READ
RECOGNITION & ENRICHMENT OF ARCHIVAL DOCUMENTS

D7.18
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	31.12.2018
Actual date of delivery	28.12.2018
Date of last update	29.12.2018
Deliverable number	D7.18
Deliverable title	Writer Identification and Retrieval
Type	report
Status & version	Final report
Contributing WP(s)	WP7
Responsible beneficiary	CVL
Other contributors	
Internal reviewers	NCSR, DUTH
Author(s)	Markus Diem, Stefan Fiel and Florian Kleber
EC project officer	Christophe DOIN
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 the newly developed methods and the current progress for a new method. The focus now lies on the user perspective; the user should be able to understand a decision made by the algorithm.

2 Learning Similarities

A new method for Writer Identification and Retrieval has been developed based on the last milestone of this work package and is published at the ICFHR 2018 [4]. It follows the scheme, which is currently used to the best of our knowledge by all state-of-the-art deep learning methods. This scheme can be seen in Figure 1.

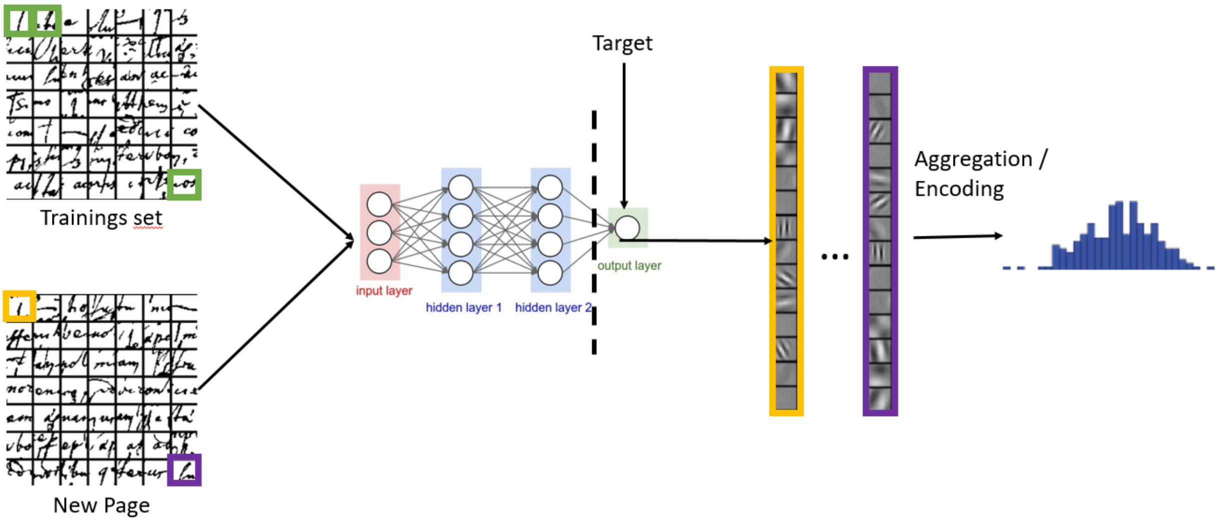

Figure 1: Writer Identification and Retrieval scheme.

First patches of the handwriting are extracted using the location of SIFT features. It has already been shown, like in [2] and [3] that the neighborhood of these location contain enough information for an successful identification of the writer. The patches extracted are than fed into a Neural Network to train a specific target. We use, like in [2], surrogate classes as target but, as a second approach, the different writers from the trainings dataset. In contrast to learning a classification task, this paper proposes to learn a similarity measurement between image patches using a triplet loss function. This means that three patches are presented to the network, two from the same writer or surrogate class and one patch from another class. The goal of the network is now

Table 1: Comparison of the method proposed to two other state-of-the-art methods.

		hard		
	MAP	Top 1	Top 2	Top3
Christlein et al.[1]	88.0	99.4	81.0	61.8
Fiel and Sablatnig[3]	67.4	94.5	48.0	25.7
Keglevic et. al[4]	86.1	98.9	77.9	56.4

to learn an embedding in which the distance between the two samples is minimized, whereas at the same time the distances two the other patch should be maximal.

