ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Distributed Computing

Prof. R. Wattenhofer

SA:

Collaborative OCR for Tabularized Data

This document describes the subject and the general time schedule of the Semester thesis of *Jan Bernegger* in the autumn term 2012. Adaptations or changes can be agreed upon by the advisors.

Motivation and Informal Description

Our Smart Shopping project aims to provide (among others) a convenient way to scan and archive shopping receipts. To this end, we use the open-source OCR engine Tesseract, but due to the special tabularized structure and the usual bad print quality of receipts, Tesseract performs rather bad – compared to its impressive performance for running text.

The goal of this thesis is to increas Tesseract's recognition performance by improving frame conditions in which Tesseract has to work. This means, for example, preprocessing the images before giving them to Tesseract, training Tesseract with a given training set, creating dictionaries for common items on receipts, etc. This data should be made available to all users of the system to enable the users to collaborate with each other in order to improve the OCR performance.

There are many other ways how we think we can obtain a better OCR performance, but the objective of this thesis is also that **you** come up with ideas that help our purpose. So if you have interesting ideas on how to tackle our problem, feel free to contact us!

Requirements: Creative thinking, advanced programming skills, basic skills in image processing, and the ability to work independently are necessary to work on this topic successfully.

Contacts

- Tobias Langner: tobias.langner@tik.ee.ethz.ch, ETZ G61.4
- Jochen Seidel: jochen.seidel@tik.ee.ethz.ch, ETZ G61.1

Detailed Project Outline

We denote the following primary tasks mandatory (on the right side you find a rough estimate for the time that we allocate to the respective task):

- Getting familiar with the topic ($\star$)
- Preprocessing stage (rotate/de-skew image, adjust contrast, crop image) ($\star\star$)
- Image analysis and logo detection ($\star\star$)
- Use detected logo and corresponding layout to invoke Tesseract ($\star\star$)
- Develop collaborative training approach (Android app, server, protocol) ($\star\star\star$)
- Evaluate impact of individual components on recognition performance ($\star$)
- Write a report documenting the design and development process as well as the final status of the project and prepare a final presentation ($\star\star$)

The Students' Duties

- Weekly meetings with the advisors to discuss current matters
- Regular check-ins into the provided *revision control system*
- A final presentation (15 min) of the work and results obtained in the project
- A final report (English), presenting work and results