Improving Access to Text
IMPACT

Technical deliverable documentation

Dissemination level: P U (P u b l i c)

D-TR4.2 - TYPEWRITTEN OCR PROTOTYPE | TR4

Document history

Revisions

Version	Status	Author	Date	Changes
0.1	Draft	Stefan Pletschacher Apostolos Antonacopoulos	27.01.2011	Created
1.0	Final	Stefan Pletschacher Apostolos Antonacopoulos	24.02.2011	Feedback from Günter Mühlberger, Clemens Neudecker, Lotte Willms, and Sebastian Colutto incorporated

Approvals

This document requires the following approvals:

Version	Date of approval	Name	Role in project	Signature
0.1	18 February 2011	Clemens Neudecker	Technical Project Manager	OK
0.1	21 February 2011	Lotte Wilms	Internal reviewer for CB5	OK
0.1	21 February 2011	Günter Mühlberger	SP Leader TR	OK
1.0	25 February 2011	Hildelies Balk	General PM	OK

Distribution

This document was sent to:

Version	Date of sending	Name	Role in project
0.1	17.02.2011	Günter Mühlberger	SP Leader TR
	17.02.2011	Clemens Neudecker	Technical Project Manager
	17.02.2011	Lotte Wilms	Internal reviewer: CB5 Tool Patron
	17.02.2011	Sebastian Colutto	WP TR4 member
1.0	25.02.2011	Hildelies Balk	General Project Manager
	28.02.2011	Liina Munari	EC Project Office

KB IMPACT is supported by the European Community under the FP7 ICT Work Programme. The project is coordinated by the National Library of the Netherlands

IMPACT – Technical deliverable documentation

D-TR4.2 - Typewritten OCR Prototype

1. Partners

USAL

2. Deliverable

The deliverable comes as a zip file containing the command line executable **`TypewrittenOCR-0-1-60.exe`** for Windows operating systems including required third party DLLs and a short documentation. Moreover, there is an example enclosed (images and a PAGE file describing text regions) with two batch files for demonstration purposes ("`Examples (start from image).bat`" and "`Examples (start from regions).bat`"). In order to run the Typewritten OCR prototype directly on document images (in contrast to running it on isolated text regions which is its primary purpose) there is a very basic region segmenter included as well. If you are interested in this tool please get in touch with the PRImA Research Group at http://www.primaresearch.org/cont.php.

3. Background

Typewritten documents are unique among machine-printed documents in the way they are created. Each character is produced independently of the others by pressing a key on the typewriter and ink is mechanically transferred on the paper proportionally to the force of the keystroke. This results in non-uniformity of the intensity of the printed areas. Even within a single word, there can be characters that are faint (lightly pressed) while others are strongly pressed resulting in much darker, blurred and filled-in characters. These problems are worse in carbon copies (of which many exist as primary sources). Another peculiarity lies in the nature of the content of typewritten documents. They mostly originate from correspondence or administrative tasks involving less natural language and instead more frequently names, abbreviations, numbers etc which render lexicon aided recognition approaches less useful.

In order to meet these challenges, a new approach which integrates background knowledge about this special type of document has been implemented in the Typewritten OCR prototype. The main idea is to precisely extract and compute individually enhanced glyph images from the original material prior to the actual classification stage. Whereas block and text line segmentation can be handled independently in most cases, it is crucial to treat glyph segmentation, enhancement, and recognition together in one combined system due to their strong interdependencies. Figure 1 shows an overview of the architecture of the Typewritten OCR prototype. Region and text line segmentation are considered external processing steps but for ease of use an extended version of the USAL segmenter (part of D-TR2 "Segmentation and classification toolkit") has been included in this deliverable in order to demonstrate the whole process chain.

Broken characters are repaired by making a small number of hypotheses as to the location of the missing parts. A number of features are then extracted from each glyph image and, using a combination of classifiers, the character is

recognised. The implemented approach is completely language independent[1] as it does not rely on any kind of lexicon-based post correction.

