Learning-based classification of informative endoscopic frames with applications in laryngoscopy

Gabriele Omodeo Vanone, 837090

Advisor: Elena De Momi, PhD
Co-advisors: Dott. Ing. Sara Moccia
Veronica Penza, PhD

A.Y. 2016-2017
OUTLINE

- Clinical Background
- State of the Art
- Aim of the Thesis
- Materials
- Methods, Results and Discussion
- Conclusion and Future Work
OUTLINE

• Clinical Background
• State of the Art
• Aim of the Thesis
• Materials
• Methods, Results and Discussion
• Conclusion and Future Work
CLINICAL BACKGROUND
Laryngeal cancer data

European distribution of estimated Age-Standardized incidence Rates (ASR) per 100,000 individuals for laryngeal cancer in both sexes, assessed in 2012
[K. Markou et al., 2013]

In terms of histopathology, 95% to 98% of laryngeal cancer is squamous cell carcinoma (SCC)
[J. Ferlay et al., 2013]
Narrow-Band Imaging (NBI) Endoscopy

• Optical technique that allows optimal visualization of neo-angiogenic changes associated with early-stage malignant lesions

• Enables preventive screening of patients at risk

[C. Piazza et al., 2012]
**Narrow-Band Imaging (NBI) Endoscopy**

- Optical technique that allows optimal visualization of neo-angiogenic changes associated with early-stage malignant lesions
- Enables preventive screening of patients at risk

[C. Piazza et al., 2012]

**Drawbacks**

- Narrow field of view
- Hard to navigate
- Time-consuming & focus-intensive video review

[M. Schuster et al., 2012]
CLINICAL BACKGROUND
Computer-Aided tools

Techniques to relieve clinicians’ workload available for the laryngeal context:

- Image stitching

[M. Schuster et al., 2012]
[S. Moccia et al., 2016]
Techniques to **relieve clinicians’ workload** available for the laryngeal context:

- Image stitching
- Computer-Aided Diagnosis (CAD)

[A. Verikar et al., 2006]
[S. Moccia et al., 2017]
Techniques to relieve clinicians’ workload:

- Image stitching
- Computer-aided diagnosis (CAD)
- …

They all need frame selection:

- To reduce input data size
- To provide for better outputs
OUTLINE

• Clinical Background
• **State of the Art**
• Aim of the Thesis
• Materials
• Methods, Results and Discussion
• Conclusion and Future Work
## STATE OF THE ART

### Frame selection in endoscopy

<table>
<thead>
<tr>
<th>Uniform frame sampling</th>
<th>Blur amount metrics</th>
<th>Color-space features</th>
<th>Keypoint transform-based features</th>
</tr>
</thead>
</table>
| - Reduces input data size | - Single parameter thresholding  
  - Blur as the primary source for lost information | - Single parameter thresholding  
  - Color histogram is used to identify different image content | - Keypoint transform is a high-end computer vision tool  
  - Useful for image stitching oriented frame selection |

[F. Crete et al., 2007]  
[S. Moccia et al., 2016]  
[R. Estrada et al., 2011]  
[V. Hai et al., 2015]  
[I. Mehmood et al., 2014]  
[M. Lux et al., 2010]  
[T. Bergen et al., 2013]
OUTLINE

- Clinical Background
- State of the Art
- **Aim of the Thesis**
- Materials
- Methods, Results and Discussion
- Conclusion and Future Work
Machine Learning for Automatic frame classification and subsequent selection.
• 18 NBI laryngoscopic videos from 18 different patients, (Olympus NBI system, provided by San Martino Hospital, Genova)

• 4 classes considered:
  a) informative frames (I)
  b) blurred frames (B)
  c) frames with saliva or specular reflections (S)
  d) underexposed frames (U)

• 720 frames labelled build the ground-truth database

Examples of class types
OUTLINE

- Clinical Background
- State of the Art
- Aim of the Thesis
- Materials
- **Methods, Results and Discussion**
- Conclusion and Future Work
METHODS
Flowchart of the proposed approach

