ICDAR2017 Competition on Historical Document Writer Identification (Historical-WI)

S. Fiel, F. Kleber and M. Diem
Computer Vision Lab
TU Wien
Vienna, Austria
Email: fiel@caa.tuwien.ac.at

V. Christlein
Pattern Recognition Lab
FAU Erlangen-Nürnberg
Erlangen, Germany
Email: vincent.christlein@fau.de

G. Louloudis, N. Stamatopoulos and B. Gatos
Computational Intelligence Laboratory
National Center for Scientific Research "Demokritos"
Athens, Greece
Email: louloud@iit.demokritos.gr

***Abstract*—The ICDAR 2017 Competition on Historical Document Writer Identification is dedicated to record the most recent advances made in the field of writer identification. The goal of the writer identification task is the retrieval of pages, which have been written by the same author. The test dataset used in this competition consists of 3600 handwritten pages originating from 13^{th} to 20^{th} century. It contains manuscripts from 720 different writers where each writer contributed five pages. This paper describes the dataset, as well as the details of the competition. Five different institutions submitted six methods which were ranked using identification and retrieval metrics. The paper describes the competition details including the dataset, the evaluation measures used as well as a short description of each submitted method.**

I. Introduction

Writer identification refers to the problem of assigning the correct writer for a given query document image by comparing it with document images for which the writers are known. The similarity of the handwriting can be computed and a ranking based on this similarity can be generated. This ranking is used to retrieve all documents of the corresponding writer. Thus, people who are working in the humanities can use these algorithms to analyze their manuscripts to determine whether a specific author has written other documents or determine the writer of a specific document. In the past years, several scientific datasets have been released [1], [2], [3], [4]. These datasets have been used for the evaluation of several techniques [5], [6], [7] which reported very high and similar performance. Thus, existing datasets cannot help for efficiently comparing writer identification methods and there is a need for a more competitive dataset. In ICDAR 2017 Competition on Historical Document Writer Identification (Historical-WI), a real world test dataset consisting of 3600 pages of historical documents was created from the digital archive of the Universitätsbibliothek Basel[1]. Sample pages of the dataset can be seen in Figure 1. The dataset consists of color as well as of binary images. Existing competition datasets were generated in a restricted environment and as a result they have several characteristics such as uniform background and non-overlapping text lines. In contrast, the current dataset consists of historical documents which do not

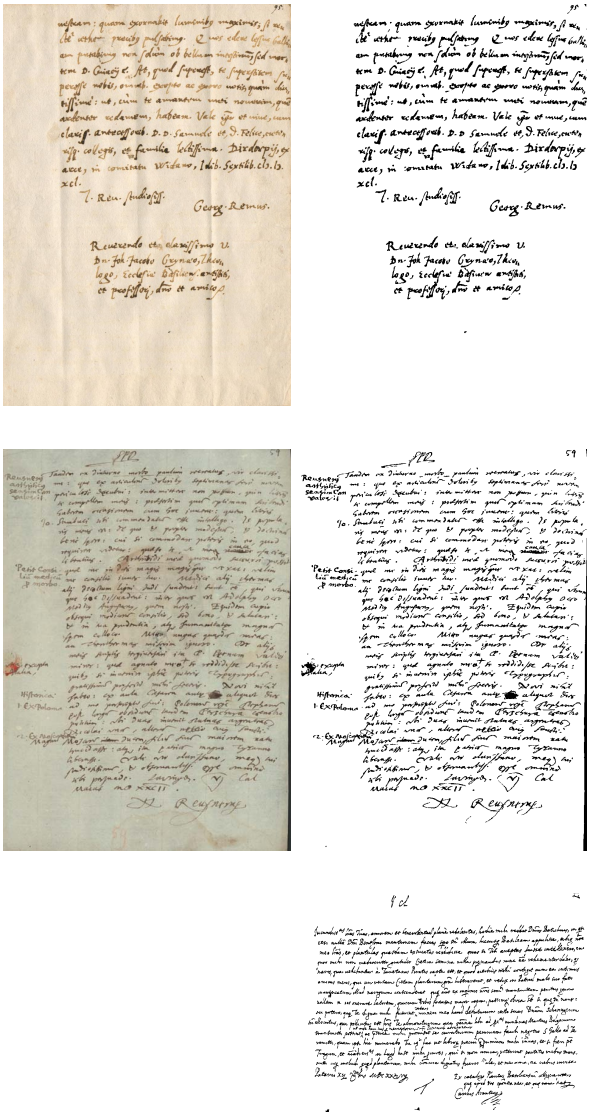

Fig. 1. Three sample pages of the Historical-WI dataset. In the left column the color image with the respective binarized image in the right column.

