DEEP LEARNING MODELS FOR INFORMATIVE-FRAME SELECTION IN LARYNGOSCOPIC VIDEOS

Supervisor:  Prof. Elena DE MOMI
Co-supervisors:  Sara MOCCIA, PhD
Leonardo S. MATTOS, PhD

Candidate:  Ilaria PATRINI, 873620
LARYNGEAL CANCER: a malignancy of the laryngeal tract belonging to the family of head and neck cancers.

From 95% to 98% of laryngeal cancers take the form of **squamous cell carcinoma (SCC)** [K. Markou et al., 2013] with the following main alterations in tissues morphology:

- Thickening and whitening of the epithelial layer.
- Dot-like vessels in the mucosa vascular tree.
- Longitudinal hypertrophic vessels.

- Late onset of symptoms;
- Alteration in the mucosa vascular tree at early stage.
LARYNGEAL CANCER: a malignancy of the laryngeal tract belonging to the family of head and neck cancers.

From 95% to 98% of laryngeal cancers take the form of **squamous cell carcinoma (SCC)** [K. Markou et al., 2013] with the following main alterations in tissues morphology:

- Thickening and whitening of the epithelial layer.
- Dot-like vessels in the mucosa vascular tree.
- Longitudinal hypertrophic vessels.

- Late onset of symptoms;
- Alteration in the mucosa vascular tree at early stage.

**Early-stage diagnosis** of SCC and **timely start of treatments** are crucial to **lower the mortality rate** and **preserve the quality of life**: maintenance of both the subdistricts anatomy and the vocal fold function [J. Unger et al., 2014].
Laryngeal endoscopy is the most widespread method to inspect the laryngeal tract.

**Narrow-band imaging (NBI)** endoscopy is currently considered as the gold standard procedure for non-invasive visual laryngeal early SCC screening.

Before hitting the tissue, the endoscope white-light illumination source is passed through a narrow-band filter that only allows the passage of wavelengths that match the absorption peak of hemoglobin:

- **415 nm** (blue light);
- **540 nm** (green light).
Maximum contrast between blood vessels and the surrounding tissue;
Optimal visualization of neo-angiogenic pattern of vasculature associated with early-stage malignant lesions;
Preventive screening [C. Piazza et al., 2012].
CLINICAL BACKGROUND – Open issues

- Maximum contrast between blood vessels and the surrounding tissue;
- Optimal visualization of neo-angiogenic pattern of vasculature associated with early-stage malignant lesions;
- Preventive screening [C. Piazza et al., 2012].

- The video reviewing process is a labour-intensive and time-consuming operation [M. Schuster et al., 2012];
- Presence of uninformative video portions.
Several machine-learning approaches have been proposed, which are mainly based on **HANDCRAFTED FEATURES** (e.g., features based on intensity or textural information).

<table>
<thead>
<tr>
<th>Method</th>
<th>Handcrafted-feature set</th>
<th>Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hai et al., 2006</td>
<td>Color, texture and motion displacement</td>
<td>Decision trees</td>
</tr>
<tr>
<td>Mehmood et al., 2014</td>
<td>Image moment, curvature and color histogram</td>
<td>Threshold-based approach</td>
</tr>
<tr>
<td>Schoemann et al., 2015</td>
<td>Keypoints</td>
<td>Threshold-based approach</td>
</tr>
</tbody>
</table>

**VIDEO CLUSTERING AND KEYFRAME EXTRACTION**

<table>
<thead>
<tr>
<th>Method</th>
<th>Handcrafted-feature set</th>
<th>Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Armin et al., 2015</td>
<td>Motion, intensity and image derivatives</td>
<td>Random forest</td>
</tr>
<tr>
<td>Perperidis et al., 2017</td>
<td>Texture</td>
<td>Gaussian mixture model</td>
</tr>
<tr>
<td>Moccia et al., 2018</td>
<td>Intensity, entropy, keypoints and texture</td>
<td>Support vector machines</td>
</tr>
</tbody>
</table>
Several machine-learning approaches have been proposed, which are mainly based on **HANDCRAFTED FEATURES** (e.g., features based on intensity or textural information).

