Informative frame classification of endoscopic videos using convolutional neural networks and hidden Markov models

Joost van der Putten[1], Jeroen de Groof [2], Fons van der Sommen [1], Maarten Struyvenberg [2], Svitlana Zinger [1], Wouter Curvers [3], Erik J. Schoon [3], Jacques .J.G.H.M. Bergman [2], Peter H.N. de With [1]

[1]Eindhoven University of Technology, Video Coding and Architectures Research group, the Netherlands
[3]Academic Medical Center, Amsterdam, the Netherlands
[4]Catharina Hospital Eindhoven, the Netherlands

Abstract

The goal of endoscopic analysis is to find abnormal lesions and determine further therapy from the obtained information. However, the procedure produces a variety of non-informative frames and lesions can be missed due to poor video quality. Especially when analyzing entire endoscopic videos made by non-expert endoscopists, informative frame classification is crucial to e.g. video quality grading. This work concentrates on the design of an automated indication of informativeness of video frames. We propose an algorithm consisting of state-of-the-art deep learning techniques, to initialize frame-based classification, followed by a hidden Markov model to incorporate temporal information and control consistent decision making. Results from the performed experiments show that the proposed model improves on the state-of-the-art with an F1-score of 91%, and a substantial increase in sensitivity of 10%. Additionally, the algorithm is capable of processing 261 frames per second.

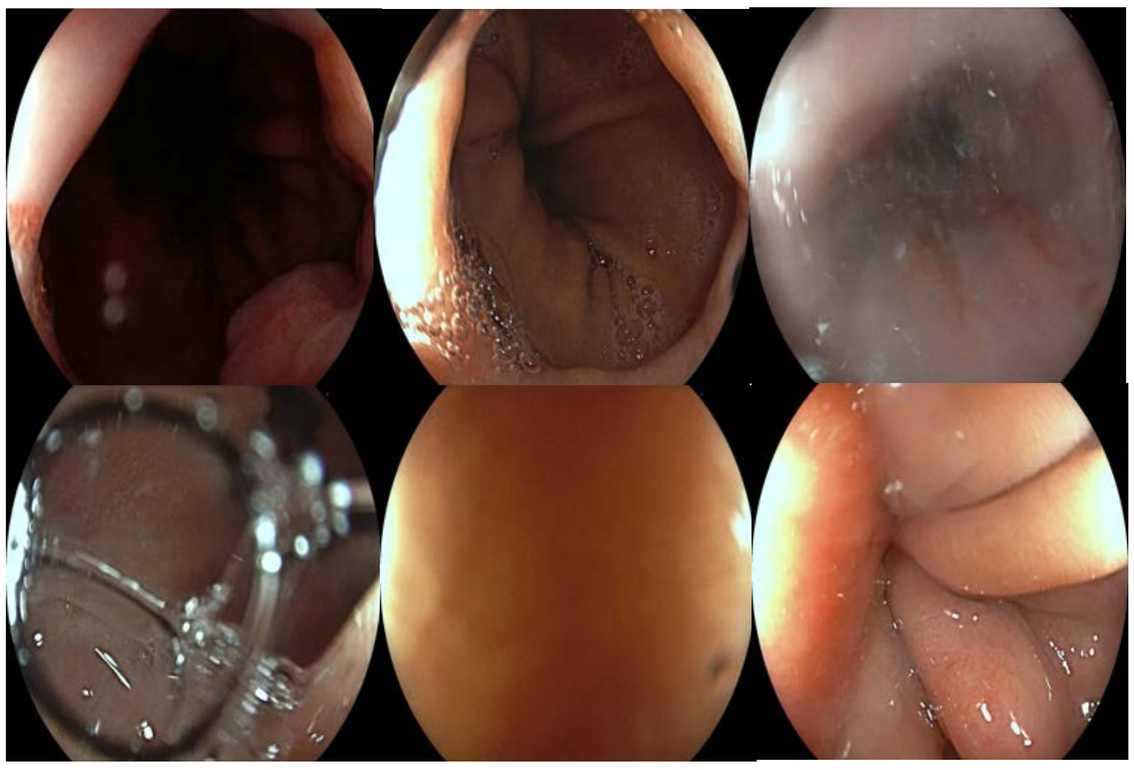

Figure 2: Examples of non-informative frames. (a) under-illumination (b) over illumination (c) Motion blur (d) bubbles (e) out of focus (f) contractions

