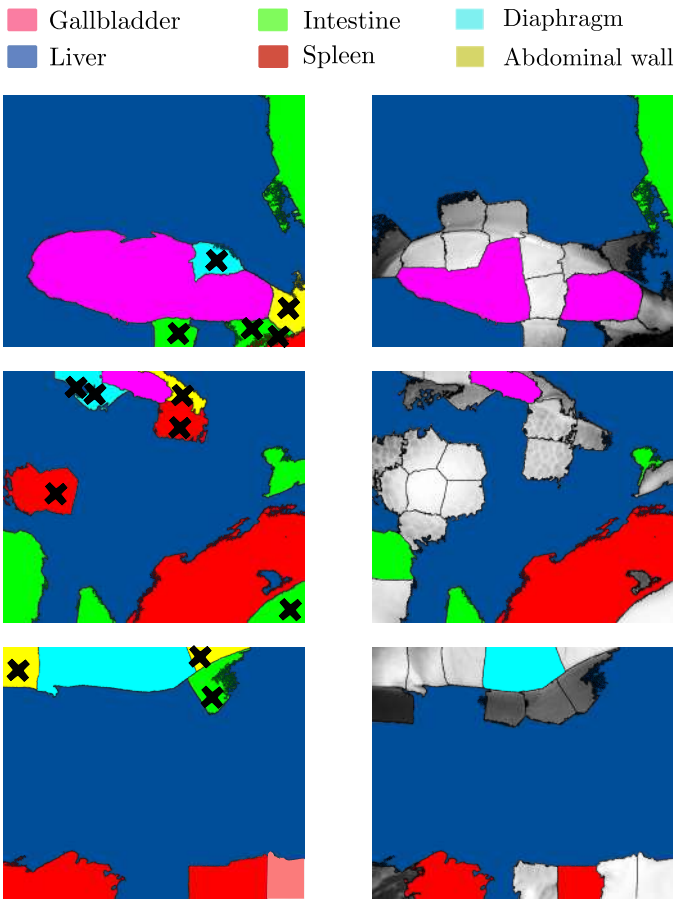


Fig. 9: Image tagging examples for *Base* case (left) and following introduction of confidence measure with $\tau = 0.9$ on the Gini coefficient (right). The low-confidence superpixels (in gray) are excluded from the image tagging. The crosses indicate erroneously classified superpixels.

was regarded so challenging by participants of the MICCAI 2017 endoscopic vision challenge (https://endovis.grand-challenge.org/) that only a single team decided to submit results for the sub-challenge deadline with kidney boundary detection. In this work, we tackled this problem by two previously unexplored approaches:

- Accuracy: We slightly changed the image acquisition process using a multispectral camera as opposed to a standard RGB camera in order to increase the quality of the input data (for the classifier). The effect of this measure was an increase in accuracy of 11% for the task of organ classification and an increase of 23% for the task of automatic image tagging.
- Robustness: We derived superpixel-based measures of confidence to increase the reliability of image tagging. The result was a boost in accuracy of 38% (RGB) and 20% (MI) absolute.

With our validation dataset, we showed that MI significantly outperforms standard RGB imaging in classifying abdominal tissues. Indeed, as the absorption and scattering of light in tissue is highly dependent on (i) the molecules present in the tissues, and (ii) the wavelength of the light, the multispectral image stack was able to encode the tissue-specific optical information, enabling higher accuracy in distinguish different abdominal structures in comparison to standard RGB.

