Interactive tools for the transcription of handwritten documents

Team: Loki (Inria Lille – Nord Europe & CRISiAL)
Advisors: Thomas Pietrzak (thomas.pietrzak@univ-lille.fr) & Stéphane Huot (stephane.huot@inria.fr)

Description

Public and private archives contain hundreds of precious ancient documents for historic and genealogical research: Civil records, church records, military records, population censuses. These archives are regularly scanned for preservation and sharing. However accessing information is tedious when the corpus is not indexed nor transcribed. Several project aim at indexing and transcribing such documents automatically, such as Transkribus [6] or Himanis [5]. Despite their efforts, transcribing documents automatically remain complicated and unsatisfactory for historians [4]. Automatic methods are not able to interpret complex writings (irregular, with overlaps, bad scan quality) [3]. Moreover these algorithms require hand-made learning database and machine learning experts to adapt the tool to a specific document corpus.

The candidate’s work will consist in designing, implementing and evaluating interactive tools for helping manual transcription of scanned handwritten documents. For this purpose, the candidate will study the historians and genealogists activities and workflow, in order to frame their needs and methods; search for existing manual and automatic methods for transcribing handwritten documents; and analyze problems that can be solved by humans, machines, and how to effectively combine these methods.

The first task on this thesis will be to continue previous work on advanced selection techniques for handwritten text[1]. In this study we designed a pixel selection tool similar to the magic wand [2]. The problem of regular magic wands is that it propagates to the whole adjacent pixels, which leads to incorrect selections in our case. We solve this problem with a combination of a magic wand and a brush. This work is in progress, and requires tuning, as well as a user evaluation.

Later work will focus on annotation techniques that support collaborative work, as well as studying the combination of interactive and automatic approaches. This tool will be invaluable for bootstrapping the transcription of large corpuses, as well as helping transcribing small corpuses.

This thesis is part of a larger project in which we would like to combine interactive and automatic methods to get the best of both approaches. The candidate will collaborate with historians, computer scientists specialized in document analysis, and Geneanet, a leading genealogy website with thousands of users and millions of documents.

Required skills

A successful candidate must hold a MSc in computer science or equivalent, and show a great interest in performing high quality research in Human-Computer Interaction. Knowledge in OCR and machine learning is helpful. He or she must speak and write English fluently, and experience or strong interest in software development. Creativity, independence, team working and communication skills are valuable advantages.

Working environment

The thesis will be carried out in the Loki team in Lille, France, joint between Inria – Lille Nord Europe and the CRISiAL (UMR CNRS 9189) laboratory of the University of Lille.

References