[1]http://www.e-manuscripta.ch/ - accessed July 2017

2379-2140/17 $31.00 © 2017 IEEE
DOI 10.1109/ICDAR.2017.225

Fig. 2. Sample document images of the unprocessed dataset which included photographs, envelopes, blank pages with bleed-through ink, illustrations, small pieces of written text, and technical drawings.

Fig. 3. Document image with the corresponding SIFT features. The SIFT features are binned with 50 column and row wise and the rows and columns below a certain threshold (1/3 of the maximal value) are skipped. Then the area of the remaining rows and columns is calculated, which is the estimated area for the writing zone.

have a uniform background, the text lines often overlap and words differ among pages.

This paper is organized as follows: First, in Section II, an overview of the dataset is given, including the page selection process from all pages of the digital archive of the Universitätsbibliothek Basel. Section III presents the participating methods while Section IV provides an analytical description of the evaluation protocol used. The results of the competition are presented in Section V and finally, conclusions are drawn in Section VI.

II. Dataset

The initial dataset is the current electronic library of the Universitätsbibliothek Basel. It consists of 140 000 images, which are released under Public Domain Mark. The images not only contain document images, but also drawings, music scores, photographs, blank pages, envelopes, small pieces of handwritten pages and technical drawings are also part of the dataset. See Figure 2 for some samples. The document images consist mainly of correspondences, but also some notes and books are included. The documents are written in different languages, most of the times German and French are used, but also Arabic handwriting occurs in the dataset.

The process of filtering out pages is mainly done automatically. The first step is the analysis of the METS files, which are provided with the images. We check if an author name is stored in the file. If not, we filter out these images. For authors whose year of birth / death is available, we require all dates and the name to match in order to guarantee that the scribe was actually the same person.

The next step to reduce the dataset is by filtering the images according to their text occurrence. For an estimation of the text region, several heuristics like the distribution of SIFT features are taken into account. This is done since according to Brink et al. [8] at least 100 characters are needed when using strong features. Figure 3 shows the heuristic to estimate the text occurrence based on SIFT features.

	Test set	*Training set*
Number of images	3600	1182
Writers	720	394
Pages per Writer	5	3
dpi	300	300

TABLE I
Most important properties of the READ Historical-WI dataset.

At the end of the selection process and based on the final number of images, the size of a test set and a training set has been defined. Table I shows the most important properties of the dataset. Finally, the test set consists of five document images per individual writer and three document images are available for training. Note that no writer of the training set has any page in the test set. 720 writers contributed to the test set, resulting in 3600 pages. For the training set 394 writers remained, which give a total of 1182 pages. All images have a quality of 300 dpi and are stored in jpeg format respectively png format for the binarized images. The dataset was made publicly available after the end of the competition[2].

[2]https://zenodo.org/record/854353

III. Methods and Participants

Five (5) research groups have participated with six (6) different methods for writer identification. In the following, brief descriptions of the respective submission are given. The order of appearance is alphabetical.

A. Barcelona: Computer Vision Centre (CVC), Universitat Autònoma de Barcelona (Anguelos Nicolaou and Dimosthenis Keratzas)

The method is based on [9], it is totally learning free and uses grayscale images as input. Sparse Radial Sampling Local Binary Patterns (SRS-LBP) histograms at radii up to 12 are extracted for the full images and pooled globally to form an embedding of 3072. The features are then normalized and projected to 200 dimensions with a PCA transform. The method can be reproduced from the webpage[3].

B. Fribourg: DIVA, University of Fribourg, Switzerland (Vinaychandran Pondenkandath and Marcus Liwicki), and MindGarage, TU Kaiserslautern, Germany (Muhammed Zeshan Afzal)

The method uses a deep convolutional neural network (CNN), trained using the triplet margin loss metric [10] to transform a given input into a space where inputs belonging to the same class (writer) are close to each other. We use triplets which consist of the anchor, positive and negative samples. The anchor and positive samples belong to the same class, and the negatives belong to any of the other several different classes. The CNN used is a ResNet18 [11] model which is pretrained on the ImageNet dataset for the ImageNet Large Scale Visual Recognition Challenge[12]. The individual samples for the triplet consist of cropped (256×256) sub-images from the input images. We use standard data augmentation methods during training. At testing time, we generate a vector for each input image by averaging the embeddings produced by multiple random crops on the same input. Finally, the pairwise cosine distance between all input images are computed and the images are ordered in decreasing similarity to a given query image.

