Medical & Biological Engineering & Computing manuscript No.
(will be inserted by the editor)

Transfer learning for informative-frame selection in laryngoscopic videos through learned features

Ilaria Patrini* · Michela Ruperti* · Sara Moccia · Leonardo S. Mattos · Emanuele Frontoni · Elena De Momi

Received: date / Accepted: date

Abstract Narrow-band imaging (NBI) laryngoscopy is an optical-biopsy technique used for screening and diagnosing cancer of the laryngeal tract, reducing the biopsy risks 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 develop a deep-learning-based strategy for the automatic selection of informative laryngoscopic-video frames, reducing the amount of data to process for diagnosis.

I. Patrini, M. Ruperti, E. De Momi
Department of Electronics, Information and Bioengineering, Politecnico di Milano
piazza Leonardo Da Vinci 32
Milan (Italy)

S. Moccia
Department of Information Engineering, Università Politecnica delle Marche
via Brecce Bianche 12
Ancona (Italy)
Department of Advanced Robotics, Istituto Italiano di Tecnologia
via Morego 30
Genoa (Italy)

L. S. Mattos
Department of Advanced Robotics, Istituto Italiano di Tecnologia
via Morego 30
Genoa (Italy)

E. Frontoni
Department of Information Engineering, Università Politecnica delle Marche
via Brecce Bianche 12
Ancona (Italy)

* These authors equally contributed to this paper

The strategy leans on the transfer learning process that is implemented to perform learned-features extraction using six different convolutional neural networks (CNNs) pre-trained on natural images. To test the proposed strategy, the learned features were extracted from the NBI-InfFrames dataset. Support vector machines (SVMs) and CNN-based approach were then used to classify frames as informative (I) and uninformative ones such as blurred (B), with saliva or specular reflections (S) and underexposed (U).

The best-performing learned-feature set was achieved with VGG 16 resulting in a recall of I of 0.97 when classifying frames with SVMs and 0.98 with the CNN-based classification. This work presents a valuable novel approach towards the selection of informative frames in laryngoscopic videos and a demonstration of the potential of transfer learning in medical image analysis.

Keywords Informative-frame selection · learned features · deep learning · transfer learning · laryngoscopy

1 Introduction

Optical imaging, such as microscopy and endoscopy, supports clinicians and surgeons in performing diagnosis and treatment [1]. Tissue analysis from optical images is crucial in several fields, such as ophthalmology [2, 3], laryngology [4, 5], and neurosurgery [6, 7].

In this paper we address the issue of enabling the application of surgical data science (SDS) methods in laryngology. In this field, the quality of laryngoscopic video frames can strongly affect the output of SDS tools. Indeed, the analysis of low-quality uninformative frames during endoscopy-based diagnosis, may increase the overall computational time required by SDS algorithms without providing any useful information. Moreover, there could be wrong classification outcomes when processing frames with low informative content, such as image with insufficient illumination [8].

A possible solution to identify and discard uninformative images consists in performing preliminary visual assessment of image quality. However, this operation is qualitative, prone to human error and usually time consuming [9]. A reasonable alternative to visual assessment is the automatic selection of informative frames, which is however not always trivial due to variability in image characteristics (e.g., noise level and resolution), image acquisition protocols, and tissue anatomy [10, 11].

To accomplish this task, several machine-learning (ML) approaches have been proposed (Sec. 2), which are mainly based on handcrafted features (e.g., features based on intensity or textural information). However, deep-learning algorithms may outperform standard learning approaches for image analysis, as shown by researchers in other SDS fields [12, 13, 14]. With deep learning, handcrafted features are replaced by learned features, which are automatically learned during a training process (i.e. without the need of manually defining

Table 1 State-of-the-art approaches to informative-frame selection.

Method	Year	Anatomical district	Feature set	Classification
Bashar et al. [18]	2010	Gastro-intestinal tract	Intensity and texture	Support vector machines
Atasoy et al. [19]	2012	Gastro-intestinal tract	Image power-spectrum histogram	Clustering
Park et al. [20]	2012	Colon	Anatomy-related	Conditional random fields
Maghsoudi et al. [21]	2014	Gastro-intestinal tract	Intensity	k-means
Ishijima et al. [22]	2015	Oral cavity-esophagus	Intensity, entropy and keypoints	Statistical comparison
Armin et al. [23]	2015	Colon	Motion, intensity and image derivatives	Random forest
Perperidis et al. [11]	2017	Lungs	Texture	Gaussian mixture model
Moccia et al. [10]	2018	Larynx	Intensity, entropy, keypoints and texture	Support vector machines

the mathematical formulation of the feature set) [15]. The most popular approach to automatic feature learning is using convolutional neural networks (CNNs), which showed remarkable performance in classifying skin cancers [13] and predicting cardiovascular risk factors from retinal fundus photographs [16]. A CNN is a neural network that consists of a series of different kinds of specialized layers, such as convolutional and pooling. The first (upper) CNN layers learn how to extract image features directly from the training images, thus the training set is commonly made of million of images to satisfactorily encode variability in the images (e.g., ImageNet [17], a dataset for natural-image classification, is made of more of 14-million images).

Collecting such a high number of labeled images is challenging in the medical field [1], despite the efforts of international organizations[1]. This problem may be overcome through transfer learning, in which CNNs store knowledge while solving one problem (e.g., natural-image classification) and apply it to a different one (such as medical-image classification) [16, 13]. Transfer learning has already been shown to be successful in several medical fields [24, 15, 25] such as the classification of chronic obstructive pulmonary disease (COPD) [26] or colorectal polyps [27], but no applications can be found in the field of informative-frame selection.

Thus, the specific aim of this work is to investigate if features learned with CNNs (pre-trained on natural images) can be exploited for informative-frame selection in endoscopic videos. In particular, for feature classification, both support vector machines (SVMs) and CNN-based approaches are investigated.