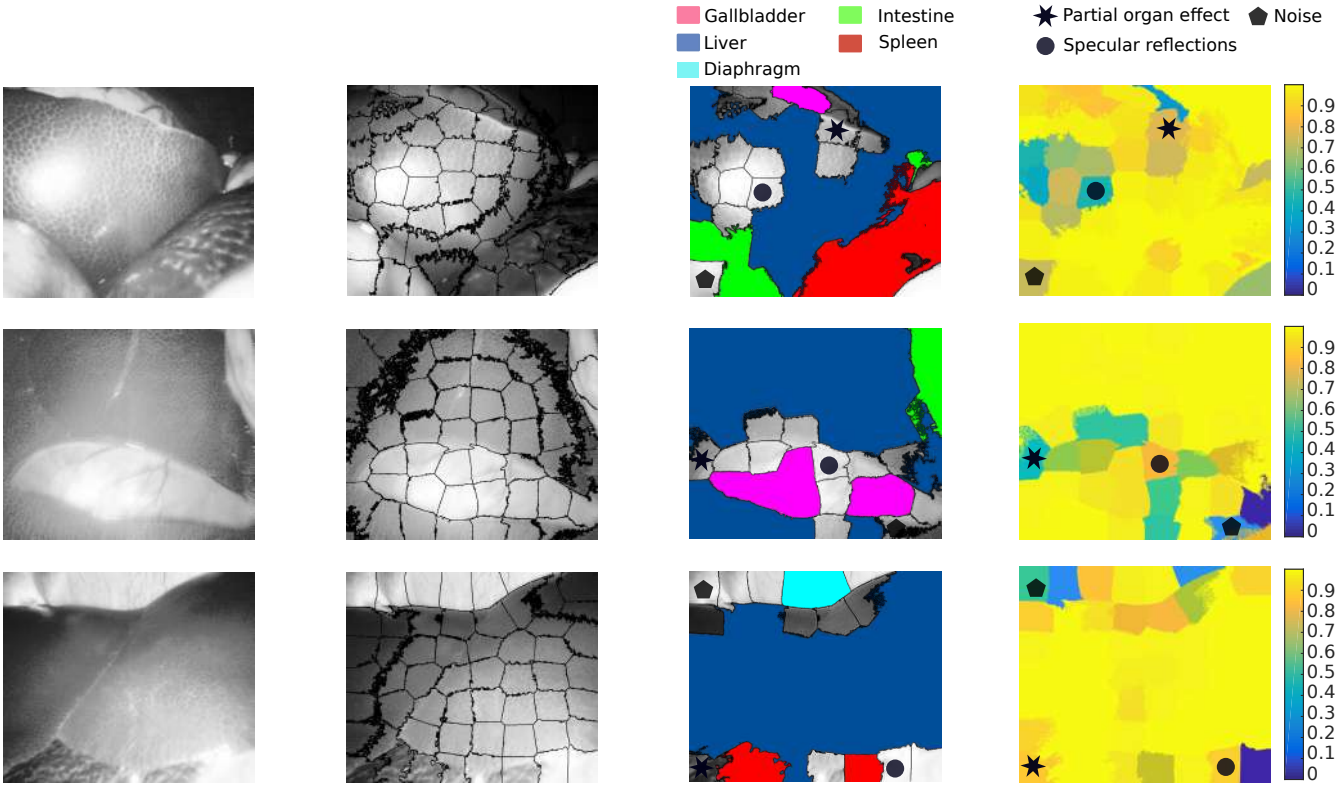
With the introduction of the confidence measure, we showed that the classification accuracy can be improved, for both RGB and MI. This happened when exploiting both GC and $PPCI$. Since no significant differences were found between GC and $PPCI$, we decided to use GC as it is more sensitive at higher values than $PPCI$ and its computation is faster. In fact, a major advantage of our method is its high classification accuracy, which attained 93% (RGB) and 99% (MI) in the regions with high confidence levels, with a significant improvement compared to the *Base* case. Few misclassifications of high-confidence Spx occurred, and where they did then this was mainly with tissues that are also challenging to distinguish between for the human eye, e.g. liver and spleen (Fig. 6). It is worth noting that GC and $PPCI$ were two examples of confidence estimation measures to investigate **H1**. We decided against using simple thresholding on the maximum (Max) value of $Pr(Spx_n = j)$ computed among the J organ classes as GC and $PPCI$ are generally known for being more sensitive at higher values [27]. This assumption was confirmed in additional experiments, where image tagging performed with confident Spx according to GC/$PPCI$ was substantially more robust than tagging based on confident Spx according to Max.

The results obtained with the introduction of the confidence measure are comparable with those obtained by Zhang et al. [16] for *ex vivo* organ classification in a controlled experimental setup. Zhang et al. reported a median classification accuracy of 98% for MI, whereas our classification accuracy for the *Base* case only achieved 90% due to the challenging nature of the *in vivo* dataset. An accuracy level comparable to the one of [16] was, however, restored for our dataset once the low-confidence Spx were excluded.

