READ

RECOGNITION & ENRICHMENT
OF ARCHIVAL DOCUMENTS

D7.17
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	31.12.2017
Actual date of delivery	12.12.2017
Date of last update	12.12.2017
Deliverable number	D7.17
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 contains information about two newly developed methods and also new datasets which have been created for the task of writer identification. One of these datasets has been used for the "ICDAR2017 Competition on Historical Document Writer Identification (Historical-WI)"[3].

2 Writer Identification and Retrieval

Two methods are currently developed. Both of them are based on deep learning approaches. The first one has been published at the ICDAR 2017 [1], while the second method is scheduled to be published to a future conference related to document image analysis.

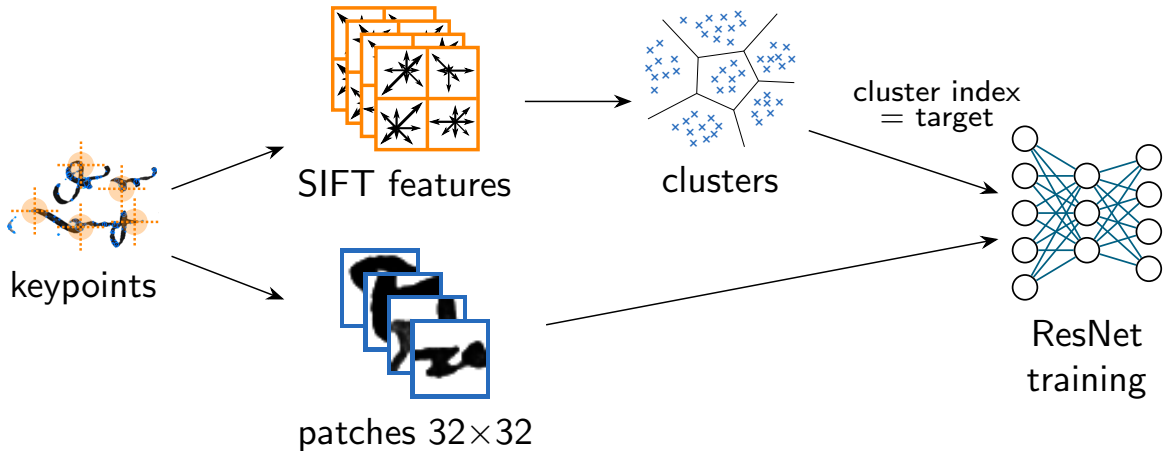

Figure 1: Overview of the unsupervised feature learning. At SIFT keypoint locations, SIFT descriptors and image patches are extracted. The cluster indices of the clustered SIFT descriptors represent the targets and the corresponding patches as input for the CNN training.

The method developed in [1] is based on deep learning using surrogate classes and the workflow is illustrated in Figure 1. Patches of the document image are extracted on the SIFT keypoint location, since our last method [2] has shown that the neighborhood of these points contains enough information for the identification of the writer. Additionally, the SIFT descriptors are extracted and clustered to a predefined number of groups to form surrogate classes. The patches can then be trained on the surrogate classes. For identification the activation of the penultimate layer is taken as feature vector. All feature vectors of one page are then combined using VLAD encoding to form

one feature vector for each page which can then be used for the identification of the writer or retrieval of the pages.

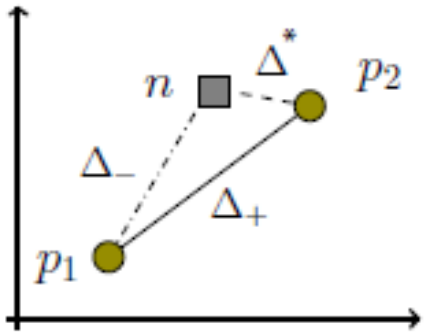

Figure 2: Illustration of the triplet learning. The distance between the positive samples p_1 and p_2 should be minimized whereas the distances to the negative sample n should be maximized.

The second method is also based on deep learning, but uses triplet-learning. Patches are again extracted at the location of the SIFT features. The network is then used to learn the similarities of patches from the same writer but simultaneously dissimilarities to patches of different writers. During the training step three patches are given to the neural network as illustrated in Figure 2. Two originating from the same writer and one patch of a different writer. The task of the neural network is to minimize the distance between the two patches of the same writer and to maximize the distance of the two patches to the third patch, which is from a different writer.