The experimental analysis is performed on the NBI-InfFrames dataset, which has been recently proposed in [10] for NBI endoscopic video-frames analysis and is available online[2]. Indeed, to the best of the authors' knowledge, it is the only labeled dataset that is publicly available in the field. All codes

[1] https://grandchallenges.org/
[2] DOI: 10.5281/zenodo.1162784

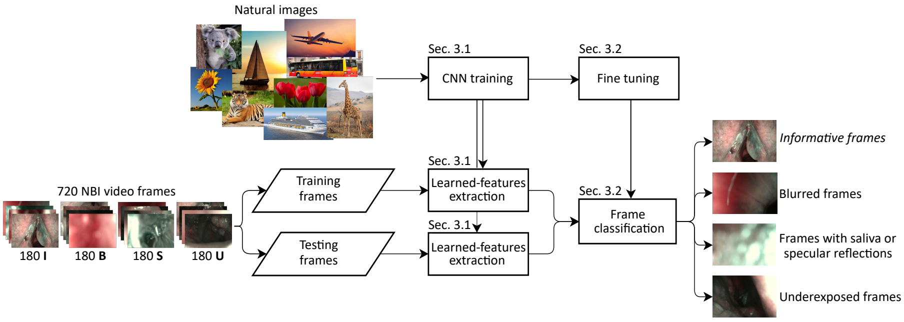


Fig. 1 Workflow of the proposed approach to informative frame selection in endoscopic videos in narrow-band imaging (NBI). Frames composing the videos are **I**: informative, **B**: blurred, **S**: with saliva and specular reflections, **U**: underexposed.

and CNN weights will be made publicly available upon publication of this work.

This paper is organized as follows: Sec. 2 surveys the approaches for informative-frame selection, Sec. 3 explains the proposed approaches to informative-frame selection. Sec. 4 deals with the experimental protocol used to test the proposed methodology. Results are presented in Sec. 5 and discussed in Sec. 6. Finally, Sec. 7 summarizes the main achievements of this work.

2 Related work

Strategies proposed in literature for informative-frame selection include simple uniform or random frame sampling (e.g., [28, 29] for bladder images), which are fast in terms of computational time but do not guarantee that all informative frames are extracted while removing the non-informative ones.

More recently, learning-based approaches have been proposed. Anatomical features are used in [20] to classify informative frames in colonoscopy videos with conditional random fields, while motion, edge and color features, along with random forests, are used in [23]. Image-frequency features from gastro-intestinal images are clustered with k-means in [19, 21], while in [18] local color-histogram features are classified with SVMs. A statistical approach to informative-frame selection in esophageal microscopy images, which exploits intensity, entropy and keypoint-based features, is proposed in [22]. Texture-based features from lung microscopy images are classified with Gaussian mixture models in [11]. In [10], a set of intensity, keypoint-based and textural features and multi-class SVMs are used to classify informative and three classes of uninformative frames in laryngoscopic videos in narrow-band imaging (NBI).

Table 1 summarizes state-of-the-art approaches to informative-frame selection.

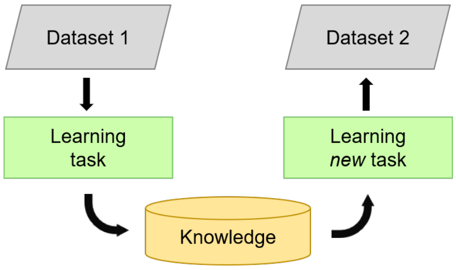


Fig. 2 Graphical representation of transfer learning approach. The knowledge (features, weights) that a model has learned from a task (e.g., natural-image classification) where a lot of labeled training data are available (dataset 1) is exploited and transferred to another task, such as medical-image classification, with less data (dataset 2).

3 Methods

In this section, the proposed strategies to feature extraction (Sec. 3.1) and classification (Sec. 3.2) are explained. The workflow of the proposed approach is shown in Fig. 1.

3.1 Transfer learning for learned-features extraction

In this paper, learned-feature extraction was performed exploiting a transfer-learning approach. As illustrated in Fig. 2, transfer learning focuses on storing the knowledge or weights of a trained neural network so that it can be reused for a further task [30].

Practically, it generalizes the knowledge (features, weights) of an existing solution to a new problem, leading to promising results also when the new task has significantly less data. This fits well in the case of problems in the computer vision domain where certain low-level features (e.g. edges, shapes, corners and intensity) can be shared across tasks, and thus enable knowledge transfer among them.

As reported in [31], the formal definition of transfer learning involves the concepts of a domain and a task. A domain $\mathcal{D}$ consists of a feature space $\mathcal{X}$ and a marginal probability distribution *P(X)* over the feature space, where X $= \{x_1, \ldots, x_n\} \in \mathcal{X}$. Given a specific domain, $\mathcal{D} = \{\mathcal{X}, P(X)\}$, a task consists of two components: a label space $\mathcal{Y}$ and an objective predictive function $f(\cdot\Delta)$ (denoted by $\mathcal{T} = \{\mathcal{Y}, f(\cdot)\}$), which is not observed but can be learned from the training data, which consist of pairs $\{x_i, y_i\}$, where $x_i \in$ X and $y_i \in \mathcal{Y}$.

Given a source domain $\mathcal{D_S}$ and learning task $\mathcal{T_S}$, a target domain $\mathcal{D_T}$ and learning task $\mathcal{T_T}$, transfer learning aims to help improve the learning of the target predictive function $f_T(\cdot)$ in $\mathcal{D_T}$ using the knowledge in $\mathcal{D_S}$ and $\mathcal{T_S}$, where $\mathcal{D_S} \neq \mathcal{D_T}$, or $\mathcal{T_S} \neq \mathcal{T_T}$.

Six pre-trained CNNs (Table 2) were investigated to perform learned-feature extraction. The CNNs used were chosen among the best performing

Table 2 Tested convolutional neural networks (CNNs) and the corresponding number of learned features. Top-1 and top-5 accuracies achieved on the ImageNet dataset are reported too. These accuracies refer to the fractions of test images for which the correct label is the first (top-1) or among the five labels (top-5) considered most probable by the model, respectively.

CNNs	Number of features	Top-1 accuracy	Top-5 accuracy
VGG 16	4096	71.5%	89.8%
Inception V4	1536	80.2%	95.2%
ResNet V1 101	2048	76.4%	93.2%
ResNet V1 152	2048	76.8%	92.9%
ResNet V2 152	2048	87.8%	94.1%
Inception - ResNet V2	1536	80.4%	95.3%

in the context of the Large Scale Visual Recognition Challenge (ILSVRC)[3]. For fair comparison, all the CNN models were pre-trained on the ImageNet dataset[4].

The tested CNN architectures, hereafter briefly described to highlight their main peculiarities, were:

VGG 16 VGG 16 was proposed by the Oxford's Visual Geometry Group (VGG) in the context of ILSVRC in 2014.