When excluding one organ from the training set, ${}^{\%}LC_{Spx}$ relative to the excluded organ was significantly higher than the number of low-confidence superpixels obtained for the remaining organs. This indicates that the confidence inclusion helped in handling situations where unknown structures appeared in the field of view of the camera.

These results are in keeping with those found in the literature for case reasoning [4], [6]. Indeed, the importance of the estimation of the level of confidence of the classification with a view to improving system performance has been widely highlighted in several research fields, such as face recognition [33], spam-filtering [34], and cancer recognition [35]. However, the use of confidence metrics had not been exploited in the context of laparoscopic image analysis, up until now.

Although several Spx misclassifications occurred at the *Base* case, which had a negative effect on tagging performance, the low-confidence Spx exclusion significantly increased tagging accuracy. Indeed, regions affected by camera sensor noise, specular reflections, and spectral channel shift due to organ movement were easily discarded based on their confidence value. The same process was implemented when the Spx segmentation failed to separate two organs. Also in this case, MI showed that it performs better than standard



(a) Test image (b) Superpixel segmentation (c) Classification for confident superpixels (d) Confidence map with associated colorbar

Fig. 10: (a) Test image, (b) test image with superpixel segmentation, (b) corresponding classification for superpixels with acceptable confidence level and (c) confidence map obtained with confidence threshold $\tau = 0.9$ on the Gini coefficient. The symbols give examples of the probable causes of uncertainty.

RGB.

While we are the first to address the challenges of *in vivo* image labeling, including the large variability of illumination, variation of the endoscope pose, the presence of specular reflections, organ movement, and the appearance of multiple organs in one image, one disadvantage of our validation setup is that our database was not recorded during real surgery. Hence, some of the challenges typically encountered when managing real surgery images were absent (e.g., blood, smoke, and occlusion). Moreover, as our camera does not provide RGB data directly, we generated a synthetic RGB image by merging three MI channels. It should be noted, however, that our RGB encodes more specific information, as the bands used to obtain these data are considerably narrower than those of standard RGB systems (FWHM = 20 nm). We also recognize that a limitation of the proposed work could be seen in the relatively small number of training images (29). However, analyzing researches on the topic of tissue classification in laparoscopy, such number is comparable with the one of Chhatkuli et al. [10], which exploited 45 uterine images, and Zhang et al. [16], which recorded 9 poses of just 12 scenes (3 pigs $\times$ 4 *ex-vivo* organs). Further, it is worth noting that our training was performed at Spx-level, meaning that the training set sample size was about 55×29, where 55 is the average number of Spx in an image.

Considering that the proposed study was not aimed at evaluating the system performance for clinical translation purpose, we did not analyze the clinical requirements of the proposed method performance. Despite the fact that we recognize the relevance of such analysis, we believe that it should be performed in relation to the specific application. For example, with reference to [19], we plan to analyze and evaluate the requirements of a context-aware AR system supported by the proposed methodology. However, when discussing with our clinical partners, it emerged that the end-to-end accuracy should be close to 100% (i.e. for recognizing the surgical state). However, it has to be further investigated how errors in image tagging affect the error of the final task.

With our MI laparoscope prototype, the image stack acquisition time (400 ms) was faster than most systems commonly presented in literature, like e.g. (e.g. [36] with $\sim$3 s), which makes it more advantageous for clinical applications. Anyway, to fully meet the clinical requirements in terms of system usability, we are currently working on further shrinking the system and speeding it up, as to achieve real-time acquisition. A further solution we would like to investigate is the use of loopy belief propagation [37], [38] as post-processing strategy to include spatial information with respect to how confident classification labels appear in the image. This would be particularly useful for images where the tagging failed due to few confident misclassified Spx surrounded by correctly classified confident Spx. Future work will also deal with the

real-time implementation of the classification algorithm, which was not the aim of this work. Recent advancements in tissue classification research suggest that the use of convolutional neural network (CNN) could be also investigated for comparison [39]. Indeed, uncertainty in deep learning is an active and relatively new field of research, and standard deep learning tools for classification do not capture model uncertainty [40]. Excluding popular dropout strategies (e.g. [41], [42]), among the most recently proposed solutions, variational Bayes by Backpropagation [43], [44] is is drawing attention of the deep learning community.

