ECOLE
POLYTECHNIQUE
DE BRUXELLES

MEMO-H504 - Mémoire de fin d'études en Informatique

Development of an automatic video classification solution for Esophagogastroduodenoscopies

Popularising Article

Stefano Donne

30th May 2022

1 Context

First, what's an Esophagogastroduodenoscopy (EGD) ? It is a very common medical exam that allows the physician to explore the upper gastro-intestinal (GI) tract of his patient (fig 1) with the help of an endoscope. In order to make a diagnosis, the physician will visually assess the state of various sites of the upper GI tract and take pictures. He can also remove some tissues samples with a biopsy forceps for further analysis. This exam is also often referred as a 'gastroscopy'.

This exam is the gold standard for the diagnosis of many diseases, such as; *oesophagitis*, *gastritis*, *ulcers* or even *gastric cancer*. In average, 409.391 EGDs [1] are carried out each year in Belgium.

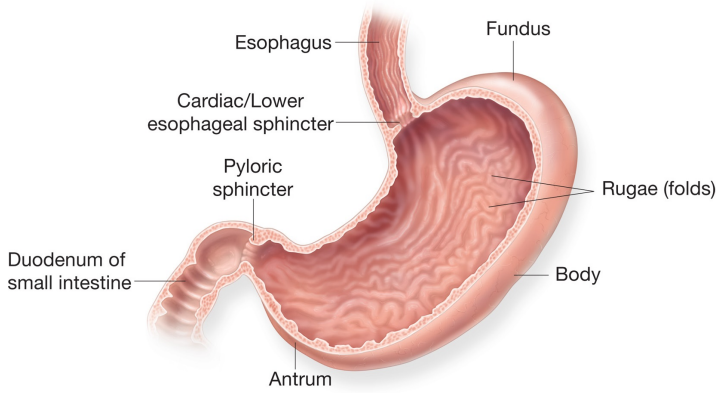

Figure 1: Upper GI tract [2]

The goal of the present work is to develop a solution that is capable of automatically classifying the frames of an EGD video, into the corresponding anatomical sites. Examples of EGD frames are given in fig 2. But what is the main motivation behind ? Such work could be the basis of the development of other solutions, such as a visual pollution detector. Visual pollution is often present in EGDs and can lead to a biased diagnosis or longer exams that discomfort the patients.

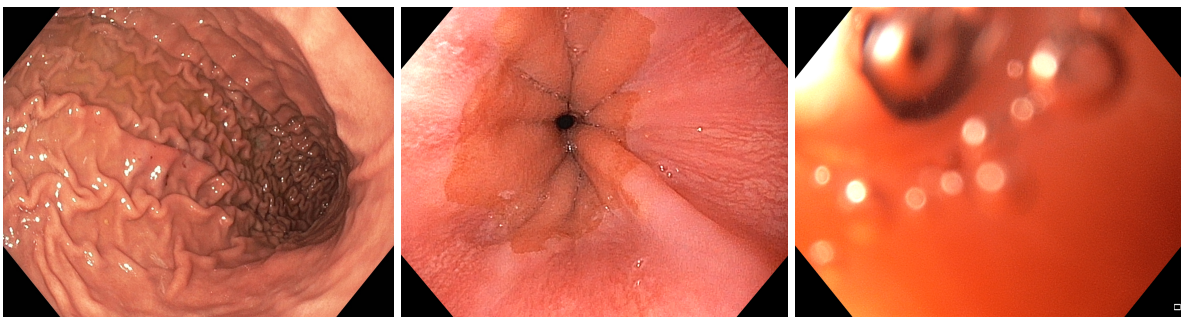

Figure 2: EGD frames examples

2 Solution

The chosen method is implemented in two steps :

1. Develop an EGD image classification method
2. Extrapolate this method for video classification

The image classification method relies on the use of Convolutional Neural Networks (CNN). These are architectures of Deep Neural Networks (DNN) that possess one or multiple convolutional layers in their network. These CNNs are a powerful tool and the current standard for image classification. The goal is to train a CNN to classify frames coming from EGD videos.

Some CNNs are selected from previous similar works and trained on a labeled data set. The data set is a collection of frames that have been manually extracted and labeled from EGD videos. The EGDs have been recorded at the CHU St-Pierre hospital between 2021 and 2022. The resulting data set is made of 3744 images, split in 7 sites (see fig 3).