VGG 16 improved the performance of previously proposed deep networks (e.g., AlexNet [32]) by replacing large-sized kernel filters with stacked kernels with dimension 3x3 pixels. The multiple stacked small-sized kernels allowed increased performance by enabling the VGG 16 to learn complex and fine-level features while making the training convergence easier and faster [33].

VGG 16 has a uniform (serial) architecture with 13 convolutional and 5 (down-sampling) max-pooling layers, followed by 3 fully-connected layers. The number of layer channels starts from 64 filters and increases by a factor of 2 after every pooling layer.

VGG 16 achieved the top-1 accuracy of 71.5% and the top-5 accuracy of 89.8% on the ImageNet dataset, where the top-1 and top-5 accuracies are the fractions of test images for which the correct label is the first (top-1) or among the five labels (top-5) considered most probable by the model, respectively.

Inception V4 The winner of ILSVRC 2014 competition was GoogLeNet (i.e., Inception V1) developed by Google LLC.

The innovative idea of GoogLeNet is the introduction of the Inception module. The input image to the module is convolved with parallel filters of different sizes (1x1, 3x3, 5x5), thus losing the CNN linear structure (such the one of VGG 16), to allow a multi-scale feature extraction. Several versions of the Inception module were proposed and the upgraded version Inception V4 was used here, as it showed the best performance [34].

The Inception V4 architecture is organized in 10 blocks, for a total of 14 inception modules linearly stacked with global average pooling at the end.

[3] http://www.image-net.org/challenges/LSVRC/

[4] http://www.image-net.org/

The 14 Inception modules are different in terms of convolutional-kernel size, number of filters and depth. This allows to process the image at varying scale when it passes through the CNN modules.

Top-1 and top-5 accuracies of Inception V4 of 80.2% and 95.2% were achieved on the ImageNet dataset, respectively.

ResNet V1 101 and ResNet V1 152 The Residual Neural Networks (ResNets) presented in [35] got the first place in the ILSVRC 2015 classification competition and the first place in ILSVRC and COCO 2015 competition in ImageNet Detection, ImageNet localization, Coco detection and Coco segmentation [35].

ResNets were introduced to solve the vanishing gradient and the gradation problem that arise when training ultra-deep CNNs. They consist of many stacked residual units (building blocks) containing skip connections to link the input and output of each unit. CNNs with residual units were shown to outperform their plain counterparts [35].

In this work, ResNets with 101 layers and 152 layers were tested. ResNet 101 consists of 4 main layers and the number of building blocks varies in each layer (3, 4, 23 and 3, respectively). Each building block is made of 3 convolutional kernels with the skip connection. ResNet 152 has the same structure but with 8 and 36 building blocks in the second and third layer, respectively. Similarly to Inception V4, both ResNets end with a global average pooling followed by a classification layer.

ResNet V1 101 and ResNet V1 152 achieved top-1 accuracies of 76.4% and 76.8% and top-5 accuracies of 93.2% and 92.9%, respectively on the ImageNet dataset.

ResNet V2 152 ResNet V2 was introduced in [36] with the goal of using pre-activation in ResNets. Pre-activation consists in using activation functions (such as the ReLU) as pre-activation of the convolutional layers, in contrast to conventional ResNets where the activation functions are used as post-activation. Pre-activation was demonstrated to have an impact both in terms of ease of optimization and improved regularization.

The version of ResNet V2 with 152 layers was tested here, which is ResNet V1 152 with pre-activation. It achieved top-1 and top-5 accuracy of 87.8% and 94.1% on ImageNet, respectively.

Inception - ResNet V2 A hybrid Inception module was proposed in [34] by Szegedy et al., which is called Inception - ResNet V2.

This architecture significantly improved the recognition performance of both ResNet V2 and Inception V4, and dramatically increased the training speed when tested on ImageNet dataset [34]. Inception - ResNet V2 was built by adding residual connections to link the input and output of the Inception V4 blocks.

Top-1 and top-5 validation accuracies of 80.4% and 95.3% were achieved on ImageNet, respectively.

Table 3 Tested conditions in this work. Condition 1 (**C1**) exploits the convolutional neural networks (CNNs) as features extractor and the extracted learned features are then classified by means of support vector machines (SVMs); in condition 2 (**C2**) the best performing CNN resulting in **C1** is fine-tuned and it is used both as feature extractor and classifier.

Tested conditions	Features extractor	Classifier
C1	All CNNs presented in Sec. 3.1	Support vector machines (SVMs)
C2	Best performing CNN of C1	Fine-tuned version of the fully connected CNN

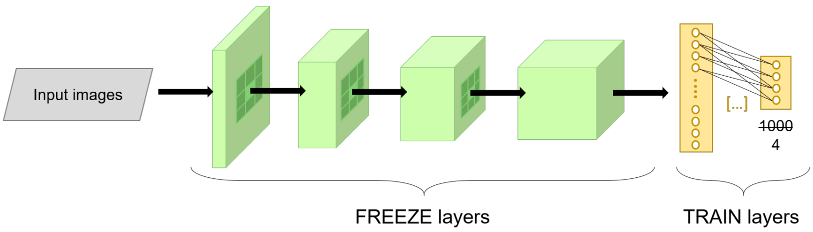


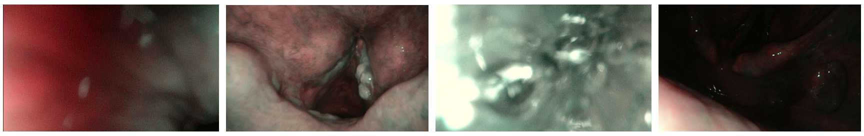
Fig. 3 Graphical representation of the fine-tuning technique. Weights of the first layers are frozen, since they refer to general features, while the weights of the last layers (or at least the ones of the fully-connected layer) are learnt on the target dataset.

3.2 Frame classification

Learned-feature matrices were standardized before classification [37]. Feature classification was first performed exploiting SVMs [37] (**C1** in Table 3) to tackle the high-dimensionality of the input features (> 1500) while being robust to noise in the features [38]. SVMs with Gaussian kernel (Ψ) were used to prevent parameter proliferation, limiting the computational complexity. To implement multi-class SVM classification, the *one-vs-rest* scheme was used. Thus, when one class was considered positive, the remaining ones were considered negative.

The SVM hyperparameters, i.e. kernel coefficient (γ) and penality parameter (C), were retrieved via grid-search and cross-validation as explained in Sec. 4.