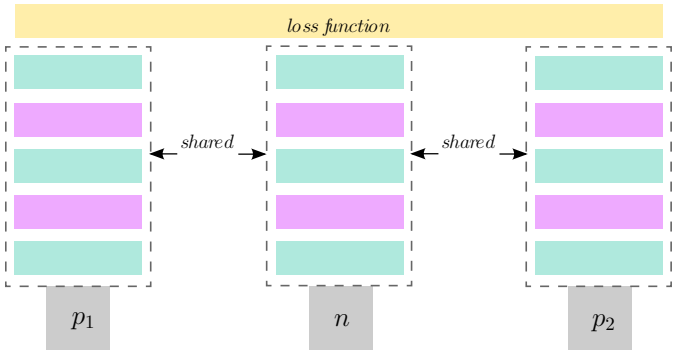

Figure 3: Architecture of the triplet network. The positive patches p and the negative patch n are feed into the network and the loss function is used to learn the similarity respectively the dissimilarity. The weights of the three parallel networks are shared, which means that all patches are processed with the same network.

The features for this training are extracted using shared weights, as illustrated in Figure 3. Thus, for identification all patches of the page are processed using one network and again these features are encoded, this time using a Fisher vector, to form one feature vector for each page. As it was mentioned above, this method is planned to be submitted to a forthcoming document image analysis conference (e.g. ICFHR, ICPR).

3 Writer Identification Competition

"ICDAR2017 Competition on Historical Document Writer Identification (Historical-WI)"[3] was organized by members of the READ team in conjunction with the ICDAR 2017 conference. The dataset was created using document images from the Universitätsbibliothek Basel. Five different institutions submitted six different methods, a number of submissions which is comparable with the number of the last competitions. The competition ran on the Scriptnet Platform(See Deliverable D3.8). The competitions follow the procedure of providing the test data and asking for result files (not executables) from the participants. The results of the competition can be seen in Table 1.

Table 1: Detailed evaluation of the participating methods. The methods are sorted alphabetically.

Method	Top-1	Hard-2	Hard-3	Hard-4	p@2	p@3	p@4	mAP
Barcelona	67.0	45.1	27.4	12.6	58.5	50.6	43.2	45.9
Fribourg	47.8	24.7	12.6	5.5	39.3	33.2	28.5	30.7
Groningen	76.1	54.9	36.4	18.5	67.5	59.4	51.2	54.2
Hamburg	67.1	46.5	29.5	14.5	59.0	51.5	44.2	46.9
Tébessa I	74.4	52.2	34.8	18.2	65.2	57.4	49.7	52.5
Tébessa II	76.4	56.6	37.8	21.3	68.4	60.3	52.7	55.6

Results of the previously described methods "SIFT+FV"[2] and "Surrogate classes" [1] and are shown in Table 2. It can be seen that both methods have a better performance then the methods submitted to the competition [4][5]. Both method were not included into the competition results, since the were developed by organizers.

Table 2: Results of two READ methods on the competition dataset.

Method	Top-1	Hard-2	Hard-3	Hard-4	p@2	p@3	p@4	mAP
SIFT+FV	81.4	63.8	46.2	27.7	74.0	66.7	59.0	62.2
Surrogate classes	88.6	77.1	64.7	46.8	78.9	72.3	59.2	74.8

4 Future work

The writer identification methods will be improved. Additionally, the methods will be investigated to their invariance of the changing of the handwriting style of a writer within decades with the new dataset which will be created in D5.9. Furthermore, a new writer identification method based on the newly proposed capsule network will be developed.

References

[1] Vincent Christlein, Martin Gropp, Stefan Fiel, and Andreas Maier. Unsupervised Feature Learning for Writer Identification and Writer Retrieval. In *2017 14th International Conference on Document Analysis and Recognition (ICDAR)*, pages 991–997, 2017.

[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] Stefan Fiel, Florian Kleber, Markus Diem, Vincent Christlein, Georgios Louloudis, Nikos Stamatopoulos, and Basilis Gatos. ICDAR2017 Competition on Historical Document Writer Identification (Historical-WI). In *2017 14th International Conference on Document Analysis and Recognition (ICDAR)*, pages 1377–1382, 2017.

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

[5] G. Louloudis, N. Stamatopoulos, and B. Gatos. ICDAR 2011 Writer Identification Contest. *2011 11th International Conference on Document Analysis and Recognition (ICDAR)*, pages 1475–1479, 2011.