VI. Conclusions

In this paper, we addressed the challenging topic of robust classification of anatomical structures in *in vivo* laparoscopic images. With the first *in vivo* laparoscopic MI dataset, we confirmed the two hypotheses: (**H1**) the inclusion of a confidence measure increases the Spx-based organ classification accuracy substantially and (**H2**) MI data are more suitable for anatomic structure classification than conventional video data. To this end, we proposed the first approach to anatomic structure labeling. The approach features an intrinsic confidence measure and can be used for high accuracy image tagging, with an accuracy of 90% for RGB and 96% for MI. In conclusion, the method proposed herein could become a valuable tool for surgical data science applications in laparoscopy due to the high level of accuracy it provides in image tagging. Moreover, by making our MI dataset fully available, we believe we will stimulate researches in the field, encouraging and promoting the clinical translation of MI systems.

Compliance with ethical standards

Disclosures

The authors have no conflict of interest to disclose.

Ethical standards

This article does not contain any studies with human participants. All applicable international, national and/or institutional guidelines for the care and use of animals were followed.

References

[1] L. Maier-Hein *et al.*, "Surgical data science for next-generation interventions," *Nature Biomedical Engineering*, vol. 1, no. 9, p. 691, 2017.

[2] K. März *et al.*, "Toward knowledge-based liver surgery: Holistic information processing for surgical decision support," *International Journal of Computer Assisted Radiology and Surgery*, vol. 10, no. 6, pp. 749–759, 2015.

[3] D. Katić *et al.*, "Bridging the gap between formal and experience-based knowledge for context-aware laparoscopy," *International Journal of Computer Assisted Radiology and Surgery*, vol. 11, no. 6, pp. 881–888, 2016.

[4] W. Cheetham and J. Price, "Measures of solution accuracy in case-based reasoning systems," in *European Conference on Case-Based Reasoning*. Springer, 2004, pp. 106–118.

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

[6] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?" *arXiv preprint arXiv:1703.04977*, 2017.

[7] J. Lee *et al.*, "Automatic classification of digestive organs in wireless capsule endoscopy videos," in *ACM Symposium on Applied Computing*. Association for Computing Machinery, 2007, pp. 1041–1045.

[8] P. W. Mewes *et al.*, "Automatic region-of-interest segmentation and pathology detection in magnetically guided capsule endoscopy," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2011*. Springer, 2011, pp. 141–148.

[9] M. S. Nosrati *et al.*, "Efficient multi-organ segmentation in multi-view endoscopic videos using pre-operative priors." in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2014*. Springer, 2014, pp. 324–331.

[10] A. Chhatkuli *et al.*, "Live image parsing in uterine laparoscopy," in *International Symposium on Biomedical Imaging ISBI 2014*. IEEE, 2014, pp. 1263–1266.

[11] K. Prokopetc *et al.*, "Automatic detection of the uterus and fallopian tube junctions in laparoscopic images," in *The 24th Biennial International Conference on Information Processing in Medical Imaging (IPMI)*. Springer, 2015, pp. 552–563.

[12] Q. Li *et al.*, "Review of spectral imaging technology in biomedical engineering: achievements and challenges," *Journal of Biomedical Optics*, vol. 18, no. 10, pp. 100 901–100 901, 2013.

[13] M. A. Afromowitz *et al.*, "Multispectral imaging of burn wounds: a new clinical instrument for evaluating burn depth," *IEEE Transactions on Biomedical Engineering*, vol. 35, no. 10, pp. 842–850, 1988.

[14] H. Akbari and Y. Kosugi, *Hyperspectral imaging: A new modality in surgery*. InTechOpen Open Access Publisher, 2009.

[15] B. Triana *et al.*, "Multispectral tissue analysis and classification towards enabling automated robotic surgery," in *SPIE Medical Imaging*. International Society for Optics and Photonics, 2014, pp. 893 527–893 527.

[16] Y. Zhang *et al.*, "Tissue classification for laparoscopic image understanding based on multispectral texture analysis," in *SPIE Medical Imaging*. International Society for Optics and Photonics, 2016, pp. 978 619–978 619.

[17] Y. Gur *et al.*, "Towards an efficient way of building annotated medical image collections for big data studies," in *Intravascular Imaging and Computer Assisted Stenting, and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*. Springer, 2017, pp. 87–95.

[18] S. Bodenstedt *et al.*, "Superpixel-based structure classification for laparoscopic surgery," in *Medical Imaging 2016*, 2016.