We also investigated the performance of the CNNs using the best-performing learned-feature for frame classification (**C2** in Table 3). To this goal, fine-tuning was implemented by freezing the weights of the first CNN layers and learning the layers of the fully-connected layers [39], as reported in Fig. 3. Indeed, the first layers contain more generic features (e.g. edge detectors or color blob detectors) that should be useful to many tasks, while the last layers become progressively more specific to the details of the classes contained in the original dataset [12]. In order to accomplish this task, Gradient Descent Optimizer (GDO) was used.



(a) **B** (b) **I** (c) **S** (d) **U**

Fig. 4 Samples of laryngeal video frames in the NBI-InfFrames. Frames were **B**: blurred, **I**: informative, **S**: with saliva and specular reflections, **U**: underexposed.

4 Experimental protocol

4.1 Dataset

The performance of the learned features extracted with the CNNs presented in Sec. 3.1 were evaluated on the NBI-InfFrames (as introduced in Sec. 1), which was built from 18 NBI endoscopic videos, referring to 18 different patients affected by squamous cell carcinoma (SCC). All 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.

The NBI-InfFrames consists of a total of 720 video frames, which are equally divided in four classes: informative (**I**), blurred (**B**), with saliva or specular reflections (**S**), and underexposed (**U**). The NBI-InfFrames dataset is balanced both at patient level and frame class level. Sample images for the four classes are shown in Fig. 4.

To extract learned features with the architectures presented in Sec. 3.1, all the images were resized to match the input size of the investigated CNN architectures. The deepest layer of each tested CNN was considered as feature extractor. The feature length for each feature set is shown in Table 2.

4.2 Parameters Tuning

For performing the classification with SVM, γ and C (the SVM hyper-parameters) were retrieved via grid-search and 10-fold cross validation. The grid-search spaces for γ and C were set to $[10^{-8}, 10]$ and $[10^{-3}, 10^{6}]$, respectively, with 10 values evenly spaced on log_{10} scale in both cases.

For performing the CNN-based classification, fine tuning was implemented using the GDO with a learning rate of 10^{-5}.

To estimate performances, 3-fold cross-validation was performed separating data at patient level, as suggested in [10].

4.3 Data analysis

The classification performance of each CNN model was evaluated computing the class-specific recall ($\mathbf{Rec_{class}} = \{Rec_{class_j}\}_{j\in[1,4]}$), the precision ($\mathbf{Prec_{class}}$

$= \{Prec_{class_j}\}_{j\in[1,4]}$), the F1 score ($\mathbf{F1_{class}} = \{F1_{class_j}\}_{j\in[1,4]}$), where:

$$Rec_{class_j} = \frac{TP_j}{TP_j + FN_j} \tag{1}$$

$$Prec_{class_j} = \frac{TP_j}{TP_j + FP_j} \tag{2}$$

$$F1_{class_j} = 2\frac{Prec_{class_j} \times Rec_{class_j}}{Prec_{class_j} + Rec_{class_j}} \tag{3}$$

being TP_j the number of true positive of the j^{th} class, FN_j the number of false negative of the j^{th} class and FP_j the number of false positive of the j^{th} class. The area (AUC) under the receiver operating characteristic (ROC) was also computed. As the classification problem was a multi-class problem (with a balanced dataset), we computed the macro-average ROC to compare the different CNN approaches to learned-feature extraction. For the best-performing feature set, i.e. the one that gave the highest recall for **I**, we performed the ROC analysis for each class.

The features learned with the investigated CNN architectures were compared with the set of features proposed in [10] in terms of classification performance. Only the method presented in [10] was considered, as it had already been shown to outperform previous literature on the topic. For the sake of completeness, we used the Wilcoxon signed-rank test (significance level = 0.05) for paired sample to assess whether the classification achieved with our best performing feature vector significantly differs from the ones achieved with the other feature sets.

Feature extraction and feature classification were implemented with Tensorflow[5] and scikit-learn[6], respectively. All the TensorFlow CNN-model files and the CNN weights were downloaded from the TensorFlow-Slim image classification model library[7].

Experiments were performed on Intel® Core TM i7-4500 CPU @ 1.80 GHz - 2.40 GHz with 8 GB of available RAM; NVIDIA GeForce GT 740 M; Microsoft Windows 10 64-bit operating system.

5 Results

For SVM-based classification (**C1** in Table 3), the macro-averaging ROC for the investigated CNN architectures are shown in Fig. 5. With the best-performing learned-feature set (obtained with VGG 16), an AUC of 0.9856 was achieved.

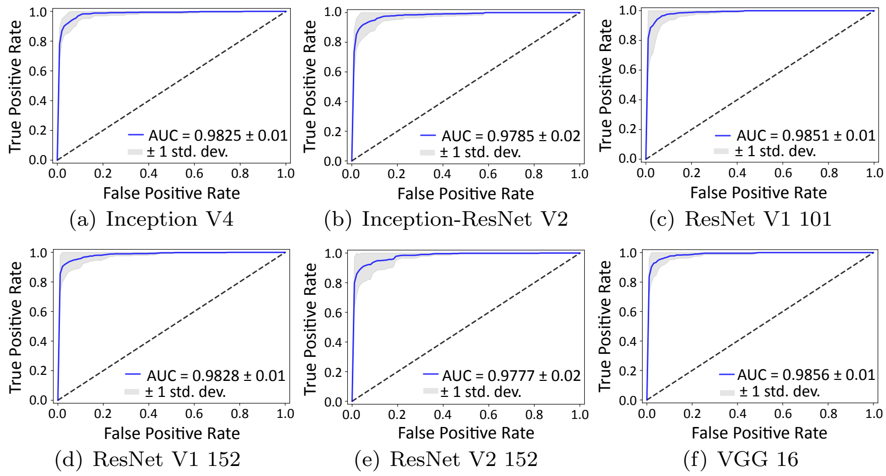
[5] http://www.tensorflow.org
[6] http://scikit-learn.org
[7] https://github.com/tensorflow/models/tree/master/research/slim

Table 4 Support vector machines (SVMs)-based classification performance in terms of class-specific precision ($\mathbf{Prec_{class}}$), recall ($\mathbf{Rec_{class}}$) and F1-score ($\mathbf{F1_{class}}$) are reported for the four different classes. **B**: blurred frames, **I**: informative frames, **S**: frames with saliva or specular reflections, **U**: underexposed frames. Results from the state of art [10] report only two significant digits.