C. Groningen: ALICE, University of Groningen, the Netherlands (Sheng He and Lambert Schomaker)

The CoHinge feature [13] is the joint distribution of the Hinge kernel on two different pixels of writing contours based on spatial joint feature distribution described in [14]. First, we extract ink contours from the binarized image. For each point on the contour, the hinge kernel with two angles (α_i,β_i) described in [15] is computed. In order to capture the spatial information, we compute the joint distribution of two hinge kernels with a fixed length on the ink contours as the CoHinge kernel: (α_i,β_i,α_j,β_j). All 4D CoHinge kernels from the ink contours are quantized into a 4D histogram as a feature vector.

[3]http://nicolaou.homouniversalis.org/2015/08/05/srslbp.html - accessed July 2017

D. Hamburg: Hamburg University, Centre for the Study of Manuscript Cultures, Germany (Hussein Mohammed, Volker Maergner, Thomas Konidaris, H. Siegfried Stiehl)

The method is based on Naive Bayes Nearest-Neighbour (NBNN) classifier and it takes into consideration the particularity of handwriting patterns by adding an orientation constraint to prevent the matching of irrelevant keypoints. SIFT algorithm is used to detect and describe keypoints in the images. No page layout analysis is applied and the binarised images are not used by the method. The method is inspired by [16] with some variations. The NBNN is used here instead of Local NBNN, and the normalization factor is not applied.

E. Tébessa I: Larbi Tebessi University, Department of Mathematics and Computer Science, Algeria (Abdeljalil Gattal and Chawki Djeddi)

In this method, the different configurations of oriented Basic Image Features (oBIFs) columns histograms [17], [18] extracted from binarized historical document samples are concatenated for generating a feature vector and the City block distance measures is used for classifying each historical document.

F. Tébessa II: Larbi Tebessi University, Department of Mathematics and Computer Science, Algeria (Abdeljalil Gattal and Chawki Djeddi)

Similar to the first method, the different configurations of oriented Basic Image Features (oBIFs) columns histograms [17], [18] extracted from smoothed binary historical document samples with low-pass filters are concatenated for generating a feature vector and the City block distance measures is used for classifying each historical document.

IV. Performance Evaluation

The mean-Average-Precision (mAP) is used for the evaluation of the Historical-WI competition since it is a very common and widely used measurement for a retrieval problem. Since for this competition the participants had to generate a ranking according to the similarity of the handwriting it can be seen as a retrieval problem. The most similar document is the identification task, and the other documents in the ranking are used for writer retrieval as described in Section I. The mAP is calculated as follows:

$$mAP = \frac{\sum_{q=1}^{Q} AveP(q)}{Q} \tag{1}$$

where Q is the set of all documents and q the current query document image, and $AveP$ the corresponding average precision. The average precision is the area under the precision-recall curve and also takes the position of the positive samples in the ranking into account. It is calculated as follows:

$$AveP = \frac{\sum_{k=1}^{n}(P(k) \times rel(k))}{number\ of\ relevant\ documents} \tag{2}$$

where P is the precision, $rel(k)$ is an indicator function equaling 1 if the item at rank k is a relevant document, zero

TABLE II
DETAILED EVALUATION OF THE PARTICIPATING METHODS. THE METHODS ARE SORTED ALPHABETICALLY.

