Generation of Synthetic Historical Documents

Manuel Drazyk

Master thesis in Computer Science

Labeled training data for historical documents of a specific time and cultural period is often hard to come by and currently can not be created reliably in an automated way. Instead of trying to automate the labeling, the approach of this thesis is to generate a document in an unsupervised manner that is similar from the original historical document in terms of its style and contains the text of a generated template document. To achieve this, in this thesis we combined a model that can recognize handwritten text with the image-to-image translation model CycleGAN to generate images that contain a given text with the same style of a arbitrary chosen historical document. The handwritten text recognition model and the generated synthetic historical documents are then evaluated qualitatively and quantitatively to review how well they perform and how useful they are as labeled training data for other machine learning models. Furthermore, the influence different template documents and methods have on the evaluation is analyzed. Our results show that the synthetic historical documents are realistic and can to a certain degree be used as labeled training data.

Prof. Rolf Ingold