	$\mathbf{Prec_{class}}$	$\mathbf{Rec_{class}}$	$\mathbf{F1_{class}}$
Moccia et al., 2018 [10]			
B	0.76	0.83	0.79
I	0.91	0.91	0.91
S	0.78	0.62	0.69
U	0.76	0.85	0.80
avg/total	0.80	0.80	0.80
Inception V4			
B	0.8883	0.9722	0.9284
I	0.9441	0.8444	0.8915
S	0.9012	0.8611	0.8807
U	0.8526	0.9000	0.8757
avg/total	0.8966	0.8944	0.8941
Inception-ResNet V2			
B	0.8737	0.9611	0.9153
I	0.9571	0.8667	0.9096
S	0.8824	0.8333	0.8571
U	0.8836	0.9278	0.9051
avg/total	0.8992	0.8972	0.8968
ResNet V1 101			
B	0.8947	0.9444	0.9189
I	0.9824	0.9278	0.9543
S	0.9128	0.8722	0.8920
U	0.9415	0.9833	0.9620
avg/total	0.9329	0.9319	0.9318
ResNet V1 152			
B	0.9259	0.9722	0.9485
I	0.9881	0.9222	0.9540
S	0.8913	0.9111	0.9011
U	0.9441	0.9389	0.9415
avg/total	0.9374	0.9361	0.9363
ResNet V2 152			
B	0.9198	0.9556	0.9373
I	0.9603	0.8056	0.8761
S	0.8316	0.9056	0.8670
U	0.9140	0.9444	0.9290
avg/total	0.9064	0.9028	0.9024
VGG 16			
B	0.9202	0.9611	0.9402
I	**0.9722**	**0.9722**	**0.9722**
S	0.9349	0.8778	0.9054
U	0.9180	0.9333	0.9256
avg/total	**0.9363**	**0.9361**	**0.9359**

The ROC curves relative to each of the four frame classes for the VGG 16-based feature set are shown in Fig. 6(a). AUC values were 0.9973 for informative frames (**I**), 0.9881 for blurred frames (**B**), 0.9862 for frames with saliva or specular reflections (**S**) and 0.9852 for underexposed frames (**U**).

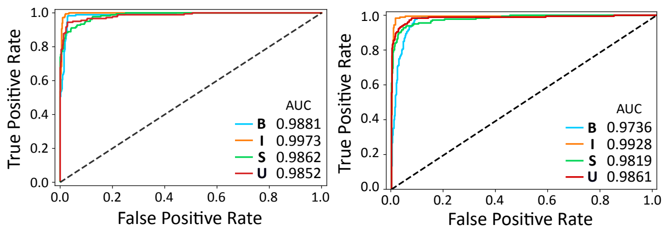
From the confusion matrix relative to VGG 16 in Fig. 7(a) , **Rec** of 0.9722 was achieved for **I**, 0.9611 for **B**, 0.8778 for **S** and 0.9333 for **U**. The median **Rec** among the four classes was 0.9361. Misclassification mainly occurred between **S** and **I**, probably due to the presence of image-intensity edges in **S** frames (e.g., saliva blobs and specular reflections) as in **I** frames.



(a) Inception V4 (b) Inception-ResNet V2 (c) ResNet V1 101

(d) ResNet V1 152 (e) ResNet V2 152 (f) VGG 16

Fig. 5 Macro-averaging receiver operating characteristic (ROC) curves for the investigated architectures coupled with support vector machines (SVMs). The area under the ROC (AUC) for each architecture is reported, too.



(a) VGG 16 (features) + SVM (b) VGG 16

Fig. 6 Receiver operating characteristic (ROC) curves for the four frame classes obtained with VGG 16. Features learned with VGG 16 are classified with (a) support vector machines (SVM) and (b) fully-connected layers. The area under the ROC (AUC) for each class is reported, too. **B**: blurred frames, **I**: informative frames, **S**: frames with saliva or specular reflections, **U**: underexposed frames.

The fine-tuned VGG 16-based classification (**C2** in Table 3) achieved values of 0.9778 for the **Rec** of **I**, 0.9333 for **B**, 0.8556 for **S** and 0.9389 for **U**. The ROC curves relative to each of the four frame classes are shown in Fig. 6(b).

Fig. 8 reports the comparison in terms of $\mathbf{Rec_{class}}$ for the tested CNN architectures and the method presented in [10]. Learned features always outperformed the handcrafted ones proposed in [10]. AUC values obtained with VGG 16-based classification were 0.9928 for informative frames (**I**), 0.9736 for blurred frames (**B**), 0.9819 for frames with saliva or specular reflections (**S**) and 0.9861 for underexposed frames (**U**).

No significant differences were found when applying the Wilcoxon signed-rank test (p-value > 0.05) to the $\mathbf{Rec_{class}}$ vectors constituted by the $\mathbf{Rec_{class}}$ of each feature sets extracted by means of each tested architecture (**C1** in

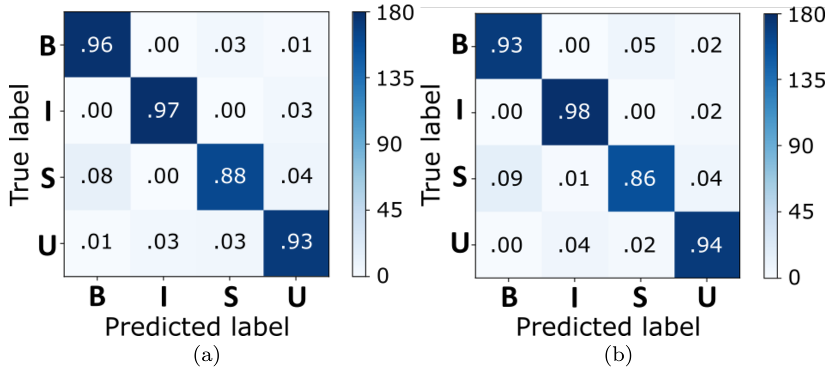


Fig. 7 Normalized confusion matrices for the best performing architecture (i.e., VGG 16): 7(a) SVMs and 7(b) fine-tuned CNN-based classification. **B**: blurred frames, **I**: informative frames, **S**: frames with saliva or specular reflections, **U**: underexposed frames. The colorbar indicates the number of images.

