Abstract

Nowadays, there are huge collections of handwritten text documents in libraries all over the world. The high demand for these resources has led to the creation of digital libraries in order to facilitate the preservation and provide electronic access to these documents. However text transcription of these documents images are not always available to allow users to quickly search information, or computers to process the information, search patterns or draw out statistics. The problem is that manual transcription of these documents is an expensive task from both economical and time viewpoints. This thesis presents a novel approach for efficient Computer Assisted Transcription (CAT) of handwritten text documents using state-of-the-art Handwriting Text Recognition (HTR) systems.

The objective of CAT approaches is to efficiently complete a transcription task through human-machine collaboration, as the effort required to generate a manual transcription is high, and automatically generated transcriptions from state-of-the-art systems still do not reach the accuracy required. This thesis is centered on a special application of CAT, that is, the transcription of old text document when the quantity of user effort available is limited, and thus, the entire document cannot be revised. In this approach, the objective is to generate the best possible transcription by means of the user effort available. This thesis provides a comprehensive view of the CAT process from feature extraction to user interaction.

First, a statistical approach to generalise interactive transcription is proposed. As its direct application is unfeasible, some assumptions are made to apply it to two different tasks. First, on the interactive transcription of handwritten text documents, and next, on the interactive detection of the document layout.

Next, the digitisation and annotation process of two real old text documents is described. This process was carried out because of the scarcity of similar resources and the need of annotated data to thoroughly test all the developed tools and techniques in this thesis. These two documents were carefully selected to represent the general difficulties that are encountered when dealing with HTR. Baseline results are presented on these two documents to settle down a benchmark with a standard HTR system. Finally, these annotated documents were made freely available to the community. It must be noted that, all the techniques and methods developed in this thesis have been assessed on these two real old text documents.

Then, a CAT approach for HTR when user effort is limited is studied and extensively tested. The ultimate goal of applying CAT is achieved by putting together three processes. Given a recognised transcription from an HTR system. The first process consists in locating (possibly) incorrect words and employs the user effort available to supervise them (if necessary). As most words are not expected to be supervised due to the limited user effort available, only a few are selected to be revised. The system presents to the user a small subset of these words according to an estimation of their correctness, or to be more precise, according to their confidence level. Next, the second process starts once these

low confidence words have been supervised. This process updates the recognition of the document taking user corrections into consideration, which improves the quality of those words that were not revised by the user. Finally, the last process adapts the system from the partially revised (and possibly not perfect) transcription obtained so far. In this adaptation, the system intelligently selects the correct words of the transcription. As results, the adapted system will better recognise future transcriptions. Transcription experiments using this CAT approach show that this approach is mostly effective when user effort is low.

The last contribution of this thesis is a method for balancing the final transcription quality and the supervision effort applied using our previously described CAT approach. In other words, this method allows the user to control the amount of errors in the transcriptions obtained from a CAT approach. The motivation of this method is to let users decide on the final quality of the desired documents, as partially erroneous transcriptions can be sufficient to convey the meaning, and the user effort required to transcribe them might be significantly lower when compared to obtaining a totally manual transcription. Consequently, the system estimates the minimum user effort required to reach the amount of error defined by the user. Error estimation is performed by computing separately the error produced by each recognised word, and thus, asking the user to only revise the ones in which most errors occur.

Additionally, an interactive prototype is presented, which integrates most of the interactive techniques presented in this thesis. This prototype has been developed to be used by palaeographic expert, who do not have any background in HTR technologies. After a slight fine tuning by a HTR expert, the prototype lets the transcribers to manually annotate the document or employ the CAT approach presented. All automatic operations, such as recognition, are performed in background, detaching the transcriber from the details of the system. The prototype was assessed by an expert transcriber and showed to be adequate and efficient for its purpose. The prototype is freely available under a GNU Public Licence (GPL).