1. Goals of the study

- Develop model to assess informativeness of endoscopic video frames
- Ask the question: what is informativeness in endoscopic video

2. Data Set

- 22,163 frames from 86 endoscopic pullback videos (5 fps)
- 64 video for training and 22 videos for testing
- Frames manually labeled by an expert endoscopist
- All frames were resized to 256X256 pixels for computational efficiency

3.1 Methods – CNN architecture

- Fully convolutional CNN with ResNet18 as base model
- Additional fully connected layer for possibility of feature extraction

Network choices	Details
Base model	Resnet 18
Optimizer	Adam + AMS-grad
Scheduler	Cyclic cosine annealing
Regularization	Batch-norm
Weight decay	10^{-5}
Data augmentation	Rotation, flipping, color permutations, random shearing and translation.

Table 2: Network details

3.2 Methods – Hidden Markov model

- A Hidden Markov Model (HMM) is employed to incorporate temporal information.
- The HMM is defined by:
  $Q = \{q_1, q_2, \ldots, q_N\}$: set of states.
  V= $\{v_1, v_2, \ldots, v_M\}$: set of possible observations.
  $A = \{a_{ij} | a_{ij} = P(s_{t+1} = q_j | s_t = q_i)\}$: *state transition* probability, where a_{ij} is the probability of transitioning from state q_i to state q_j.
  $B = \{b_j(k) | b_j(k)) = P(v_k | s_t = q_j)\}$: *state observation* probability, where $b_j(k)$ is the probability of output symbol v_k at state s_t.
  $\pi = \{\pi_i | \pi_i = P(s_1 = q_i)\}$: *state initialization* probability.
  $\lambda = \{A, B, \pi\}$: complete parameter set of the model
- In our case, $N = 2$ (informative, non-informative) and $M = 1$ (output of the classifier)
- Parameters $\lambda = \{A, B, \pi\}$ are determined on the training set using the Viterbi algorithm.

4. results

Model	Accuracy	Sensitivity	Specificity	F1	Fps
Dongen *et al.*	0.85	0.86	0.84	0.85	3
CNN-only (Ours)	0.94	0.76	0.98	0.85	260
CNN-HMM (Ours)	0.94	0.86	0.96	0.91	261

Table 2: Classification results

- Both of our results outperform state-of-the-art of Dongen *et al.*
- The CNN-HMM model increases sensitivity significantly with very little loss in specificity.
- Processing time the proposed model is easily real-time.

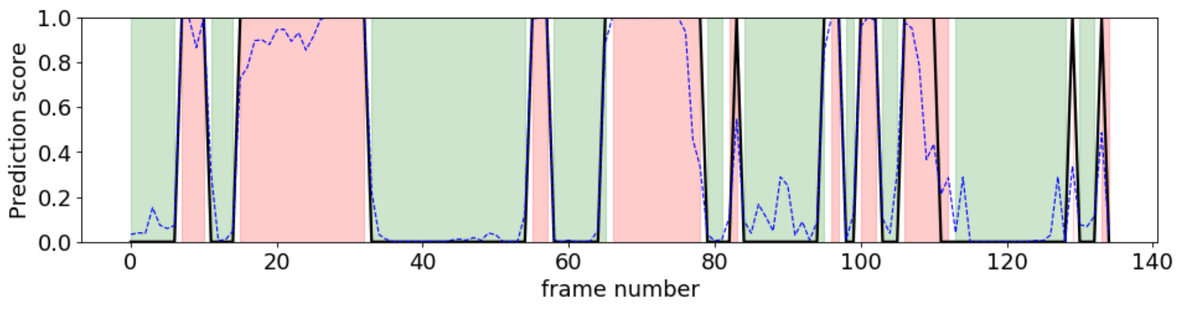

Figure 5: Video sequence indicating informativeness. Green and red backgrounds indicate the ground truth informative and non-informative frames, respectively. The dashed blue line refers to the output of the CNN and the black line indicates the response after modelling the output sequence with a hidden Markov model.

5. Discussion

- Images where output of CNN was closest to 0.5 i.e. most uncertain samples.