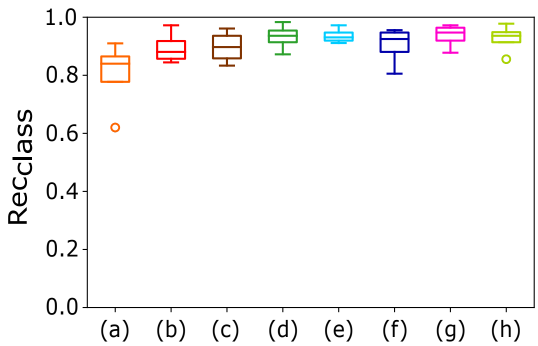


Fig. 8 Boxplots of classification recall (Rec_{class}) obtained for (a) [10], for features classified with SVMs and extracted with (b) Inception V4, (c) Inception-ResNet V2, (d) ResNet V1 101, (e) ResNet V1 152, (f) ResNet V2 152, (g) VGG 16, and (h) for features extracted and classified with VGG 16.

Table 3), including the one of the state of art, with the $\mathbf{Rec_{class}}$ vector of the resulted best feature set. The same for the comparison between the $\mathbf{Rec_{class}}$ vector of the resulted best feature set (**C1** in Table 3) and the $\mathbf{Rec_{class}}$ vector of the feature set extracted from the fine-tuned version of the best network (**C2** in Table 3).

6 Discussion

In this paper, we presented and evaluated a strategy for informative frame selection that exploits learned features automatically extracted from CNNs that were pre-trained on natural images.

When comparing the CNN architectures (**C1** in Table 3), VGG 16 outperformed all the others in terms of **Rec** of **I**, as reported in Table 4. A visual comparison of the classifications of **I** by the investigated CNN models is reported in Table 5. In Table 6, the number of misclassified **I** is reported, clearly showing that VGG 16 lowered the misclassification rate. The reason for this could be seen in the relatively simple (i.e., serial, without branches or skip connections) architecture and small depth (16 layers) of the VGG 16 architecture. This resulted in the extraction of more generalizable features and led to a more successful transfer learning.

From the comparison with the handcrafted-based approaches in the literature, and in particular with [10] (considered the state-of-the-art up to now as it outperformed all previously published methods), the learned features extracted with all the tested architectures showed higher $\mathbf{Rec_{class}}$ when classifying blurred frames, frames with saliva or specular reflections and underexposed frames. Moreover, three out of the six CNN architectures (i.e. ResNet V1 101, ResNet V1 152, and VGG 16) also showed higher value of **Rec** for informative frames. This confirmed considerations asserted in the literature of other SDS fields. In fact, deep-learning strategies coupled with transfer learning for feature extraction are often showing higher performance than standard machine learning for handcrafted-feature classification [40]. This has a crucial role in the medical field, where achieving high classification performance is necessary but labeled datasets large enough to train a CNN model from scratch are challenging to collect [3, 40].

6.1 Impact of fine-tuning technique

The fine-tuned VGG 16-based classification (**C2** in Table 3) showed higher value of **Rec** for **I** and of **Rec** for **U** compared to SVMs classification, while the other two classes (i.e., **B**, **S**) achieved comparable results. One possible reason for this could be due to the relative small size of the dataset (less than a thousand samples), for which SVM may be more suitable [12], because of the particular ability at drawing decision boundaries on a small dataset.

Hence, as future work, we aim at enlarging the dataset exploiting Generative Adversarial Networks in order to enable better fine-tuning of the proposed system to potentially solve misclassification problems.

We are also interested in investigating contentious-learning strategies that use feedback from clinicians to train and tune the classification model in real time. Furthermore, we intend to explore the performance of the proposed algorithm when applied to endoscopy and microscopy videos of different anatomical regions, such as the gastro-intestinal tract.

7 Conclusion

This paper presented a method for endoscopic informative-frame classification that exploited CNN-based learned features through transfer learning and cou-

Table 5 Sample informative frames (**I**) and relative classification for each tested convolutional neural network (CNN). The red and green boxes correspond to misclassification and correct classification, respectively. **S**: frames with saliva or specular reflections, **U**: underexposed frames.

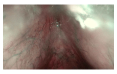
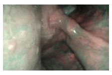
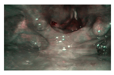
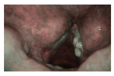
CNNs					
Inception V4	U	I	U	U	I
Inception-ResNet V2	S	I	U	U	I
ResNet V1 101	S	I	I	U	I
ResNet V1 152	S	I	I	U	I
ResNet V2 152	S	S	U	U	I
VGG 16	I	I	I	U	I

Table 6 Tested convolutional neural networks (CNNs) and corresponding number of informative frames (**I**) misclassified as frames with saliva or specular reflections (**S**) and as underexposed (**U**). No **I** frame was misclassified as blurred (**B**) by any CNN. The total number of informative frames is 180.

CNNs	**No. of I misclassified as S**	**No. of I misclassified as U**
Inception V4	6	22
Inception - ResNet V2	9	15
ResNet V1 101	7	6
ResNet V1 152	10	4
ResNet V2 152	22	13
VGG 16	**0**	**5**

pled with SVM multi-class classification, and classification after fine-tuning implementation on the pre-trained CNN. With our experimental protocol, the overall median classification recall among the four frame classes (i.e. **B**, **I**, **S**, **U**) for the best-performing learned features (VGG 16) set, coupled with transfer learning and SVM multi-class classification, was 93.61% (max recall = 97.22% for the informative frames). The overall median recall among the four frame classes achieved with the fine-tuned VGG 16-based classification was 92.64% (max recall = 97.78% for the informative frames). Both approaches outperformed the state of the art.

To conclude, this research demonstrated that using learned features obtained through transfer learning, together with SVMs or CNN-based classification, is an effective approach for the classification of informative frames in endoscopic videos.

