READ

RECOGNITION & ENRICHMENT
OF ARCHIVAL DOCUMENTS

D5.12
Page Image Explorer (PIE) P2

Markus Diem, Stefan Fiel, Florian Kleber
CVL

Distribution: http://read.transkribus.eu/

READ
H2020 Project 674943

This project has received funding from the European Union's Horizon 2020
research and innovation programme under grant agreement No 674943

Project ref no.	H2020 674943
Project acronym	READ
Project full title	Recognition and Enrichment of Archival Documents
Instrument	H2020-EINFRA-2015-1
Thematic priority	EINFRA-9-2015 - e-Infrastructures for virtual research environments (VRE)
Start date/duration	01 January 2016 / 42 Months

Distribution	Public
Contract. date of delivery	31.12.2017
Actual date of delivery	28.11.2017
Date of last update	21.12.2017
Deliverable number	D5.12
Deliverable title	Page Image Explorer (PIE) P2
Type	report
Status & version	in progress
Contributing WP(s)	WP5
Responsible beneficiary	CVL
Other contributors	CVL
Internal reviewers	NAF, ASV
Author(s)	Markus Diem, Stefan Fiel, Florian Kleber
EC project officer	Martin MAJEK
Keywords	Document Clustering, Visualization

Contents

1 Executive Summary

The Page Image Explorer (PIE) allows intuitive exploration of documents. The key idea is to access potentially unsorted document collections and connect/group their items by user defined criteria. Hence, PIE strongly focuses on user interaction and visualization of large document collections. PIE will be built upon the READ Framework[1] which is publicly available under LGPLv3.

In D5.11 the prototype for document clustering was presented. The presented prototype used features based on the previous framework system. This comprised features like *textheight*, *form type*, *text line spacing*, *writing color*, *paper color*, etc. For a detailed description see D5.11. Most of the features are currently not available as a result based on the currently developed framework and tools used in Transkribus (e.g. the layout analysis in the new framework is based on baselines where similar features will be derived for clustering).Thus, D5.12 presents the selected features, which will be used for the clustering and which are available as result of the currently developed tools.

Section 2 describes the features used for clustering and Section 4 gives an outlook to D5.13.

2 Features

The basic layout analysis and the layout analysis is presented in D6.5 and D6.11. Currently, the baselines of text are extracted in documents. The baseline information will be used to derive the following features:

- average line spacing
- text density
- text position
- presence of text vs. image

The text density will be estimated based on the detected text regions. Also the text position will be estimated by the text region information. The baseline to polygon/rectangle tool is used to estimate the text region itself (see github for the tool). Based on the polygon/rectangle additional features will be calculated:

- average x-height
- min/max x-height
- slant of the text

The writer identification will be integrated in the same way as in the previously presented clustering tool (see D5.11). Only experiments regarding the distance function will be done (e.g. cosine distance). After an evaluation phase of the clustering tool the following features can be calculated to improve the clustering if necessary:

[1] `https://github.com/TUWien/ReadFramework`

- paper color
- text color
- document class (e.g. certain table)

The evaluation will show if the listed features are needed (currently, the developed layout tools do not provide the listed features). As an option, additional features based on the developed text spotting or the document understanding can be added in the future.

3 UI Mock-Ups

The PIE visualization interface will mainly show the embedding viewport which features OpenGL for responsive zooming and panning. Figure 1 shows a mockup of the planned UI. Here, each document is represented by a dot in the embedding space. The embedding space groups similar document pages (the user defines which similarity function(s) to take). Hence, a page's position rather relies on its relationship with all other documents than typical Euclidean distances. In Figure 1, all document pages are visualized as colored dots, where the color indicates the document group. Groups are created and labeled by users with respect to their exploration results. Dock widgets (right) allow users to choose which collections to display. In addition, users can choose which groups to display and how the similarity should be computed (i.e. which features to consider for the subspace embedding).

Document pages can be selected and grouped, and labeled in the embedding space. If a group size is low enough, thumbnails of the page images are visualized rather than dots (see Figure 2).

4 Outlook

The main development based on the selected features will be implemented in D5.13. The visualization will utilize OpenGL viewports with Qt overpainting (for e.g. nice font rendering), which allows PIE to scale better compared to its precursor. Thus, the front-end for 2D spatial clustering will be the main work of D5.13. PIE will be developed publicly and is available on github[2].