Sequence of video frames → Feature extraction → Frame classification using Support Vector Machines (SVM)

i.e. Each frame becomes a point in a N-dimensional space

N = number of features

[C. J. Burges, 1998]
[Y. Lin et al., 2011]
[C. Vercellis, 2011]
### METHODS

Descriptors and Feature vector

<table>
<thead>
<tr>
<th>Numerical features</th>
<th>Vector components</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blind/referenceless image spatial quality evaluator (BRISQUE)</td>
<td>1</td>
</tr>
<tr>
<td>Sobel-Tenengrad focus evaluation function score (TEN)</td>
<td>3</td>
</tr>
<tr>
<td>Global variance of the image Laplacian</td>
<td>3</td>
</tr>
<tr>
<td>Local variance of the image intensity</td>
<td>3</td>
</tr>
<tr>
<td>Global variance of the image intensity</td>
<td>3</td>
</tr>
<tr>
<td>Shannon entropy</td>
<td>3</td>
</tr>
<tr>
<td>Histogram</td>
<td>3</td>
</tr>
<tr>
<td>Number of detected keypoints</td>
<td>1</td>
</tr>
</tbody>
</table>

**Total** 20
## METHODS

### Descriptors and Feature vector

<table>
<thead>
<tr>
<th>Numerical features</th>
<th>V.C.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blind/referenceless image spatial quality evaluator (BRISQUE)</td>
<td>1</td>
</tr>
<tr>
<td>Sobel-Tenengrad focus evaluation function score (TEN)</td>
<td>3</td>
</tr>
<tr>
<td>Global variance of the image Laplacian</td>
<td>3</td>
</tr>
<tr>
<td>Local variance of the image intensity</td>
<td>3</td>
</tr>
<tr>
<td>Global variance of the image intensity</td>
<td>3</td>
</tr>
<tr>
<td>Shannon entropy</td>
<td>3</td>
</tr>
<tr>
<td>Histogram</td>
<td>3</td>
</tr>
<tr>
<td>Number of detected keypoints</td>
<td>1</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>20</strong></td>
</tr>
</tbody>
</table>

Asymmetric generalized Gaussian distribution

\[
f(x, \alpha, \sigma_1^2, \sigma_2^2) = \begin{cases} \frac{\alpha}{(\beta_l + \beta_r)\gamma \left(\frac{1}{\alpha}\right)} \exp \left(-\left(\frac{x}{\beta_l}\right)^\alpha\right) & x < 0 \\ \frac{\alpha}{(\beta_l + \beta_r)\gamma \left(\frac{1}{\alpha}\right)} \exp \left(-\left(\frac{x}{\beta_l}\right)^\alpha\right) & x \leq 0 \end{cases}
\]

\(\alpha\) is a shape parameter, \(\sigma_1, \sigma_2\) are asymmetric scale parameters, \(\gamma\) and \(\beta_l, \beta_r\) depend on the aforementioned parameters

[A. Mittal et al., 2012]
METHODS
Descriptors and Feature vector

<table>
<thead>
<tr>
<th>Numerical features</th>
<th>V.C.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blind/referenceless image spatial quality evaluator (BRISQUE)</td>
<td>1</td>
</tr>
<tr>
<td>Sobel-Tenengrad focus evaluation function score (TEN)</td>
<td>3</td>
</tr>
<tr>
<td>Global variance of the image Laplacian</td>
<td>3</td>
</tr>
<tr>
<td>Local variance of the image intensity</td>
<td>3</td>
</tr>
<tr>
<td>Global variance of the image intensity</td>
<td>3</td>
</tr>
<tr>
<td>Shannon entropy</td>
<td>3</td>
</tr>
<tr>
<td>Histogram</td>
<td>3</td>
</tr>
<tr>
<td>Number of detected keypoints</td>
<td>1</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>20</strong></td>
</tr>
</tbody>
</table>

Sobel-Tenengrad function

\[
TEN = \sum_{m}^{M} \sum_{n}^{N} [S(m, n)]^2
\]

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

\(G_x, G_y\) are the convolution of the image \(I\), of size \(M \times N\) pixels, with the Sobel’s kernel in horizontal and vertical direction, respectively.