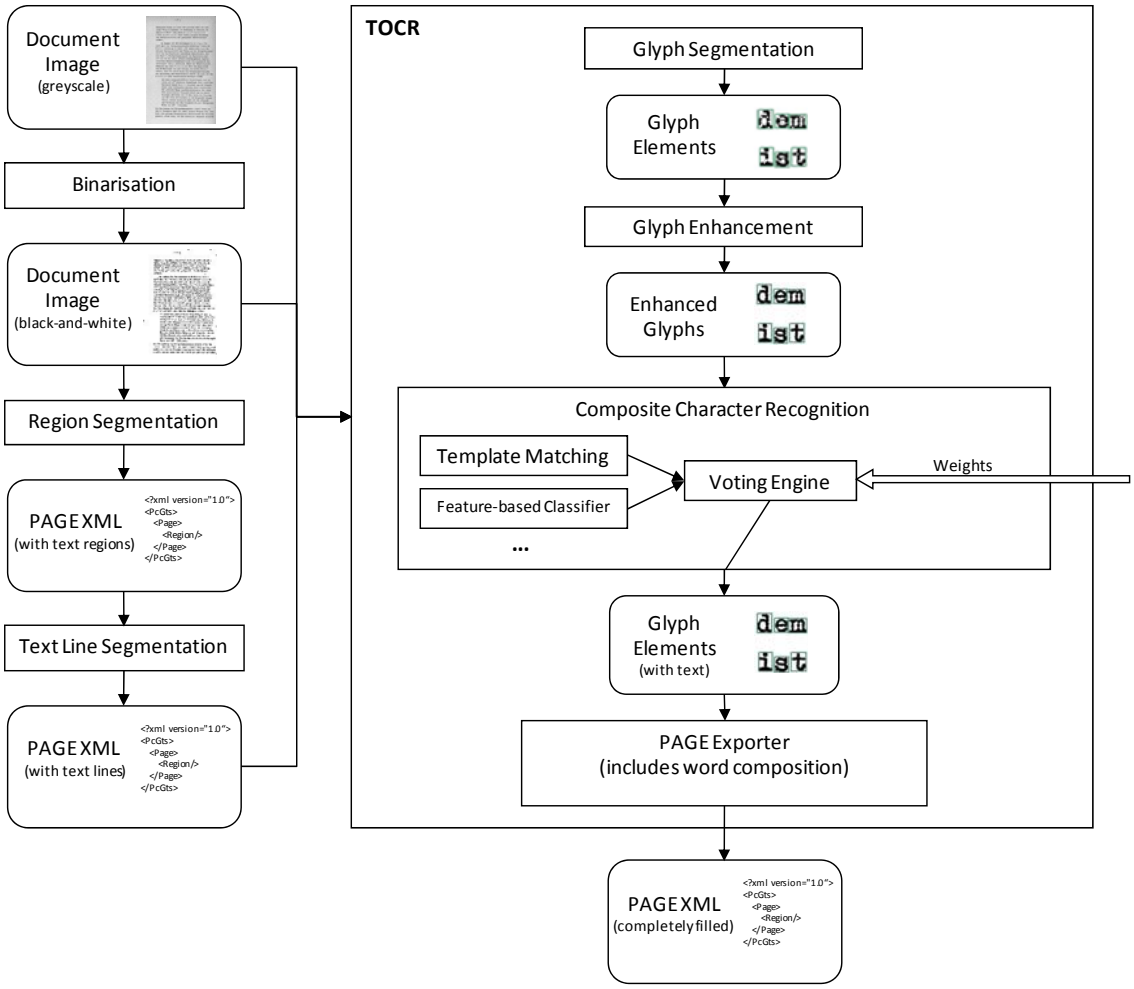

Figure 1. Structural overview of the Typewritten OCR prototype

4. Outline of functionality

In order to allow stand-alone demonstration as well as integration into the OC5 framework the Typewritten OCR prototype has been compiled into a command-line tool which can be controlled by a set of parameters. The tool takes as input a bitonal TIFF image, an optional greyscale or colour TIFF image, and a PAGE file (http://schema.primaresearch.org/PAGE/gts/pagecontent/2010-03-19/) describing the outline of text regions and lines. Output is also a PAGE file, however, completed with words, glyphs and the actual recognised Unicode-encoded text filled in on all levels. Figure 2 shows the recognition result of a typewritten document stored in PAGE and displayed in Aletheia.

[1] Apart from the fact that it has been specifically trained for Latin alphabets. Other alphabets like Cyrillic could be easily trained as well.

