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, 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 provides 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 90% (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.

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. 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) 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 ($S_{px}$)-based anatomical structure classification, which features an intrinsic confidence measure for self-performance estimation and which can be generalized to multispectral imaging (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. II-C): A comprehensive in vivo study is conducted using seven pigs to experimentally investigate hypotheses H1 and H2.

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.

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. [4], [5]). The method for multiple-organ segmentation in laparoscopy reported in [6] 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 [7] 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 [8] 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 [9]. 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. Multi-spectral 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 [9].

One of the first in vivo applications of MI was proposed by Afromowitz et al. [10], 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, Akkari et al. [11] and Triana et al. [12] 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. [13]. 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.

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). The experimental protocol designed to test the two hypotheses (H1 and H2) introduced in Sec. I is presented in Sec. II-C.

A. Uncertainty-aware tissue classification

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

1) Data acquisition: 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 light-filter central wavelengths ($\lambda_i$, where $i \in [1,8]$ is the band index) and the corresponding full widths at half maximum (FWHM) are reported in Table I. The filters were chosen according to the band selection strategy for endoscopic spectral imaging presented in [14]. The method makes use of the
Sheffield index [15], which is a theorectic band selection index 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 × 1029 × 8 for MI and 1228 × 1029 × 3 for RGB.

2) Pre-processing: To remove the influence of the dark current and to obtain the spectral reflectance image \( R(\lambda_i) \) for each band \( i = 1\ldots8 \), 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. 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 [16]:

\[
R(\lambda_i) = \frac{I(\lambda_i) - D(\lambda_i)}{W(\lambda_i) - D(\lambda_i)}
\]  

(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 [17]. 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 [18].

3) 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 [19]. 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) [19] 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. [13], 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) [20, 21], which fully meets these requirements, was here used to describe the tissue texture, where \( \tau \) and \( p \) represent the radius and the number of neighbors considered in the LBP computation, respectively. \( H_{LBP} \) 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 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 feature vector 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).

4) Superpixel-based classification: To classify the \( Spx \)-based features, we used SVM with the radial basis function and the one-against-one scheme. Prior to classification, we standardized the feature matrix within each feature dimension. As a prerequisite for our confidence estimation, we retrieved the probability \( P(Spx_n = j) \) for the \( n^{th} \) \( Spx \), \( n \in [1, N] \), to belong to the \( j^{th} \) organ \( (j \in [1, J]) \), where \( N \) and \( J \) are the numbers of considered \( Spx \) and considered organs, respectively. For the probability estimation, we trained the parameters of a sigmoid function to map the SVM outputs to probabilities [20].

5) Confidence estimation: To estimate the SVM classification performance, we used an intrinsic measure of confidence known as the Gini coefficient (GC) [21]. For each \( n^{th} \) \( Spx \), GC\((Spn)\) is the area under the Lorenz curve, which is the cumulative probability among the \( J \) outcome states rank-ordered according to the decreasing values of their individual probabilities \( P(Spn = 1), \ldots, P(Spn = J) \). GC has values of 0 and 1 for a uniform probability distribution (complete uncertainty) and for the case of a single state at 100% with the others at 0%, respectively.

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\((Spn)\) was thresholded to obtain binary confidence information. An \( Spx \) was considered to have an acceptable confidence level if \( GC(Spn) > \tau \), for the threshold \( \tau \).

C. In vivo validation

Seven pigs were used to examine the H1 and H2 introduced in Sec. I. Three pigs were used for training (29 images) and four for testing (28 images). We considered six porcine organ tissues typically encountered during hepatic laparoscopic surgery: the liver, gallbladder, spleen, diaphragm, intestines, 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

The multispectral images were pre-processed as described in Sec. I.A. The \( Spx \) segmentation with LSC was achieved using an average \( Spx \) size of 150² pixels and an \( Spx \) compactness factor of 0.1. Accordingly, 55 \( Spx \) on average were obtained for each image. The LBP were computed considering the
TABLE II: 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%)|

Fig. 2: Effect of confidence threshold $\tau$ 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.

Fig. 3: Confusion matrix for confidence threshold $\tau = 0.9$ and multispectral imaging. The values are in percentages and the colorbar indicates the number of superpixels.

Fig. 4: Image tagging accuracy ($Acc_{Tag}$) for RGB and multispectral imaging (MI) for Base case and following introduction of confidence measure ($\tau = 0.9$). The stars indicate significant differences.

III. RESULTS

The descriptive statistics of $Acc_{Spx}$ for the analyzed features are reported in Table II. 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.

As shown in Fig. 2 when $\tau$ was varied in $[0.5 : 0.1 : 1]$, the median $Acc_{Spx}$ for the MI increased monotonously to 99% ($\tau = 0.9$). The same trend was observed for the RGB, with an overall improvement of the median from 81% to 93%. For both the Base case and after introduction of the confidence measure, the MI outperformed the RGB (p-value < 0.05). Figure 3 shows the confusion matrix for MI and $\tau = 0.9$. 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%.

1) Investigation of H1: To investigate whether the 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 confidence measure. $Acc_{Spx}$ is defined as the ratio of correctly classified $Spx$ to all samples in the testing set. For image tagging, we computed the tagging accuracy ($Acc_{Tag}$) for different $\tau$, 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 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).

The feature extraction was implemented using OpenCV\(^1\). The classification was implemented using scikit-learn\(^2\).

Following $(r, p)$ combinations: (1, 8), (2, 16), and (3, 24). The SVM kernel parameters ($C = 10^4$ and $\gamma = 10^{-5}$) were retrieved during the training phase via grid-search. The grid-search spaces for $\gamma$ and $C$ were set to $[10^{-8}, 10^1]$ and $[10^1, 10^5]$, respectively, with 10 values spaced evenly on the log scale in both cases. The determined values for the hyperparameters were subsequently used in the testing phase.