Computed for each image channel in the RGB-space

[J.M. Mateos-Pérez et al., 2012]
## METHODS

### Descriptors and Feature vector

<table>
<thead>
<tr>
<th>Numerical features</th>
<th>V.C.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blind/referenceless image spatial quality evaluator (BRISQUE)</td>
<td>1</td>
</tr>
<tr>
<td>Sobel-Tenengrad focus evaluation function score (TEN)</td>
<td>3</td>
</tr>
<tr>
<td>Global variance of the image Laplacian</td>
<td>3</td>
</tr>
<tr>
<td>Local variance of the image intensity</td>
<td>3</td>
</tr>
<tr>
<td>Global variance of the image intensity</td>
<td>3</td>
</tr>
<tr>
<td>Shannon entropy</td>
<td>3</td>
</tr>
<tr>
<td>Histogram</td>
<td>3</td>
</tr>
<tr>
<td>Number of detected keypoints</td>
<td>1</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>20</strong></td>
</tr>
</tbody>
</table>

Global variances are computed for all pixels of the image in bulk.

Local variances are average results over a neighborhood of pixels around each image pixel, averaged over all image pixels.

Computed for each image channel in the RGB-space.

[J.M. Mateos-Pérez et al., 2012]
## METHODS
### Descriptors and Feature vector

<table>
<thead>
<tr>
<th>Numerical features</th>
<th>V.C.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blind/referenceless image spatial quality evaluator (BRISQUE)</td>
<td>1</td>
</tr>
<tr>
<td>Sobel-Tenengrad focus evaluation function score (TEN)</td>
<td>3</td>
</tr>
<tr>
<td>Global variance of the image Laplacian</td>
<td>3</td>
</tr>
<tr>
<td>Local variance of the image intensity</td>
<td>3</td>
</tr>
<tr>
<td>Global variance of the image intensity</td>
<td>3</td>
</tr>
<tr>
<td>Shannon entropy</td>
<td>3</td>
</tr>
<tr>
<td>Histogram</td>
<td>3</td>
</tr>
<tr>
<td>Number of detected keypoints</td>
<td>1</td>
</tr>
</tbody>
</table>

Total 20

256 bins image color histogram is described by the three quartiles

[M.K. Bashar et al., 2008]
### Descriptors and Feature vector

<table>
<thead>
<tr>
<th>Numerical features</th>
<th>V.C.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blind/referenceless image spatial quality evaluator (BRISQUE)</td>
<td>1</td>
</tr>
<tr>
<td>Sobel-Tenengrad focus evaluation function score (TEN)</td>
<td>3</td>
</tr>
<tr>
<td>Global variance of the image Laplacian</td>
<td>3</td>
</tr>
<tr>
<td>Local variance of the image intensity</td>
<td>3</td>
</tr>
<tr>
<td>Global variance of the image intensity</td>
<td>3</td>
</tr>
<tr>
<td>Shannon entropy</td>
<td>3</td>
</tr>
<tr>
<td>Histogram</td>
<td>3</td>
</tr>
<tr>
<td><strong>Number of detected keypoints</strong></td>
<td>1</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>20</strong></td>
</tr>
</tbody>
</table>

Number of keypoints detected using the oriented fast and rotated brief keypoint detector (ORB)

[E. Rublee et al., 2011]
**METHODS**

Support Vector Machines (SVM)

**Classification method:** Radial basis function SVM
- SVM are robust to noise on training data
- Kernel-trick prevents parameters proliferation

**SVM parameter tuning:**
Automatic using grid search and 10-fold cross-validation
3-fold SVM training and averaging

- 720 frames is fairly small set for this application
- 3-fold dataset split and averaged training improves stability of the results  
  [S. Moccia et al., 2017]

3 sets of frames are used to train the SVM. Each set is made of 6 videos. Each set is made of 240 frames.

I: Informative
S: With Saliva or Specular Reflections
B: Blurred
U: Underexposed

I: 60  
S: 60  
B: 60  
U: 60
3-fold SVM training and averaging

- 720 frames is fairly small set for this application
- 3-fold dataset split and averaged training improves stability of the results

3 sets of frames are used to train the SVM. Each set is made of 6 videos. Each set is made of 240 frames.

Trifold training is performed to make up for the Small number of frames.
3-fold SVM training and averaging

- 720 frames is fairly small set for this application
- 3-fold dataset split and averaged training improves stability of the results

3 sets of frames are used to train the SVM. Each set is made of 6 videos. Each set is made of 240 frames.

**Trifold training is performed to make up for the Small number of frames.**
3-fold SVM training and averaging

- 720 frames is fairly small set for this application
- 3-fold dataset split and averaged training improves stability of the results

3 sets of frames are used to train the SVM. Each set is made of 6 videos. Each set is made of 240 frames.