[2] https://github.com/TUWien/ReadFramework

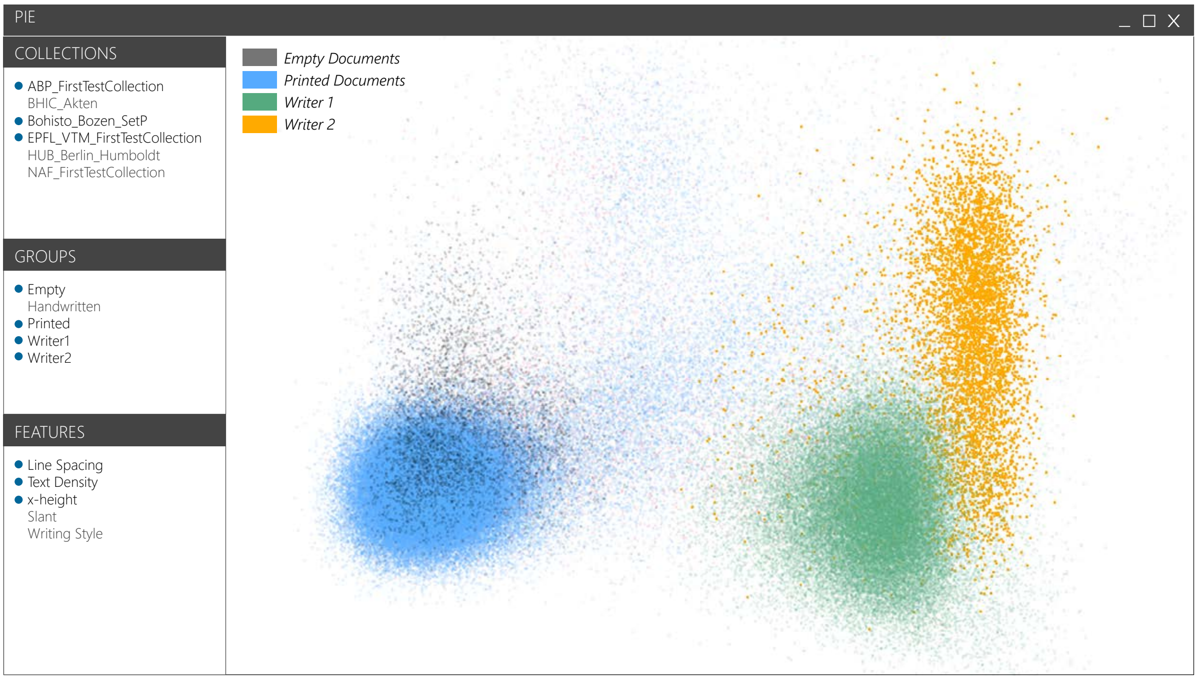

Figure 1: UI mock-up of the *all documents* view.

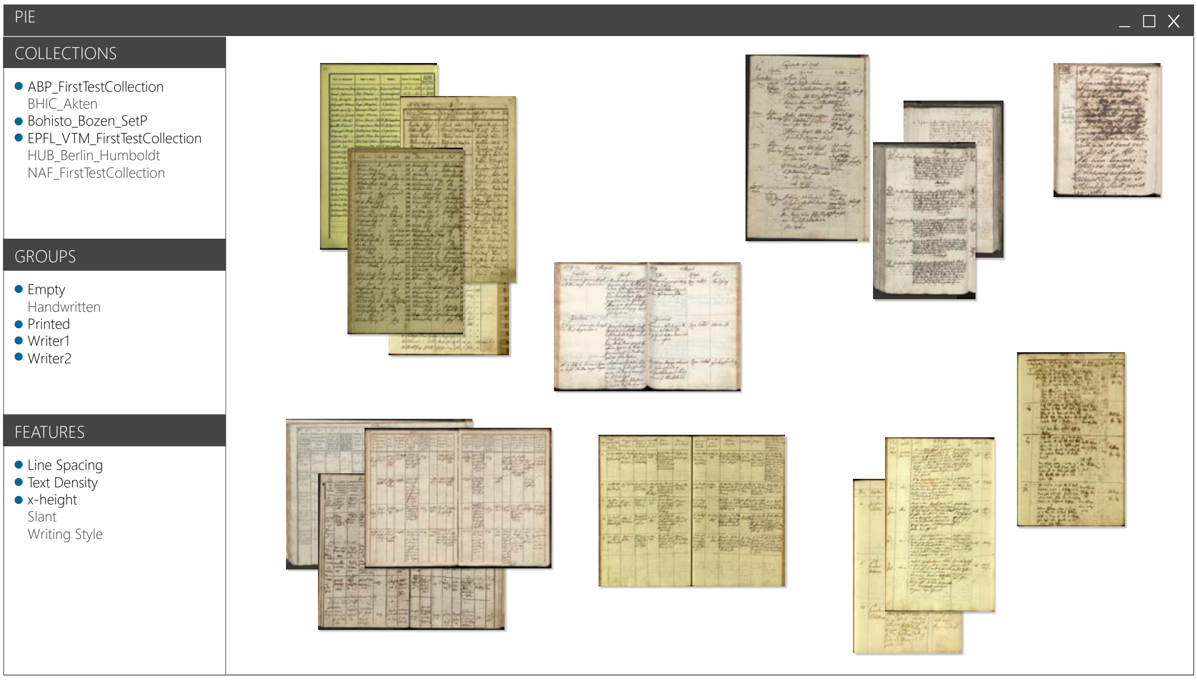

Figure 2: Thumbnail visualization of only a few selected documents.