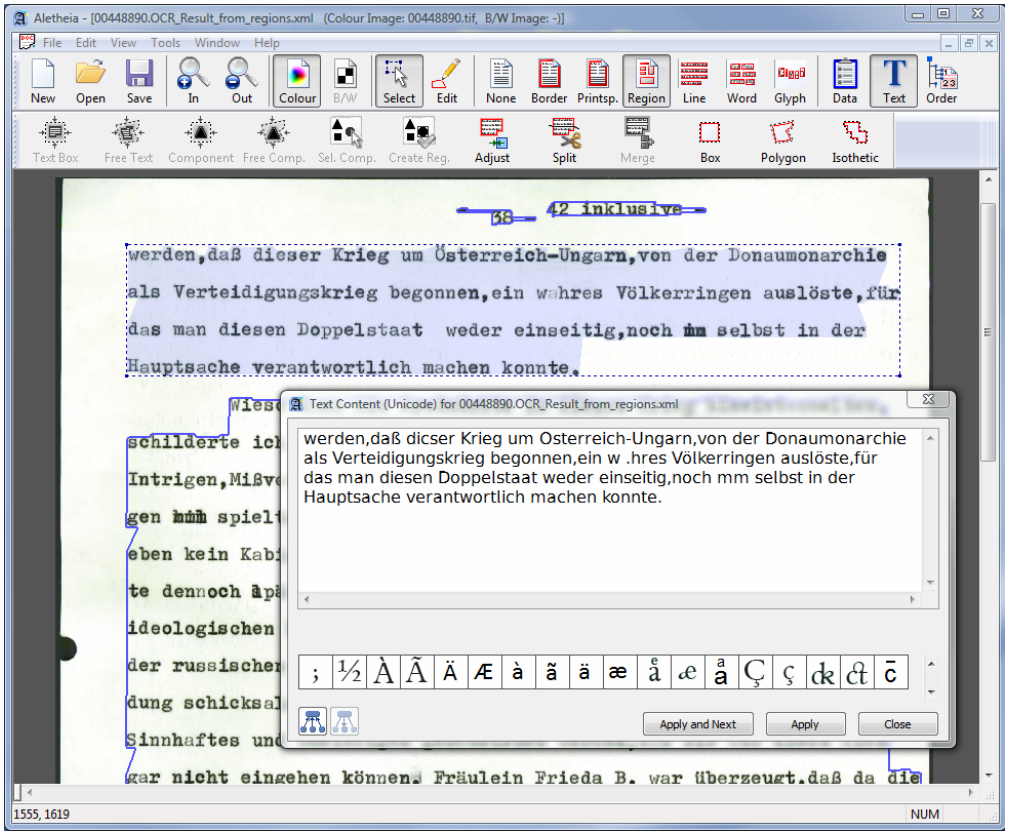

Figure 2. Recognition result of an average quality typewritten document.

It has to be noted that the overall text recognition quality depends on the accuracy of the preceding segmentation steps (i.e. correct text regions and lines are expected as input for the actual Typewritten OCR prototype).

Usage

The command line executable can be used as follows:

```
TypewrittenOCR-0-1-60.exe <binaryimage> <greyscaleimage> <colourimage> <inputxml>
<outputxml> [classifierweights]
```

Where:

`<binaryimage>`	Filename of the binary image to be analysed (required)
`<greyscaleimage>`	Filename of the greyscale image to be analysed (optional, set to -1 if not using)
`<colourimage>`	Filename of the colour image to be analysed (optional, set to -1 if not using)
`<inputxml>`	Filename of a PAGE XML file containing a region and line segmentation layout

```
<outputxml>          Filename to which a PAGE XML file containing the recognition
                     result should be written
<classifierweights>  A comma separated list (without spaces) of the weight for
                     each classifier. Currently there are two classifiers available:
                      Template matching and
                      Feature analysis
                     To weight them equally, for example, use: 1,1
```

Examples

Included in the deliverable are two batch files which can be used for a quick demonstration of the tool.

1. Starting from a segmentation result
   This example is to show the intended usage of the tool: Integrated in a workflow which performs region segmentation first and then applies the Typewritten OCR. An example segmentation result (in PAGE format) has been included in the Examples folder along with the corresponding bitonal and greyscale images. In order to run the Typewritten OCR prototype on this example simply run "**`Examples (start from regions).bat`**" from a Windows command line.

2. Starting from an image
   In order to show the Typewritten OCR prototype in a stand-alone environment (i.e. without having segmentation results from a previous workflow step available) another helper tool has been bundled with this deliverable. When running "**`Examples (start from image).bat`**" from a Windows command line the same example as above will be processed, however, this time only using the included images. This is achieved by running a basic region segmenter first which stores its results (in PAGE format) in a temp folder before the actual Typewritten OCR is started.

In both cases, the output of the Typewritten OCR is stored in PAGE format in the Results folder.

5. Testing and Evaluation

Testing

There are three layers of testing implemented to ensure stability and compatibility of the tools:

1. Internal – by the main developer (USAL)
2. Internal – by an independent person who is not directly involved in the development (USAL)
3. External – by technical project partners (within the scope of OC5 – via integration into the IMPACT Framework)

Feedback from testing is centrally collected and managed by means of a server-based bug tracker (Mantis).

Source code and releases are managed by a version control system (Subversion) which enables tracking of changes as well as reverting to earlier development stages.

