READ

RECOGNITION & ENRICHMENT
OF ARCHIVAL DOCUMENTS

D7.16
Writer Identification and Retrieval

Markus Diem, Stefan Fiel and 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	31.12.2016
Actual date of delivery	28.12.2016
Date of last update	22.12.2016
Deliverable number	D7.16
Deliverable title	Writer Identification and Retrieval
Type	report
Status & version	in progress
Contributing WP(s)	WP7
Responsible beneficiary	CVL
Other contributors	
Internal reviewers	DUTH, NCSR
Author(s)	Markus Diem, Stefan Fiel and Florian Kleber
EC project officer	Martin Majek
Keywords	Writer Identification, Writer Retrieval

Contents

1 Executive Summary

Writer Identification and Retrieval is the task of identifying the scribe of a document after creating a ranking of documents in a dataset according to the similarity of the handwriting to a reference document. These methods can be used to determine the author of documents or to search for documents in the archive where the author is not known.

The current deliverable implements the methodology presented by Fiel and Sablatnig in [1] and [2]. SIFT features are calculated and a feature vector for the handwriting is created by using the bag-of-words approach, respectively the Fisher vector. These feature vectors can then be compared to the vectors of other pages. The result of this comparison is the similarity of the handwriting. By using a nearest neighbor classification the author of a handwritten document can be identified or ranking of documents according to the similarity of the handwriting can be generated.

2 CVL Framework

The CVL Framework developed contains basic image processing algorithms and methodologies for Document Image Analysis. It is developed within the project and is the basis for the WP of CVL. It contains also the methods for writer retrieval and identification. The tasks are developed in C++ and available at github under LGPL-3.0:
`https://github.com/TUWien/ReadFramework`
`https://github.com/TUWien/ReadModules`
ReadFramework contains the implementation of the methods, whereas ReadModules contains a C++/Qt plugin which uses this code and can be loaded and executed using nomacs[1]

3 Transkribus module

The module for the integration of the writer identification and retrieval method to the Transkribus plattform is available at:
`https://github.com/TUWien/CVLModules`
It is the defined interface of the ReadFramework to the Transkribus platform.

4 Writer Identification and Retrieval

Currently two writer identification and retrieval methods are developed. The first method is the implementation of Fiel and Sablatnig [1] which is based on a bag of words approach using SIFT features. The SIFT features are slightly modified by removing the rotational independence up to 180 °. These features are calculated on the writing area of the document image and then their distance to pre-trained centers are calculated and a histogram of the closest center is built for each page. This histogram

[1] nomacs - ImageLounge `http://www.nomacs.org`

is then used for the identification of the writer respectively for retrieval of pages with similar handwriting.

The second method is based on Fiel and Sablatnig [2], which is an improvement of the first method. Instead of searching for the closest center, the distribution of the feature is interpreted as a probability function. With the use of Gaussian Mixture models this distribution is incorporated into a feature vector using the Fisher Kernel. Again, this feature vector is used for the identification of the writer or for the retrieval.

The current implementation achieves an identification rate on the ICDAR 2013 writer identification dataset [3] of 93.0%. This dataset consists of 1000 pages, written by 250 different writers. Each writer has written four pages, two in English and two in Greek. For the retrieval a performance of 45.8% for the Top 2 is achieved, which means for a reference document the 2 most similar pages returned by the system are from the same writer. The performance for having all three others are in the three most similar pages is 22.5%.

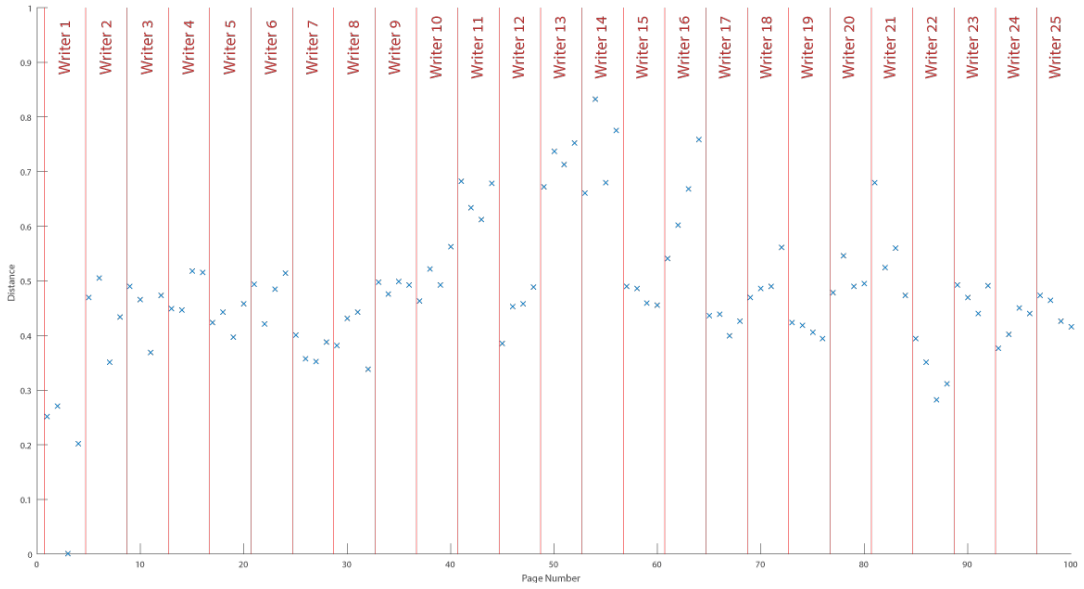

Figure 1: Plot of the similarity distances of the first 100 pages of the ICDAR 2013 dataset to reference document #3. The distance of page #3 to itself is zero. It can be seen that pages of the same writer often have a similar distance, which indicates that an identification or retrieval is possible.

Figure 1 shows a visualization of the similarity distances of the first 100 pages of the ICDAR 2013 dataset. All distances shown are calculated with respect to the reference page # 3. The red vertical lines indicate a change of the writer, whereas the 4 crosses between two red lines are the 4 documents of the particular writer in the dataset. It can be seen that page # 3 has a distance of 0 to itself and the smallest distance to the other 3 documents of Writer 1. The documents of the other writers have similar distances to the reference document, which indicates that the similarity of the handwriting is calculated correctly.

Additional, a new dataset consisting of historical documents has been created, which is presented in D5.8. First experiments on this new dataset have been carried out.

An identification rate of 72.6% was achieved,although the Fisher vector was created on modern handwriting and also the parameters of the methods have not been changed. Currently a new approach for writer retrieval is developed using deep learning. Patches are extracted on the text-lines and features are learned by presenting triplets of these patches to the deep learning method, which tries to minimize the interclass distance and maximize the intraclass distance. The encoding of these features is still work in progress.

References

[1] S. Fiel and R. Sablatnig. Writer Retrieval and Writer Identification Using Local Features. In Michael Blumenstein, Umapada Pal, and Seiichi Uchida, editors, *2012 10th IAPR International Workshop on Document Analysis Systems (DAS)*, pages 145 –149. IEEE, march 2012.

[2] S. Fiel and R. Sablatnig. Writer Identification and Writer Retrieval Using the Fisher Vector on Visual Vocabularies. In *2013 12th International Conference on Document Analysis and Recognition (ICDAR)*, pages 545–549, 2013.

[3] G. Louloudis, B. Gatos, N. Stamatopoulos, and A. Papandreou. ICDAR 2013 Competition on Writer Identification. In *12th International Conference on Document Analysis and Recognition (ICDAR)*, pages 1397–1401, Aug 2013.