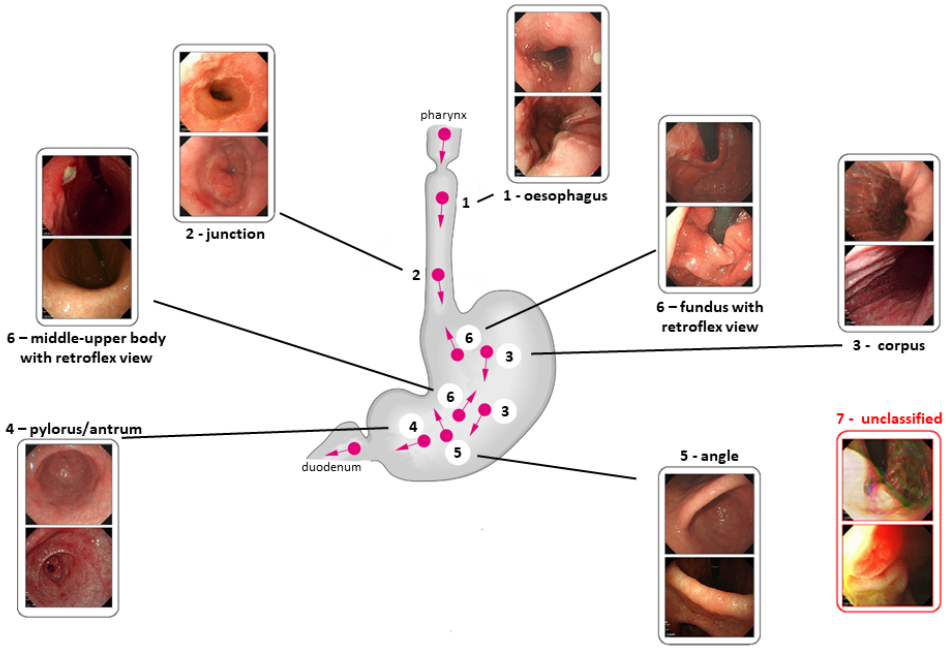

Figure 3: Data Set categories (modified from [3])

Then, the best CNN model found is selected for video application. The classification capabilities of the model are extrapolated for video usage in the following way : the CNN classifies 5 frames per second, the content of all 5 classification is then averaged. Therefore, the anatomical site detected at a frame i depends of the CNN outputs made in the previous second of video. It gives a stable throughput to the predictions. For a given frame, the CNN outputs a prediction value for each site (7), an example of the average made for videos is shown in fig 4.

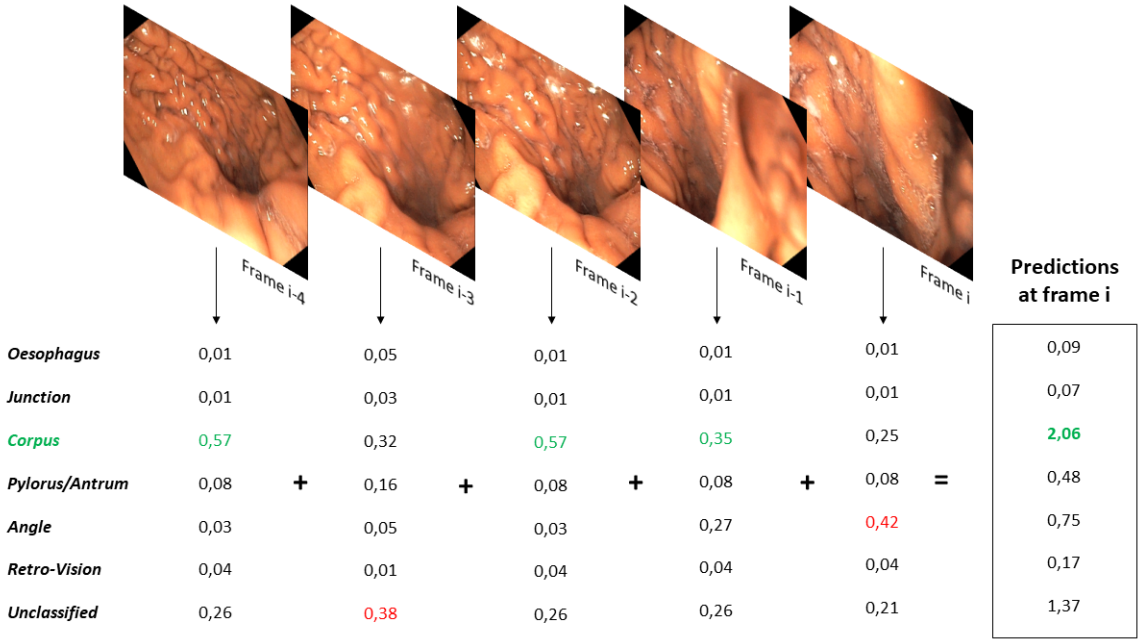

Figure 4: CNN predictions used for video classification

3 Results

The result of the training of the CNN has been conclusive. All selected models performed well. But one model called *DenseNet121* performed better than the others on the validation set, with 97.89% of accuracy. This is better or as well as the outcomes of previous other works. The main difference is that the current work focuses on EGD (video) frames and not pictures carefuly taken by the physician : more diversity and blurry images in our case. We also consider less anatomical sites (7 vs 11 in the work of Q. He et al. [3]).

This model has then been selected for video application. The main issue consisted in obtained a stable throughput. Indeed, a phenomenon called *flickering* can occur and make the prediction jumps from one frame to the other in the video. This is caused by the ambiguity existing between several sites. The solution consisting in averaging previous the previous predictions of the CNN model has given good results.

Unfortunately, no labeled video set was made so it was not possible to determine the accuracy of the solution. But a qualitative assessment has shown that the solution performed well most of the time. Two outputs from an EGD video treatment are made : (1) a time-line of the visited sites and (2) a selections of image samples for each site. Examples are given in fig 5 and 6.