Evaluation

A prerequisite for evaluating the Typewritten OCR prototype was the collection of a suitable dataset followed by the creation of ground truth for a representative subset. The typewritten dataset currently consists of 2627 document images (http://www.prima.cse.salford.ac.uk:8080/impact-dataset/private/specificsets.php -> Typewritten -> All) and for 300 of which ground truth is available with text and polygonal outlines on region level (http://www.prima.cse.salford.ac.uk:8080/impact-dataset/private/specificsets.php -> Typewritten -> Typewritten (withGT)).

Preliminary evaluation trials for a small number of documents (10 typewritten pages from different sources and of different types, containing ca 9'000 characters – ground truth arrived very late for more evaluation to be carried out and included at the time of writing this document) showed already very promising results. Especially documents of low quality, which is quite common for typewritten material due to use and ageing issues, and documents posing difficulties to lexicon-based post correction methods can greatly benefit from this enhancement-based approach. Figure 3 shows an example of how the new glyph extraction and enhancement methods as integral part of the Typewritten OCR prototype can improve the completeness of character segmentation (no missed glyphs) and hence the overall accuracy of recognition results.

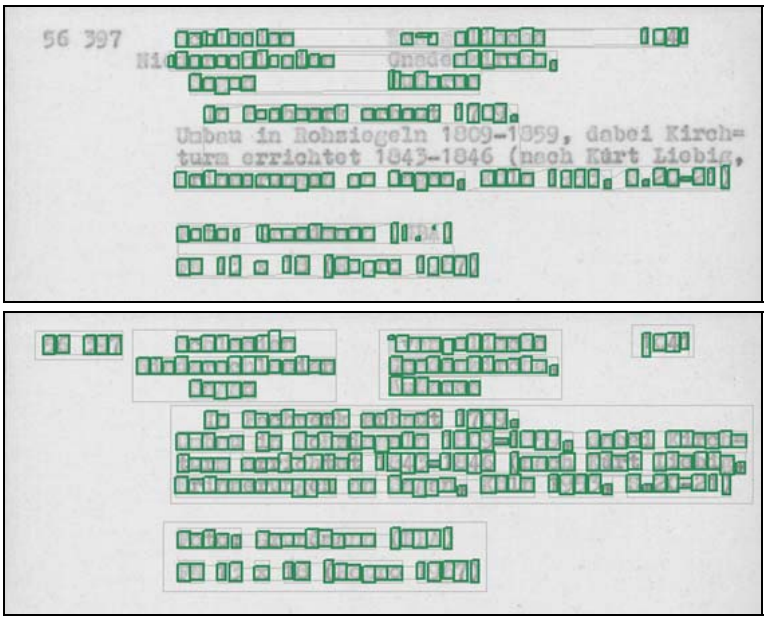

Figure 3. Glyph detection (green boxes): Top – FineReader Engine 9, Bottom – Typewritten OCR prototype

With an average character recognition accuracy of 87.5% for the first version of the Typewritten OCR prototype it was already possible to surpass the current state-of-the-art (FineReader Engine 9) with 85.2%.

A large scale evaluation of the Typewritten OCR prototype is still ongoing since the required ground truth became only available in January 2011. Updated evaluation results will be made available as soon as this is finished. The availability of a considerable amount of ground truth data will also allow for further training and optimisation of the prototype which will be reflected in the final evaluation.

Early work on the Typewritten OCR prototype was published and presented at the 11th International Conference on Document Analysis and Recognition (ICDAR 2009, Barcelona) in "A New Framework for Recognition of Heavily Degraded Characters in Historical Typewritten Documents Based on Semi-Supervised Clustering" by S. Pletschacher, J. Hu, and A. Antonacopoulos. More publications on the final prototype are envisaged for ICDAR 2011 in Beijing and subsequent conference and journal articles.

A number of potential extensions and improvements were discovered during the research and development activities and if resources permit will be followed up in the final year of the project. In particular it is planned to further explore and improve internal feedback mechanisms between the core components (glyph segmentation, enhancement, and recognition) in order to iteratively improve recognition results.

Contribution towards project expectations:

☐	*Software tool is fit to be put into productive use and is supported by the necessary installation guides*
☐	*Software tool can be made available in a productive environment with further development which is clearly defined*
☒	*Software tool demonstrates potential functionality and is available in a publicly accessible environment*
☒	*Report of findings of research available (for experimental tools only)*

6. License and IPR protection

Due to the research-oriented (proof of concept) nature of this work package, the final deliverable is a prototype and therefore licensing is not currently envisaged in the present form of the tool. Commercialisation will be considered but will require additional work in order to develop the prototype further into a real product.

For copyright of the used third party libraries libxml2 (http://xmlsoft.org/), LibTIFF (http://www.remotesensing.org/libtiff/) and OpenCV (http://opencv.willowgarage.com/) see the respective websites.