\(^1\)http://opencv.org/\(^2\)http://scikit-learn.org/
When applied to endoscopic image tagging, the mean $\text{Acc}_{\text{Tag}}$ 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. The descriptive statistics are reported in Fig. 4. In this instance, the MI also outperformed the RGB both in the Base case and with the confidence measure (p-value < 0.05). Figure 6 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 are shown in Fig. 5. 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.

IV. DISCUSSION

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 distinguishing 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. Indeed, 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. 3).

The results obtained with the introduction of the confidence measure are comparable with those obtained by Zhang et al. [13] 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 [13] was, however, restored for our dataset once the low-confidence Spx were excluded.

These results are in keeping with those found in the literature for case reasoning [23], [24]. 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 [25], spam-filtering [26], and cancer recognition [27]. 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
confidentiality value. The same process was implemented when the $S_{px}$ segmentation failed to separate two organs. Also in this case, MI showed that it performs better than standard 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).

V. 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 $S_{px}$-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.

Future work will 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 [23].

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.

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, S. Vedula, S. Speidel, N. Navab, R. Kikinis, A. Park, M. Eisenmann, H. Feussner, G. Forestier, S. Giannarou et al., “Surgical Data Science: Enabling next-generation surgery,” arXiv preprint, arXiv:1701.06482, 2017.

[2] K. März, M. Hafezi, T. Weller, A. Saffari, M. Nolden, N. Fard, A. Majlesara, S. Zelzer, M. Maleshkova, M. Volovoky 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, J. Schuck, A.-L. Wekerle, H. Kennett, B. P. Müller-Stich, R. Dillmann, and S. Sengel, “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] J. Lee, J. Oh, S. K. Shah, X. Yuan, and S. J. Tang, “Automatic classification of digestive organs in wireless capsule endoscopy videos,” in ACM Symposium on Applied Computing. Association for Computing Machinery, 2007, pp. 1041–1045.

[5] P. W. Mewes, D. Neumann, O. Licegevic, J. Simon, A. L. Juloski, and E. Angelooulou, “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.

[6] M. S. Nosrati, J.-M. Peyrat, J. Abi-Nahed, D. Al-Alao, A. Al-Ansari, R. Abagheriheb, and G. Hamarnen, “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.

[7] A. Chhatkuli, A. Bartoli, A. Malti, and T. Collins, “Live image parsing in uterine laparoscopy,” in International Symposium on Biomedical Imaging ISBI 2014. IEEE, 2014, pp. 1263–1266.

[8] K. Prokopete, T. Collins, and A. Bartoli, “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.

[9] Q. Li, X. He, Y. Wang, H. Liu, D. Xu, and F. Guo, “Review of spectral imaging technology in biomedical engineering: achievements and challenges,” Journal of Biomedical Optics, vol. 18, no. 10, pp. 100901–100901, 2013.

[10] M. A. Afremowitiz, J. B. Callis, D. M. Heimbach, L. A. DeSoto, and M. K. Norton, “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.

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

[12] B. Triana, J. Cha, A. Shademan, A. Krieger, J. U. Kang, and P. C. Kim, “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.

[13] Y. Zhang, S. J. Wirkert, J. Izzat, H. Kennett, M. Wagner, B. Mayer, C. Steck, N. T. Clancy, D. S. Elson, and L. Maier-Hein, “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.

[14] S. J. Wirkert, N. T. Clancy, D. Stoyanov, S. Arya, G. B. Hanna, H.-P. Schlemmer, P. Sauer, D. S. Elson, and L. Maier-Hein, “Endoscopic Sheffield index for unsupervised in vivo spectral band selection,” in International Workshop on Computer-Assisted and Robotic Endoscopy, Springer, 2014, pp. 110–120.

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

[16] A. Majlesara, F. Marzani, and P. Gouton, “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.

[17] D.-J. Kroon, C. H. Slump, and T. J. Maal, “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.

[18] S. Moccia, V. Penza, G. O. Vanone, E. De Momi, and L. S. Mattos, “Automatic workflow for narrow-band laryngeal video stitching,” in Computer Vision and Pattern Recognition (CVPR), Conference on. IEEE, 2015, pp. 1254–1263.

[19] T.-F. Wu, C.-J. Lin, and R. C. Weng, “Probability estimates for multi-class classification by pairwise coupling,” Journal of Machine Learning Research, vol. 5, no. Aug, pp. 975–1005, 2004.
[21] B. G. Marcot, “Metrics for evaluating performance and uncertainty of
Bayesian network models,” Ecological Modelling, vol. 230, pp. 50–62,
2012.

[22] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion,
O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg et al.,
“Scikit-learn: Machine learning in Python,” Journal of Machine Learn-
ing Research, vol. 12, no. Oct, pp. 2825–2830, 2011.

[23] 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.

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

[25] S. J. Delany, P. Cunningham, D. Doyle, and A. Zamolotskikh, “Gener-
ating estimates of classification confidence for a case-based spam filter,”
in International Conference on Case-Based Reasoning. Springer, 2005,
pp. 177–190.

[26] J. Orozco, O. Rudovic, F. X. Roca, and J. Gonzalez, “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.

[27] C. Zhang and R. L. Kodell, “Subpopulation-specific confidence design-
ation for more informative biomedical classification,” Artificial Intelli-
gence in Medicine, vol. 58, no. 3, pp. 155–163, 2013.

[28] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao,
D. Mollura, and R. M. Summers, “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.