Proposal for Tutorial on Tesseract at DAS2016

Title: Tesseract Blends Old and New OCR Technology

Presenter: Ray Smith, Google Inc, rays@google.com

Abstract:

This tutorial will cover the algorithms, design and implementation of the open source OCR engine known as Tesseract.

Designed largely in secret, the methods used in Tesseract are not well known, yet it remains a formidable force in OCR, and continues to improve. The layout analysis was second in the 2009 ICDAR competition, it supports more than 100 languages, including Chinese, and several Indic languages, and recent changes have allowed easy plug-in of new classifiers, including a new Deep LSTM addition.

This tutorial will lay all the cards on the table, covering the following topics:

- Background/history
- Overall architecture
- Internal data structures
- Character classification
- Segmentation and language models
- New Deep LSTM Implementation and Integration with Tesseract’s Language Model
- Challenges of truly multilingual OCR
- Live demos

Target Audience:

- Academic groups that use or want to use Tesseract in their research.
- Students wanting to start a Tesseract-based research project.
- Industrial groups interested in using Tesseract in their products.
- OCR researchers wanting to learn from a lot of experience.

Expected Attendance:

The Tesseract tutorial at DAS 2014 was presented to a full house.

Motivation and Learning Outcomes:

Tesseract is a widely used open source OCR engine that is also used as a baseline for many academic papers. At the previous DAS, a tutorial on Tesseract was well attended and generated a lot of useful discussion and questions. The renewed Tutorial that will be presented at DAS 2016 will cover a new Deep LSTM implementation that will be added to the open source version of Tesseract in March 2016. Audiences attending the DAS tutorial will therefore benefit from being the first to learn about the leading edge of OCR technology, as it is applied to the most popular and best open source OCR engine.

Optional Hands-on opportunity:
This tutorial runs live demos throughout. Attendees who bring their own latop will be able to have a hands-on experience in which they get to build and run the latest Tesseract on their own machine, to follow along with the demos.
Bring along your own laptop with the following configuration to take part:

- Hardware: Laptop **with external Mouse!! (Scroll wheel needed.)**
- Ram/Disk: Most laptops under 10 years old should handle it with ease.
- Operating system: Linux or Windows. No specific version needed.
- C++ compiler: Linux: gcc, Windows: Visual Studio Express 2010 OR Mingw.
- Java runtime. Version not important.
- Working WiFi or USB port for downloading software and data.
- Mac users: The Tesseract demos are **very hard to use without a scrollwheel.** If you can emulate that somehow, bring it along, and we will give it a try! No guarantees though.

Topic and Description

Tesseract is a well-known open source OCR system that is used by many academic and industrial users. The previous English version has had more than 1.7m downloads, and 10 languages more than 100k downloads. It would be impossible to cover all of Tesseract in a half-day tutorial, and there are new and exciting topics to cover in the Deep LSTM addition, so this tutorial will cover a limited number of topics in depth: the original feature extraction, classification, and beam search scheme, the new Deep LSTM system and its integration, and the effect of the LSTM system on accuracy over a broad spectrum of languages. The presenter will be emphasizing lessons learned in building a full general-purpose, multilingual OCR system, and comparing to current technologies, where appropriate, which should be of use to a significant number of the DAS audience.

Outline Program:

Half-day format, split into seven roughly 30-minute segments as below. The ordering is designed so that attendees can install and build Tesseract on their own machines during the coffee break and run it themselves during the sessions in the second half:

1. **Introduction and history.** Taken largely from the keynote on Tesseract presented at DRR in February 2013, this section will set the scene for the rest of the tutorial with project motivation, some engineering history, and some useful lessons learned.
2. **Overall architecture, coordinate spaces and data structures.** Essential to understanding the detail of the design, the major data structures (page hierarchies, API, classification results) and how they fit into the overall architecture are covered. This section is a prerequisite for successfully understanding the subsequent modules.
3. **The Importance of Language Independence.** Tesseract supports about 100

languages, but how language-independent is it, and how well does it recognize them? This section ends with a guide to installing and building the Tesseract source code, for those that have laptops and wish to participate in the hands-on option.
**** Coffee break, Installing and building the code. ****

4. **Character segmentation, language models and beam search.** This section talks about the character chopper, the chop-then-join philosophy vs. more recent methodologies that aim to avoid segmentation, the integration with the language models, and the beam search. **Includes live graphical demos of the character segmentation search and beam search.**
5. **Character classifiers.** Detailed description of feature extraction and the 2-stage character classifier, its strengths, weaknesses, and how it relates to past and contemporary machine learning. **Includes live graphical demos of the character classifier in action.**
6. **New Deep LSTM implementation.** Description of a new generic Deep Net/LSTM implementation inside Tesseract.
7. **Deep Neural Net integration.** How the Deep Net implementation is integrated with the language model and beam search. **Includes live graphical demos of the new Deep LSTM engine in action.**

Previous Edition

A previous edition of this tutorial was presented to DAS 2014.
Slides: https://drive.google.com/file/d/0B7I10Bj_LprhbUlIUFICdGtDYkE/edit?usp=sharing
The tutorial for DAS 2016 will replace 3 sections to put significant emphasis on the new Deep LSTM engine as follows:
3. Training -> Challenges of multilanguage OCR and a presentation of results covering a broad spectrum of languages, comparing the older Tesseract engine with the newer LSTM-based system.
7. Layout Analysis -> New LSTM implementation, covering the network description language, the layer components available, and some usable network architectures.
8. Modernization efforts -> New LSTM integration, covering how the Deep LSTM system fits with the rest of Tesseract, makes use of the language model, and allows the use of the old Tesseract engine as a fall-back.

Material

The tutorial website will include all slides used in the tutorial, and attendees will also be able to download and build Tesseract from the source code if they wish.

Equipment

Standard presentation equipment will be adequate for the tutorial.

Brief Resume of Presenter

Ray Smith
Senior Staff Software Engineer
Google Inc.
1600 Amphitheatre Pkwy
Mountain View
CA 94043
USA
rays@google.com
+1 650 335 5369

Ray spent 8 years at HP Labs Bristol, developing the Tesseract OCR engine, including classifier technology, textline finding, visualization tools, distributed test system, OCR for compression. The next 3 years were spent developing HP PrecisionScan: HP’s premier document-oriented scanning software. This was followed by 7 years at Caere Corporation/ScanSoft, re-architecting Omnipage to make better use of multiple OCR engines in combination, making substantial accuracy improvements between version 10 and version 15. Ray has spent the last 10 years at Google, once again working on Tesseract, to make it a truly multilingual OCR system covering more than 100 languages, and most recent adding LSTM technology to it.
Ray has published several papers and patents on topics related to OCR, winning Best Industrial Paper Award at ICDAR 2011, and presented a keynote talk on Tesseract at DRR 2013.