### VIDEO CLUSTERING AND KEYFRAME EXTRACTION

<table>
<thead>
<tr>
<th>Method</th>
<th>Handcrafted-feature set</th>
<th>Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hai et al., 2006</td>
<td>Color, texture and motion displacement</td>
<td>Decision trees</td>
</tr>
<tr>
<td>Mehmoood et al., 2014</td>
<td>Image moment, curvature and color histogram</td>
<td>Threshold-based approach</td>
</tr>
<tr>
<td>Schoemann et al., 2015</td>
<td>Keypoints</td>
<td>Threshold-based approach</td>
</tr>
</tbody>
</table>

### UNINFORMATIVE FRAME REMOVAL

<table>
<thead>
<tr>
<th>Method</th>
<th>Handcrafted-feature set</th>
<th>Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Armin et al., 2015</td>
<td>Motion, intensity and image derivatives</td>
<td>Random forest</td>
</tr>
<tr>
<td>Perperidis et al., 2017</td>
<td>Texture</td>
<td>Gaussian mixture model</td>
</tr>
<tr>
<td>Moccia et al., 2018</td>
<td>Intensity, entropy, keypoints and texture</td>
<td>Support vector machines</td>
</tr>
</tbody>
</table>
Several machine-learning approaches have been proposed, which are mainly based on **HANDCRAFTED FEATURES** (e.g., features based on intensity or textural information).

### VIDEO CLUSTERING AND KEYFRAME EXTRACTION

<table>
<thead>
<tr>
<th>Method</th>
<th>Handcrafted-feature set</th>
<th>Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hai et al., 2006</td>
<td>Color, texture and motion displacement</td>
<td>Decision trees</td>
</tr>
<tr>
<td>Mehmood et al., 2014</td>
<td>Image moment, curvature and color histogram</td>
<td>Threshold-based approach</td>
</tr>
<tr>
<td>Schoemann et al., 2015</td>
<td>Keypoints</td>
<td>Threshold-based approach</td>
</tr>
</tbody>
</table>

### UNINFORMATIVE FRAME REMOVAL

<table>
<thead>
<tr>
<th>Method</th>
<th>Handcrafted-feature set</th>
<th>Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Armin et al., 2015</td>
<td>Motion, intensity and image derivaties</td>
<td>Random forest</td>
</tr>
<tr>
<td>Perperidis et al., 2017</td>
<td>Texture</td>
<td>Gaussian mixture model</td>
</tr>
<tr>
<td>Moccia et al., 2018</td>
<td>Intensity, entropy, keypoints and texture</td>
<td>Support vector machines</td>
</tr>
</tbody>
</table>
Several machine-learning approaches have been proposed, which are mainly based on **HANDCRAFTED FEATURES** (e.g., features based on intensity or textural information).

### VIDEO CLUSTERING AND KEYFRAME EXTRACTION

<table>
<thead>
<tr>
<th>Method</th>
<th>Handcrafted-feature set</th>
<th>Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hai et al., 2006</td>
<td>Color, texture and motion displacement</td>
<td>Decision trees</td>
</tr>
<tr>
<td>Mehmood et al., 2014</td>
<td>Image moment, curvature and color histogram</td>
<td>Threshold-based approach</td>
</tr>
<tr>
<td>Schoemann et al., 2015</td>
<td>Keypoints</td>
<td>Threshold-based approach</td>
</tr>
</tbody>
</table>

### UNINFORMATIVE FRAME REMOVAL

<table>
<thead>
<tr>
<th>Method</th>
<th>Handcrafted-feature set</th>
<th>Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Armin et al., 2015</td>
<td>Motion, intensity and image derivatives</td>
<td>Random forest</td>
</tr>
<tr>
<td>Perperidis et al., 2017</td>
<td>Texture</td>
<td>Gaussian mixture model</td>
</tr>
<tr>
<td>Moccia et al., 2018</td>
<td>Intensity, entropy, keypoints and texture</td>
<td>Support vector machines</td>
</tr>
</tbody>
</table>

Deep-learning algorithms may outperform standard learning approaches for image analysis [Goodfellow et al., 2016], by replacing handcrafted features with **LEARNED FEATURES**, which are automatically learned during a training process.
To develop a **deep learning-based strategy** to investigate if features learned with convolutional neural networks (CNNs), pre-trained on natural images, can be exploited for **informative-frame selection** in laryngoscopic videos.

The proposed solution extends the classification process to **four classes** to inform the clinician on the quality of images he is acquiring in real time.
## AIM OF THE WORK – Tested conditions

<table>
<thead>
<tr>
<th>Tested conditions</th>
<th>Feature extractor</th>
<th>Classifier</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>Six different CNNs</td>
<td>Support vector machines (SVMs)</td>
</tr>
<tr>
<td>C2</td>
<td>Best performing CNN of C1</td>
<td>Fine-tuned fully connected CNN</td>
</tr>
<tr>
<td>C3</td>
<td>Best performing CNN of C1 trained from scratch</td>
<td>Best performing CNN of C1 trained from scratch</td>
</tr>
</tbody>
</table>
METHODS – Workflow

Methods

- CNN training
- Learned-feature extraction
- Fine tuning
- Frame classification

- Informative frame (I)
- Blurred frame (B)
- Frame with saliva or specular reflections (S)
- Underexposed frame (U)