For applying the neural network to the task of writer identification or retrieval, one single feature vector has to be generated for each patch. Thus, a aggregation or encoding of patch features has to be made. So the last layer of the network, which is responsible for the assignment of the class, is cut of and the activation of the second last layer is used as feature vector for a patch. A VLAD encoding is used to generate the feature vector of one page, which is then used for the identification of the writer. Table 1 shows the results compared with other methods. It can be seen that current method performs slightly worse than others, which can be explained that dataset contains Greek and English documents. Since the alphabet of these two languages differ, it shows that the proposed method is currently not invariant to the change of the alphabet.

3 Improving the Comprehensibility

The next focus lies on the comprehensibility of the results which are presented to the user. As already mentioned above, currently the writer identification and writer retrieval methods generate one feature vector for each page. The distance of feature vectors of two pages is then calculated to determine whether it is the same writer or not. Is is also possible to sort the examined pages according to the distances, which is the result of writer retrieval. So pages which look very different to the human eye have a very small distance, whereas pages of the same writer can have a large one (without the try of counterfeiting). Thus a new method is in development which should overcome these problem and makes it possible to present a visualization of the decision process to the user. When comparing two pages, patches are extracted from page A. These patches contain characters or combination of characters. Page B is then searched for patches with similar characteristics. Figure 2 shows 2 patches from 4 writers with the characters "ve" or "ne" on it. A network, similar to the one which is presented above, is then trained to tell whether the same writer or not has produced a patch pair.

Such patch pairs are found all over the two pages and for each of these pairs a decision is made. Figure 3 shows how these pairs are generated. On the left page an image patch is selected, which is the character "a" in this case and on the right page similar patches are found. In is not important, that all "a"s are found in the right page, as long as there are enough pairs of the page. The number of patch pairs found may also indicate if the documents have been written by the same writer, but currently this is not taken into account.

Figure 2: 8 patches of 4 different writers which have been classified as similar. The network for the decision if it is the same writer is trained with triplets of these patches.

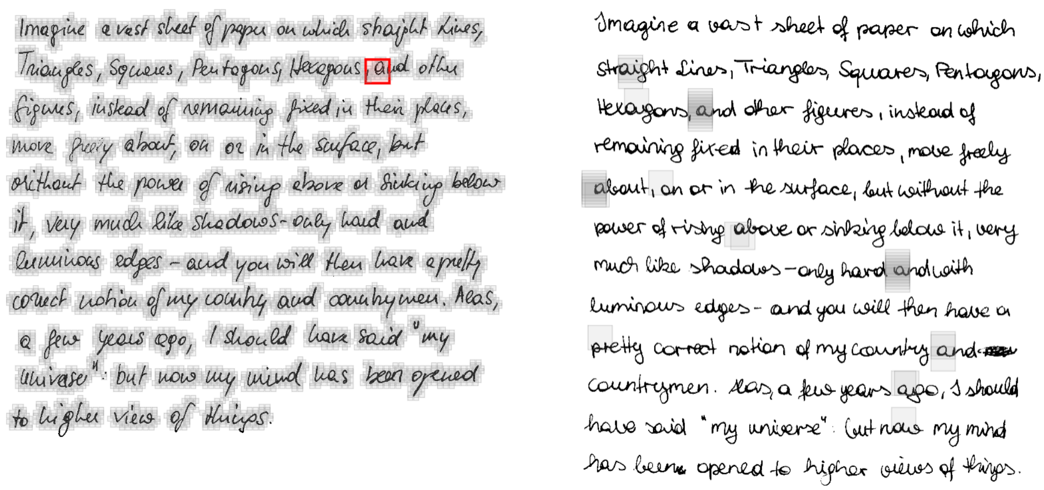

Figure 3: Left: Query image with all possible patches (gray) and one selected patch (red). Right: All patches which are similar to the red patch on the left side

At the end a voting over all these pairs is made to determine if the same person has written the two pages. Since this method is computational expensive it is planed to use older methods, like [1], [3], [2], to make a ranking of the similarity and then only examine the first couple pages of the ranking. If a user now wants to analyze why one page is assigned to a writer the voting can be visualized. The user is than able to see which patches in both pages look similar and which have been identified as belonging to the same writer. This method is currently in development and we can not presented any results here.

References

[1] Vincent Christlein, David Bernecker, and Elli Angelopoulou. Writer Identification Using VLAD Encoded Contour-Zernike Moments. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 906–910, 2015.

[2] 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.

[3] 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.

[4] Manuel Keglevic, Stefan Fiel, and Robert Sablatnig. Learning Features for Writer Retrieval and Identification using Triplet CNNs. In *16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 211–216, 2018.