References

1. Maier-Hein L, Vedula SS, Speidel S, Navab N, Kikinis R, Park A, Eisenmann M, Feussner H, Forestier G, Giannarou S, et al (2017) Surgical data science for next-generation interventions. Nature Biomedical Engineering 1(9):691
2. Campochiaro PA (2015) Molecular pathogenesis of retinal and choroidal vascular diseases. Progress in Retinal and Eye Research

3. Moccia S, De Momi E, El Hadji S, Mattos LS (2018) Blood vessel segmentation algorithms – Review of methods, datasets and evaluation metrics. Computer Methods and Programs in Biomedicine 158:71–91
4. Piazza C, Del Bon F, Peretti G, Nicolai P (2012) Narrow band imaging in endoscopic evaluation of the larynx. Current Opinion in Otolaryngology & Head and Neck Surgery 20(6):472–476
5. Moccia S, De Momi E, Guarnaschelli M, Savazzi M, Laborai A, Guastini L, Peretti G, Mattos LS (2017) Confident texture-based laryngeal tissue classification for early stage diagnosis support. Journal of Medical Imaging 4(3):034,502
6. Essert C, Fernandez-Vidal S, Capobianco A, Haegelen C, Karachi C, Bardinet E, Marchal M, Jannin P (2015) Statistical study of parameters for deep brain stimulation automatic preoperative planning of electrodes trajectories. International Journal of Computer Assisted Radiology and Surgery 10(12):1973–1983
7. Moccia S, Foti S, Routray A, Prudente F, Perin A, Sekula RF, Mattos LS, Balzer JR, Fellows-Mayle W, De Momi E, et al (2018) Toward improving safety in neurosurgery with an active handheld instrument. Annals of Biomedical Engineering pp 1–15
8. Gómez P, Semmler M, Schützenberger A, Bohr C, Döllinger M (2019) Low-light image enhancement of high-speed endoscopic videos using a convolutional neural network. Medical & Biological Engineering & Computing pp 1–13
9. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP (2004) Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing 13(4):600–612
10. Moccia S, Vanone GO, De Momi E, Laborai A, Guastini L, Peretti G, Mattos LS (2018) Learning-based classification of informative laryngoscopic frames. Computer Methods and Programs in Biomedicine 158:21–30
11. Perperidis A, Akram A, Altmann Y, McCool P, Westerfeld J, Wilson D, Dhaliwal K, McLaughlin S (2017) Automated detection of uninformative frames in pulmonary optical endomicroscopy. IEEE Transactions on Biomedical Engineering 64(1):87–98
12. Goodfellow I, Bengio Y, Courville A, Bengio Y (2016) Deep learning, vol 1. MIT press Cambridge
13. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S (2017) Dermatologist-level classification of skin cancer with deep neural networks. Nature 542(7639):115–118
14. Wang Q, Zheng Y, Yang G, Jin W, Chen X, Yin Y (2018) Multiscale rotation-invariant convolutional neural networks for lung texture classification. IEEE Journal of Biomedical and Health Informatics 22(1):184–195
15. Nanni L, Ghidoni S, Brahnam S (2017) Handcrafted vs. non-handcrafted features for computer vision classification. Pattern Recognition 71:158–172
16. Poplin R, Varadarajan AV, Blumer K, Liu Y, McConnell MV, Corrado GS, Peng L, Webster DR (2018) Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. Nature Biomedical

Engineering p 1
17. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, et al (2015) Imagenet large scale visual recognition challenge. International Journal of Computer Vision 115(3):211–252
18. Bashar MK, Kitasaka T, Suenaga Y, Mekada Y, Mori K (2010) Automatic detection of informative frames from wireless capsule endoscopy images. Medical Image Analysis 14(3):449–470
19. Atasoy S, Mateus D, Meining A, Yang GZ, Navab N (2012) Endoscopic video manifolds for targeted optical biopsy. IEEE Transactions on Medical Imaging 31(3):637–653
20. Park SY, Sargent D, Spofford I, Vosburgh KG, A-Rahim Y (2012) A colon video analysis framework for polyp detection. IEEE Transactions on Biomedical Engineering 59(5):1408
21. Maghsoudi OH, Talebpour A, Soltanian-Zadeh H, Alizadeh M, Soleimani HA (2014) Informative and uninformative regions detection in WCE frames. Journal of Advanced Computing 3(1):12–34
22. Ishijima A, Schwarz RA, Shin D, Mondrik S, Vigneswaran N, Gillenwater AM, Anandasabapathy S, Richards-Kortum R (2015) Automated frame selection process for high-resolution microendoscopy. Journal of Biomedical Optics 20(4):046,014
23. Armin MA, Chetty G, Jurgen F, De Visser H, Dumas C, Fazlollahi A, Grimpen F, Salvado O (2015) Uninformative frame detection in colonoscopy through motion, edge and color features. In: International Workshop on Computer-Assisted and Robotic Endoscopy, Springer, pp 153–162
24. Kumar A, Kim J, Lyndon D, Fulham M, Feng D (2017) An ensemble of fine-tuned convolutional neural networks for medical image classification. IEEE Journal of Biomedical and Health Informatics 21(1):31–40
25. Yoo TK, Choi JY, Seo JG, Ramasubramanian B, Selvaperumal S, Kim DW (2019) The possibility of the combination of oct and fundus images for improving the diagnostic accuracy of deep learning for age-related macular degeneration: a preliminary experiment. Medical & Biological Engineering & Computing 57(3):677–687
26. Cheplygina V, Pena IP, Pedersen JH, Lynch DA, Sørensen L, de Bruijne M (2018) Transfer learning for multicenter classification of chronic obstructive pulmonary disease. IEEE Journal of Biomedical and Health Informatics 22(5):1486–1496
27. Zhang R, Zheng Y, Mak TWC, Yu R, Wong SH, Lau JY, Poon CC (2017) Automatic detection and classification of colorectal polyps by transferring low-level CNN features from nonmedical domain. IEEE Journal of Biomedical and Health Informatics 21(1):41–47
28. Behrens A (2008) Creating panoramic images for bladder fluorescence endoscopy. Acta Polytechnica 48(3)
29. Weibel T, Daul C, Wolf D, Rösch R, Guillemin F (2012) Graph based construction of textured large field of view mosaics for bladder cancer

diagnosis. Pattern Recognition 45(12):4138–4150
30. Yosinski J, Clune J, Bengio Y, Lipson H (2014) How transferable are features in deep neural networks? In: Advances in Neural Information Processing Systems, pp 3320–3328
31. Pan SJ, Yang Q (2009) A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering 22(10):1345–1359
32. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp 1097–1105
33. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:14091556
34. Szegedy C, Ioffe S, Vanhoucke V, Alemi AA (2017) Inception-v4, inception-resnet and the impact of residual connections on learning. In: Association for the Advancement of Artificial Intelligence, vol 4, p 12
35. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 770–778
36. He K, Zhang X, Ren S, Sun J (2016) Identity mappings in deep residual networks. In: European Conference on Computer Vision, Springer, pp 630–645
37. Burges CJ (1998) A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery 2(2):121–167
38. Lin Y, Lv F, Zhu S, Yang M, Cour T, Yu K, Cao L, Huang T (2011) Large-scale image classification: fast feature extraction and SVM training. In: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, pp 1689–1696
39. Tajbakhsh N, Shin JY, Gurudu SR, Hurst RT, Kendall CB, Gotway MB, Liang J (2016) Convolutional neural networks for medical image analysis: full training or fine tuning? IEEE transactions on Medical Imaging 35(5):1299–1312
40. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, van der Laak JA, Van Ginneken B, Sánchez CI (2017) A survey on deep learning in medical image analysis. Medical Image Analysis 42:60–88

Ilaria Patrini (B.Sc. 2016) is a M.Sc. student of Biomedical Engineering in the Electronic Information and Bioengineering Department (DEIB) of Politecnico di Milano. She is currently doing her M.Sc. thesis at the Neuroengineering and Medical Robotics Laboratory (NearLab).