- NBI video frames

- Training frames
- Testing frames
METHODS – C1: Workflow

TRANSFER LEARNING AND SVMs-BASED CLASSIFICATION

C1

NBI video frames

Training frames

Testing frames

CNN training

Learned-feature extraction

Learned-feature extraction

Fine tuning

Frame classification

Informative frame (I)

Blurred frame (B)

Frame with saliva or specular reflections (S)

Underexposed frame (U)
METHODS – C1: Transfer learning and SVMs-based classification

FEATURE EXTRACTION

- Inception V4
- Inception – ResNet V2
- ResNet V1 101
- ResNet V1 152
- ResNet V2 152
- VGG 16

- Best performing CNNs in the context of the Image Large Scale Visual Recognition Challenge (ILSVRC) [http://www.imagenet.org/challenges/LSVRC/].
**METHODS – C1: Transfer learning and SVMs-based classification**

**FEATURE EXTRACTION**

- Inception V4
- Inception – ResNet V2
- ResNet V1 101
- ResNet V1 152
- ResNet V2 152
- VGG 16

- Pre-trained on the **ImageNet dataset**, released in the context of ILSVRC 2012 [http://www.image-net.org/];
- Best performing **CNNs** in the context of the Image Large Scale Visual Recognition Challenge (ILSVRC) [http://www.imagenet.org/challenges/LSVRC/].

**SUPPORT VECTOR MACHINES – based classification**

- To be consistent with [Moccia et al., 2018];
- Robust to noise in training data;
- Kernel-trick prevents parameter proliferation;
- **Multi-class SVM classification** with one-vs-rest scheme: when one class was considered positive, the remaining ones were considered negative;
- SVMs hyperparameters tuning retrieved via grid-search and cross-validation.
METHODS – C2: Workflow

TRANSFER LEARNING AND FINE-TUNED CNN-BASED CLASSIFICATION

- CNN training
- Fine tuning
- Learned-feature extraction
- Frame classification

NBI video frames

- Training frames
- Testing frames

Informative frame (I)
Blurred frame (B)
Frame with saliva or specular reflections (S)
Underexposed frame (U)
FEATURE EXTRACTION

Best performing learned-feature set obtained in C1.

FINE-TUNED CNN-based classification

- First layers contain more generic features 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 [I. Goodfellow et al., 2016].
**METHODS – C3: Workflow**

C3

**TRAINING FROM SCRATCH**

- **NBI video frames**
  - Training frames
  - Testing frames

  - CNN training
  - Learned-feature extraction
  - Fine tuning
  - Frame classification

**Informative frame (I)**
- Blurred frame (B)
- Frame with saliva or specular reflections (S)
- Underexposed frame (U)
TRAINING FROM SCRATCH

To prove and quantify the transfer learning effect:

- The best performing architecture of C1 was trained from scratch;
- All the stored knowledge or weights were initialized and learned only from the task under analysis;
- Weights initialization performed with Glorot initialization.
18 NBI endoscopic videos (25 fps) of 18 patients affected by SCC.

For each video, frames were extracted randomly and presented to evaluators to be annotated in four classes (I, B, S, U).

720 frames equally divided in the four different classes:

- 180 x I
- 180 x B
- 180 x S
- 180 x U

B: Blurred frame; I: Informative frame; S: Frame with saliva or specular reflections; U: Underexposed frame.
18 NBI endoscopic videos (25 fps) of 18 patients affected by SCC.

For each video, frames were extracted randomly and presented to evaluators to be annotated in four classes (I, B, S, U).

720 frames equally divided in the four different classes:

- **180 x I**
- **180 x B**
- **180 x S**
- **180 x U**

One video sequence (18.24 s) was fully labeled: 456 frames

**B**: Blurred frame; **I**: Informative frame; **S**: Frame with saliva or specular reflections; **U**: Underexposed frame.
All the videos were rearranged in three sub-folders: data separation in the folds is performed both at class- and patient-level.

240 frames for each fold (60 frames per class)

3-FOLD CROSS VALIDATION – NBI-InfFrames dataset

1) ▪ 720 frames is fairly small set;
2) ▪ 3-fold average training is performed to make up for the small number of frames and to improve the stability of the results.
3) ▪ Training: 480 frames
      ▪ Test: 240 frames
All the videos were rearranged in **three sub-folders**: data separation in the folds is performed both at class- and patient-level.