**Trifold training is performed to make up for the Small number of frames.**
METHODS
Performance Evaluation Metrics

Evaluation metrics:

\[ \text{Rec}_{\text{class},j} = \frac{TP_j}{TP_j + FN_j} \quad j \in [I, B, S, U] \]

\( TP \) = True Positive \quad \( FN \) = False Negative

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

- Gold standard: human classified labels
- Tri-fold cross-validation
- Tested against 3 state of the art competitor methods
- Wilcoxon signed-rank test (0.05 significance level)
METHODS
Performance Evaluation Metrics

Evaluation metrics:

\[ \text{Rec}_{\text{class},j} = \frac{TP_j}{TP_j + FN_j} \quad j \in [I, B, S, U] \]

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

Comparison methods:

• Crete et al. blur assessment metric
• Bashar et al. color-space feature metric
• Bergen et al. keypoint-transform based metric

Gold standard: human classified labels
Tri-fold cross-validation
Tested against 3 state of the art competitor methods
Wilcoxon signed-rank test (0.05 significance level)

[T. Bergen et al., 2013]
[F. Crete et al., 2007]
[M.K. Bashar et al., 2008]
Comparison with State of the Art:

- 3 competitor methods selected
  - A blur metric
  - A color-space feature metric
  - A keypoint-based metric
- Applied to the same ground-truth dataset
- State of the art features used to train SVM with the same protocol as before
- Comparison made on SVM classification performances
RESULTS AND DISCUSSION
Performance and State of the Art comparison

Average Classification performance of the proposed method

Recall scoring: comparison with the State of the Art

<table>
<thead>
<tr>
<th>True label</th>
<th>I: Informative frames</th>
<th>B: Blurred frames</th>
<th>S: Frames with saliva or specular reflections</th>
<th>U: Underexposed frames</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted label</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>.91</td>
<td>.00</td>
<td>.03</td>
<td>.06</td>
</tr>
<tr>
<td>B</td>
<td>.00</td>
<td>.83</td>
<td>.06</td>
<td>.11</td>
</tr>
<tr>
<td>S</td>
<td>.07</td>
<td>.22</td>
<td>.62</td>
<td>.09</td>
</tr>
<tr>
<td>U</td>
<td>.02</td>
<td>.05</td>
<td>.08</td>
<td>.85</td>
</tr>
</tbody>
</table>

Boxplot showing the comparison of Recall scoring with the State of the Art for different datasets (Bashar, Crete, Bergen, Proposed)
RESULTS AND DISCUSSION
Performance and State of the Art comparison

ROC curves for the proposed method and the three comparison state of the art methods

Stars indicate significative differences
METHODS
Realistic Application: image stitching

Application mock-up:
• Image stitching protocol based on the exploratory work by Schuster et al.
• Use of a commercial image stitching software called Autostitch
• 3 protocols, with and without frame selection and with frame selection and video segmentation based on Schoeffman et al.

[M. Schuster et al., 2012]
[M. Schoeffman et al., 2015]
METHODS
Realistic Application: image stitching

Without Frame Selection (NO FS)
- Start
- All video Frames
- Autostitch
- Evaluation
- End

With Frame Selection (FS)
- Start
- All video Frames
- Frame selection
- Autostitch
- Evaluation
- End

With Frame Selection and Video Segmentation (FS + VS)
- Start
- All video Frames
- Frame selection
- Video segmentation
- Autostitch
- Evaluation
- End

[M. Schoeffman et al., 2015]
RESULTS AND DISCUSSION
Realistic Application: image stitching

Execution time of the image stitching protocols
How visually good are the stitching results?

Quality score in [0, 5].

0 = Autostitch took longer than 30 min
1 = very bad output
5 = very good output

Based on geometrical considerations.
OUTLINE

• Clinical Background
• State of the Art
• Aim of the Thesis
• Materials
• Methods, Results and Discussion
• Conclusion and Future Work
CONCLUSION AND FUTURE WORK

Conclusion:
• Highly accurate classification method overcoming state of the art competition
• Robust to inter-patient variability and parameter independent

Limitation:
• Trained on a limited number of images (but on a large number of patients)

Future work:
• Application to Computer Aided Diagnosis
• Application to other body districts
THANKS FOR YOUR CURIOSITY