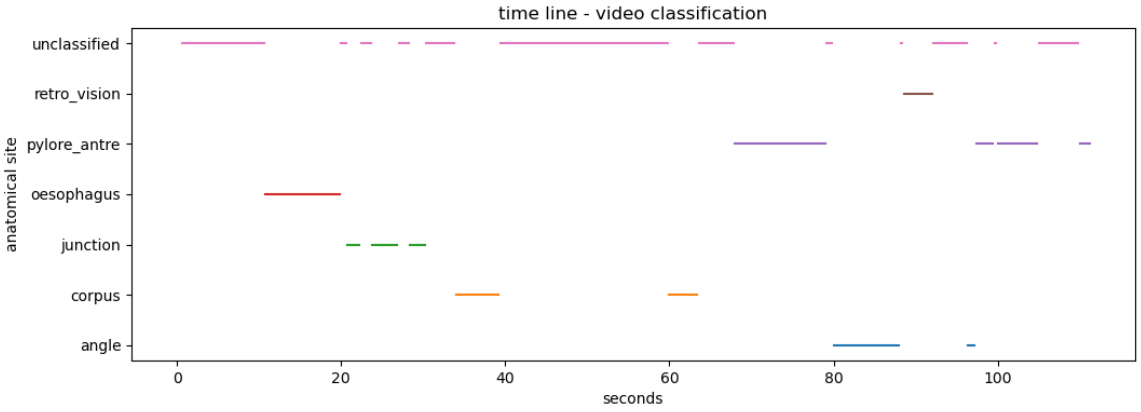

Figure 5: Example : time-line outputed at the end of an EGD processing

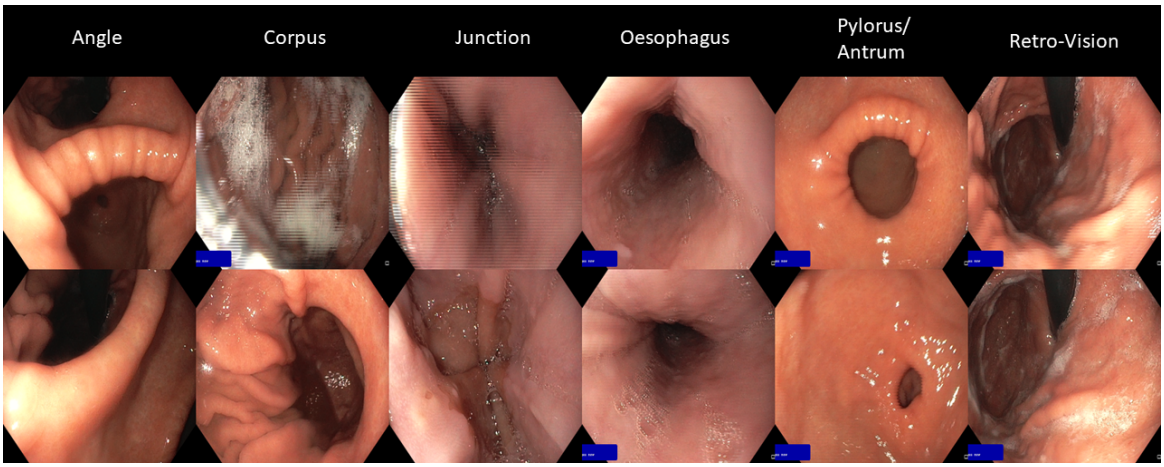

Figure 6: Example : image samples outputed at the end of an EGD processing

4 Conclusion

The main goal of the work has been met : a solution to automatically classify sequences of an EGD into their respective anatomical site has been developed.

However, some improvements could be made. Indeed, some ambiguity between several sites has been occasionally observed during video classification. It is supposed that the following changes could improve the current solution :

1. Update the current anatomical sites categories
2. Improve the data collection to include more diversity in the selected frames
3. Collect more data concerning unusual anatomies and specific pathologies
4. Create a labeled video data set of EGDs
5. Explore less trivial time-based architecture such as 3D-CNNs or RNNs [4][5]

(word-count: 807)

Bibliography

[1] SPF Santé publique, *Endoscopie digestive haute*, 20-03-2021
https://www.belgiqueenbonnesante.be/fr/variations-de-pratiques-medicales/systeme-digestif/gastro-intestinal/endoscopie-digestive-haute

[2] *Anatomy of the stomach*, 18-03-2021
https://anatomy-medicine.com/digestive-system/29-the-stomach.html

[3] Q. He, S. Bano, O.F. Ahmad, et al. *Deep learning-based anatomical site classification for upper gastrointestinal endoscopy.*, International Journal of Computer Assisted Radiology and Surgery, 2020.

[4] H. Kwon, M. Kim, S. Kwak, M. Cho. *MotionSqueeze: Neural Motion Feature Learning for Video Understanding*, 2020.

[5] S. Zha, F. Luisier, W. Andrews, N. Srivastava, R. Salakhutdinov. *Exploiting Image-trained CNN Architectures for Unconstrained Video Classification*, 2015.