QC for Digital Pathology – Scan Duplicate Classification

Bachelor's Thesis
Supervisor: Prof. Schüffler (https://schuefflerlab.org)

The Computational Pathology Lab of the TUM Institute for Pathology is offering a bachelor's thesis titled "QC for digital pathology – scan duplicate classification". In this work, a classifier is to be trained for the classification of scan duplicated into "same slide" or "different slide", depending on the preview images being similar or not. The data set of 4000 preview images (2000 duplicates) will be provided and can be extended if needed. See below for detailed description.

Summary

The TUM Institute for Pathology digitizes 200.000 tissue slides per year using high-resolution slide scanner. This process is semi-automatic: physical slides are produced in our lab, loaded into the slides scanners and automatically scanned to *whole slide images (WSI)*. The WSI is then automatically sent to our laboratory information system / database and registered to the right patient via a unique barcode on the slide.

However, errors happen during the manual creation of the slide, e.g. using the same barcode for different slides. Also, a slide can be scanned twice (e.g., due to scanning artifacts) resulting in two WSI with the same barcode. Although duplicates are easily identified at the time of patient sign-out, automated detection of such duplicates is wanted to keep the database clean and facilitate downstream research with clinical data.

This project aims to build an automated duplicate detector based on WSI. The detector will run on our retrospective database differentiating duplicates into rescans of the same slide (one of them can be deleted) or erroneous re-use of the same barcode on different slides (need manual resolution by our lab).

Dataset

Figure 1 illustrates the problem: each row shows a two scans with the same barcode. A duplicate can arise from a rescan (then the two images are similar) or from a wrong barcode (then the two images are dissimilar). The task is to automatically differentiate the two scenarios. We will use a dataset of 4000 images (2000 Training, 2000 Testing). Note that the images are not manually annotated and the annotation into "same slide" or "different slide" is part of the thesis (however, for human eyes it is a quite simple task).

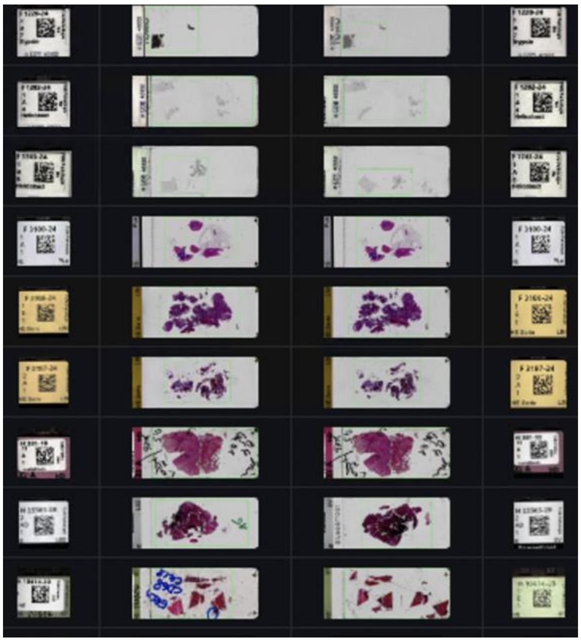

Figure 1: Examples of duplicates (WSI with the same barcode). Each row shows two scans with the same barcode. The rescan shows a similar preview image (e.g., rows 1, 2, 4, 5, 6, 7) indicating it originates from the same slide, or a dissimilar preview image (e.g., rows 3,8, 9) indicating it originates from a different slide.

Methods

First, the student creates an annotated dataset from 2000 image pairs indicating "same slide" or "different slide" by visual inspection.

Then, the student can use methods from computer vision (handcrafted, similarity based) and/or from machine learning (supervised or unsupervised) or compare both to fulfill the task.

- Computer vision methods include tissue segmentation, mask creation, similarity score definition.
- Machine learning methods include tissue segmentation, Feature extraction, model learning (e.g., random forest, deep learning) and/or clustering.

The method will then be tested on 2000 test images.

Successful outcome

A successful outcome would be a successful binary classification of two images into "rescans of same slide" or "rescans of different slide".

Depending on results, a publication is aimed in the area of quality control for digital pathology.