Michela Ruperti (B.Sc. 2016) is a M.Sc. student of Biomedical Engineering in the Electronic Information and Bioengineering Department (DEIB) of Politecnico di Milano. She is currently doing her M.Sc. thesis at the Neuroengineering and Medical Robotics Laboratory (NearLab).

Sara Moccia (B.Sc. 2012, M.Sc. 2014, Ph.D. 2018) was born in Bari (Italy) on September 1990. She graduated cum laude in Biomedical Engineering at Politecnico di Milano (Milan, Italy) in December 2014, with a thesis entitled: "Statistical-segmentation techniques of liver metastases and necroses in FGD-PET for the automatic evaluation of pre and post thermoablation PET/CT studies". In May 2018, she obtained the European PhD cum laude in Bioengineering from Istituto Italiano di Tecnologia, Department of Advanced Robotics (Genoa, Italy) and Politecnico di Milano, Department of Electronics, Information and Bioengineering with a thesis entitled "Supervised tissue classification in optical images: Towards new applications of surgical data science". During her PhD, she was hosted at the Department of Computer-Assisted Medical Interventions at the German Cancer Research Center (Heidelberg, Germany). Sara is currently Postdoc at Università. Politecnica delle Marche, Department of Information Engineering (Ancona, Italy) and Affiliated Researcher at Istituto Italiano di Tecnologia.

Leonardo S. Mattos (B.Sc. 1998, M.Sc. 2003, Ph.D. 2007) is a Permanent Researcher and Head of the Biomedical Robotics Laboratory at the Italian Institute of Technology (IIT, Genoa). His research background ranges from robotic microsurgery and assistive human-machine interfaces to computer vision and micro-biomanipulation. Leonardo received his Ph.D. degree in electrical engineering from the North Carolina State University (NCSU, USA), where he worked as research assistant at the Center for Robotics and Intelligent Machines (CRIM) from 2002 until 2007. Since then he has been a researcher at the IIT's Department of Advanced Robotics. Dr. Mattos leads a team of 22 researchers and technicians at IIT and collaborates closely with other institutions, including hospitals and industry. So far, Leonardo has been a key player in securing over 8 million Euros for research through externally funded projects. He was the PI and coordinator of the European project μRALP – Micro-Technologies and Systems for Robot-Assisted Laser Phonomicrosurgery, and is currently the PI and coordinator of the translational project Robotic Microsurgery and of the TEEP-SLA project, which is dedicated to the creation of new interfaces and assistive systems for ALS patients. Dr. Mattos is also participating to the project Sistemi Cibernetici Collaborativi - Teleoperation as coordinator and work package leader. He is the author or co-author of more than 130 peer-reviewed publications and has been the chair and main organizer of several international scientific events, including the 9th Joint Workshop on New Technologies for Computer/Robot Assisted Surgery (CRAS 2019), the 4th Joint Workshop on New Technologies for Computer/Robot Assisted Surgery (CRAS 2014), the IEEE BioRob 2014 Workshop on Robotic Microsurgery and Image-Guided Surgical Interventions, and the IEEE BioRob 2012 Workshop on Robot-Assisted Laryngeal Microsurgery. Leonardo has also been part of scientific and program committees of major international conferences, serving as Associate Editor for ICRA, BioRob, MFI, and the Hamlyn Symposium. He is also a regular reviewer for several computer-assisted surgery and robotics conferences and journals.

Emanuele Frontoni (M.Sc. 2003, Ph.D. 2006) is Associate Professor in the Department of Information Engineering of Università Politecnica delle Marche. He received the doctoral degree in electronic engineering from the University of Ancona, Italy, in 2003. In the same year he joined the Dept. of Ingegneria Informatica, Gestionale e dell'Automazione (DIIGA) at the Università Politecnica delle Marche, as a Ph.D. student in "Intelligent Artificial Systems". He obtained his PhD in 2006 discussing a thesis on Vision Based Robotics. His research focuses on applying computer science, artificial intelligence and computer vision techniques to mobile robots and innovative IT applications. He is a member of IEEE and AI*IA, the Italian Association for Artificial Intelligence.

Elena De Momi (M.Sc. 2002, Ph.D. 2006) is Associate Professor in the Department of Electronic Information and Bioengineering (DEIB) of Politecnico di Milano. She is co-founder of the Neuroengineering and Medical Robotics Laboratory, in 2008, being responsible of the Medical Robotics section. IEEE Senior Member, she is currently Associate Editor of the Journal of Medical Robotics Research, of the International Journal of Advanced Robotic Systems, Frontiers in Robotics and AI and Medical & Biological Engineering & Computing. From 2016 she has been an Associated Editor of IEEE ICRA, IROS and BioRob, Area Chair of MICCAI and she is currently Publication Co-Chair of ICRA 2019. She is responsible for the lab course in Medical Robotics and of the course on Clinical Technology Assessment of the MSc degree in Biom. Eng. at Politecnico di Milano and she serves in the board committee of the PhD course in Bioengineering. Her academic interests include computer vision and image-processing, artificial intelligence, augmented reality and simulators, teleoperation, haptics, medical robotics, human robot interaction. She participated to several EU funded projects in the field of Surgical Robotics (ROBOCAST, ACTIVE and EuRoSurge, where she was PI for partner POLIMI). She is currently PI for POLIMI of the EDEN2020 project, aimed at developing a neurosurgery drug delivery system and of the ATLAS MSCA-ITN-2018-EJD, and coordinator of the MSCA-IF-2017 – Individual Fellowships. She has been evaluator and reviewer for the European Commission in FP6, FP7 and H2020.