240 frames for each fold (60 frames per class)

**Complete video sequence**

- All the frames belonging to the three folds (excluding the ones of the fully labeled video) were used for training.

| Training: 663 frames | Test: 456 frames |
EVALUATION METRICS

- $\text{Rec}_{\text{class}_j} = \frac{TP_j}{TP_j + FN_j}$

- $\text{Prec}_{\text{class}_j} = \frac{TP_j}{TP_j + FP_j}$

- $F1_{\text{class}_j} = 2 \frac{\text{Prec}_{\text{class}_j} \times \text{Rec}_{\text{class}_j}}{\text{Prec}_{\text{class}_j} + \text{Rec}_{\text{class}_j}}$

- Area (AUC) under the receiver operating characteristic curve (ROC)

- Class-specific ROC analysis for the best performing CNN
## RESULTS – Tested conditions

<table>
<thead>
<tr>
<th>Tested conditions</th>
<th>Feature extractor</th>
<th>Classifier</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>Six different CNNs</td>
<td>Support vector machines (SVMs)</td>
</tr>
<tr>
<td>C2</td>
<td>Best performing CNN of C1</td>
<td>Fine-tuned fully connected CNN</td>
</tr>
<tr>
<td>C3</td>
<td>Best performing CNN of C1 trained from scratch</td>
<td>Best performing CNN of C1 trained from scratch</td>
</tr>
</tbody>
</table>
RESULTS – C1: Transfer learning and SVM-based classification

- **B**: Blurred frame; **I**: Informative frame; **S**: Frame with saliva or specular reflections; **U**: Underexposed frame.
RESULTS – C1: Transfer learning and SVM-based classification

B: Blurred frame; I: Informative frame; S: Frame with saliva or specular reflections; U: Underexposed frame.
RESULTS – C2: Transfer learning and fine-tuned CNN-based classification

B: Blurred frame; I: Informative frame; S: Frame with saliva or specular reflections; U: Underexposed frame.
RESULTS – C2: Transfer learning and fine-tuned CNN-based classification

C2

VGG 16 feature-extraction
SVMs-based classification

VGG 16 feature-extraction
Fine-tuned VGG 16-based classification

B: Blurred frame; I: Informative frame; S: Frame with saliva or specular reflections; U: Underexposed frame.
RESULTS – C2: Transfer learning and fine-tuned CNN-based classification

C2

Complete video sequence

Automatic removal of 146 uninformative frames out of 154.

B: Blurred frame; I: Informative frame; S: Frame with saliva or specular reflections; U: Underexposed frame.
RESULTS – C3: Frame classification with training from scratch

VGG 16-based classification trained from scratch

<table>
<thead>
<tr>
<th></th>
<th>B</th>
<th>I</th>
<th>S</th>
<th>U</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>.87</td>
<td>.01</td>
<td>.09</td>
<td>.03</td>
</tr>
<tr>
<td>I</td>
<td>.00</td>
<td>.78</td>
<td>.21</td>
<td>.01</td>
</tr>
<tr>
<td>S</td>
<td>.11</td>
<td>.12</td>
<td>.76</td>
<td>.01</td>
</tr>
<tr>
<td>U</td>
<td>.05</td>
<td>.00</td>
<td>.04</td>
<td>.91</td>
</tr>
</tbody>
</table>

Transfer learning and fine-tuned VGG 16-based classification

<table>
<thead>
<tr>
<th></th>
<th>B</th>
<th>I</th>
<th>S</th>
<th>U</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>.93</td>
<td>.00</td>
<td>.05</td>
<td>.02</td>
</tr>
<tr>
<td>I</td>
<td>.00</td>
<td>.98</td>
<td>.00</td>
<td>.02</td>
</tr>
<tr>
<td>S</td>
<td>.09</td>
<td>.01</td>
<td>.86</td>
<td>.04</td>
</tr>
<tr>
<td>U</td>
<td>.00</td>
<td>.04</td>
<td>.02</td>
<td>.94</td>
</tr>
</tbody>
</table>

B: Blurred frame; I: Informative frame; S: Frame with saliva or specular reflections; U: Underexposed frame.
CONCLUSIONS

- Approaches outperformed the state-of-the-art. Results 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;

  Highlighting the potentiality of Deep Learning

- **First attempt** to use DL-based strategy to informative frame selection in endoscopic videos;

- Results are expected to provide major contributions towards lowering the degree of manual intervention required by computer-assisted systems intended to analyze and summarize the endoscopic video content.

IMPACT

- The methodology is expected to impact positively on the detection of early-stage tumors, being able to perform a more accurate screening.
CONCLUSIONS (2/2)

LIMITATION

- Dimension of the evaluation dataset (but high number of patients).

FUTURE WORK

- Enlarge the dataset exploiting Generative Adversarial Networks (GAN) for data augmentation;
- Investigate contentious-learning strategies;
- Application to other body districts.

WORKSHOP AND PUBLICATION

- Presented at Computer/Robot Assisted Surgery (CRAS) workshop (March, 2019);
- Published on Medical & Biological Engineering & Computing (March, 2020).
THANK YOU
FOR YOUR KIND ATTENTION!

ilaria.patrini@mail.polimi.it