[19] D. Katić *et al.*, "Context-aware augmented reality in laparoscopic surgery," *Computerized Medical Imaging and Graphics*, vol. 37, no. 2, pp. 174–182, 2013.

[20] T. Neumuth *et al.*, "Validation of knowledge acquisition for surgical process models," *Journal of the American Medical Informatics Association*, vol. 16, no. 1, pp. 72–80, 2009.

[21] A. Mansouri *et al.*, "Development of a protocol for CCD calibration: application to a multispectral imaging system," *International Journal of Robotics and Automation*, vol. 20, no. 2, pp. 94–100, 2005.

[22] D.-J. Kroon *et al.*, "Optimized anisotropic rotational invariant diffusion scheme on cone-beam CT," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2010*. Springer, 2010, pp. 221–228.

[23] S. Moccia *et al.*, "Automatic workflow for narrow-band laryngeal video stitching," in *Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the*. IEEE, 2016, pp. 1188–1191.

[24] Z. Li and J. Chen, "Superpixel segmentation using linear spectral clustering," in *Computer Vision and Pattern Recognition (CVPR), Conference on*. IEEE, 2015, pp. 1356–1363.

[25] T.-F. Wu *et al.*, "Probability estimates for multi-class classification by pairwise coupling," *Journal of Machine Learning Research*, vol. 5, no. Aug, pp. 975–1005, 2004.

[26] J. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Advances in Large Margin Classifiers*, vol. 10, no. 3, pp. 61–74, 1999.

[27] B. G. Marcot, "Metrics for evaluating performance and uncertainty of Bayesian network models," *Ecological Modelling*, vol. 230, pp. 50–62, 2012.

[28] S. J. Wirkert *et al.*, "Endoscopic Sheffield index for unsupervised in vivo spectral band selection," in *International Workshop on Computer-Assisted and Robotic Endoscopy*. Springer, 2014, pp. 110–120.

[29] C. Sheffield, "Selecting band combinations from multispectral data," *Photogrammetric Engineering and Remote Sensing*, vol. 51, pp. 681–687, 1985.

[30] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[31] R. Sznitman *et al.*, "Fast part-based classification for instrument detection in minimally invasive surgery," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2014, pp. 692–699.

[32] M. Allan *et al.*, "Toward detection and localization of instruments in minimally invasive surgery," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 4, pp. 1050–1058, 2013.

[33] S. J. Delany *et al.*, "Generating estimates of classification confidence for a case-based spam filter," in *International Conference on Case-Based Reasoning*. Springer, 2005, pp. 177–190.

[34] J. Orozco *et al.*, "Confidence assessment on eyelid and eyebrow expression recognition," in *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on*. IEEE, 2008, pp. 1–8.

[35] C. Zhang and R. L. Kodell, "Subpopulation-specific confidence designation for more informative biomedical classification," *Artificial Intelligence in Medicine*, vol. 58, no. 3, pp. 155–163, 2013.

[36] N. T. Clancy *et al.*, "Multispectral imaging of organ viability during uterine transplantation surgery," in *Progress in Biomedical Optics and Imaging-Proceedings of SPIE*, vol. 8935, 2014.

[37] K. P. Murphy *et al.*, "Loopy belief propagation for approximate inference: An empirical study," in *Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers Inc., 1999, pp. 467–475.

[38] A. T. Ihler *et al.*, "Loopy belief propagation: Convergence and effects of message errors," *Journal of Machine Learning Research*, vol. 6, no. May, pp. 905–936, 2005.

[39] H.-C. Shin *et al.*, "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1285–1298, 2016.

[40] Y. Gal, "Uncertainty in deep learning," Ph.D. dissertation, PhD thesis, University of Cambridge, 2016.

[41] N. Srivastava *et al.*, "Dropout: a simple way to prevent neural networks from overfitting." *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[42] A. Kendall and R. Cipolla, "Modelling uncertainty in deep learning for camera relocalization," in *International Conference on Robotics and Automation*. IEEE, 2016, pp. 4762–4769.

[43] C. Blundell *et al.*, "Weight uncertainty in neural networks," *arXiv preprint arXiv:1505.05424*, 2015.

[44] N. Pawlowski *et al.*, "Implicit weight uncertainty in neural networks," *arXiv preprint arXiv:1711.01297*, 2017.