Method	Top-1	Hard-2	Hard-3	Hard-4	Soft-5	Soft-10	p@2	p@3	p@4	mAP
Barcelona	67.0	45.1	27.4	12.6	76.9	80.1	58.5	50.6	43.2	45.9
Fribourg	47.8	24.7	12.6	5.5	62.1	68.3	39.3	33.2	28.5	30.7
Groningen	76.1	54.9	36.4	18.5	83.9	85.8	67.5	59.4	51.2	54.2
Hamburg	67.1	46.5	29.5	14.5	76.6	80.2	59.0	51.5	44.2	46.9
Tébessa I	74.4	52.2	34.8	18.2	82.1	85.0	65.2	57.4	49.7	52.5
Tébessa II	76.4	56.6	37.8	21.3	84.1	86.6	68.4	60.3	52.7	55.6

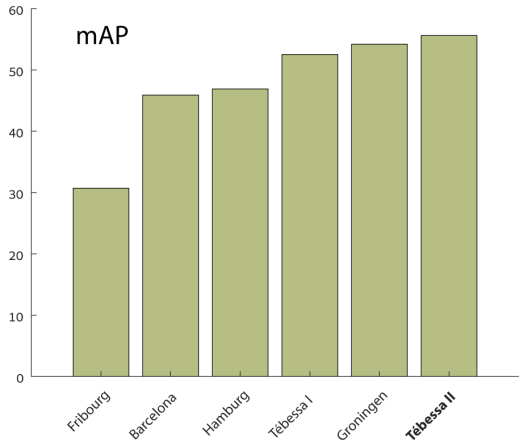

Fig. 4. Result of the Historical-WI. The ranking of the methods is according to the mAP score.

otherwise, and n is the number of all documents in the dataset. For the Historical-WI, n equals 3600 and the "number of relevant documents" is 5, since every writer has 5 documents in the test set (thus the $rel(k)$ functions has only 5 times the result 1). The mAP is a good indicator for the quality of the ranking, but has the disadvantage that it is not intuitive. Thus, other evaluations metrics are also presented in this paper. The Top-1 precision is the percentage of the document images, where the author of the first page in the ranking equals the query image. Similar to the ICDAR 11 and 13 Competition on Writer Identification [1][2], we present also the hard and soft criteria. The hard criterion means that the first k pages in the ranking have to be from the same writer as the reference document. We present the hard criterion up to rank $k = 4$. For the soft criterion, at least one document image in the first k ($k = 5$ and $k = 10$) documents in the ranking has to be written by the same writer. The precision at k, which corresponds to the percentage of correct documents within the first k pages, is also given for the values 2, 3 and 4.

V. RESULTS

For the submission of the results, the ScriptNet[4] platform was used. The participants had to download the dataset and afterwards generate a ranking according to the similarity of the handwriting. The ranking file is then submitted via the platform where the evaluation is carried out and the participants retrieve the mAP and Top-1 precision. Results of the competitors were kept secret.

Figure 4 shows the result of the Historical-WI competition. The ranking is generated according to the mAP score. The second method of Tébessa has the best mAP followed by Groningen and Tébessa I. The numbers of the mAP can be seen in the detailed evaluation, which is presented in Table II. The difference between Tébessa II and Groningen is only 1.4. Interestingly, the only system relying on deep learning presented by the team Fribourg performed worst with a mAP score of 30.7. The difficulty of training deep learning based features for this dataset has also been observed in a parallel work [19]. The Top-1 precision is 76.4% for Tébessa, which is only 0.3% better than the performance of Groningen. For the Hard-4 criterion the best performance is 21.3%, which means that only for every fifth page the method was able to find all other four pages of the writer. The soft criterion reveals that for nearly 14% of all reference pages no other document of the same writer is in the first ten documents of the ranking. Figure 5 shows on the left side three pages of the test set, where all methods failed to fulfill the Soft-10 criterion. The three images on the right are three different pages from the corresponding writer in the test set. It can be seen that for the first writer both pages originate from the same correspondence. In total there were 239 pages (6.6%) where all methods failed to find a page of the same writer within the first 10 pages of the ranking. For 103 pages (2.9%) all methods achieved to find all four other pages of the writer within the dataset. Figure 6 shows one of these pages and a page of the same writer.

VI. CONCLUSION

The ICDAR 2017 Competition on Historical Document Writer Identification (Historical-WI) is dedicated to the advances in the field of writer identification. The test dataset consists of 3600 document images of the Unversitätsbibliothek Basel. They originate from the 13^{th} to 20^{th} century. The document images show mainly one page of correspondences, but also some pages of books and notes are included. The participants were evaluated by means of mAP, but also a detailed evaluation with the soft and hard criterion, percentage at rank k, and the Top-1 precision are presented. The best performing method has been submitted by Abdeljalil Gattal and Chawki Djeddi from the Larbi Tebessi University, Department of Mathematics and Computer Science.

[4]https://scriptnet.iit.demokritos.gr/competitions/ - accessed July 2017

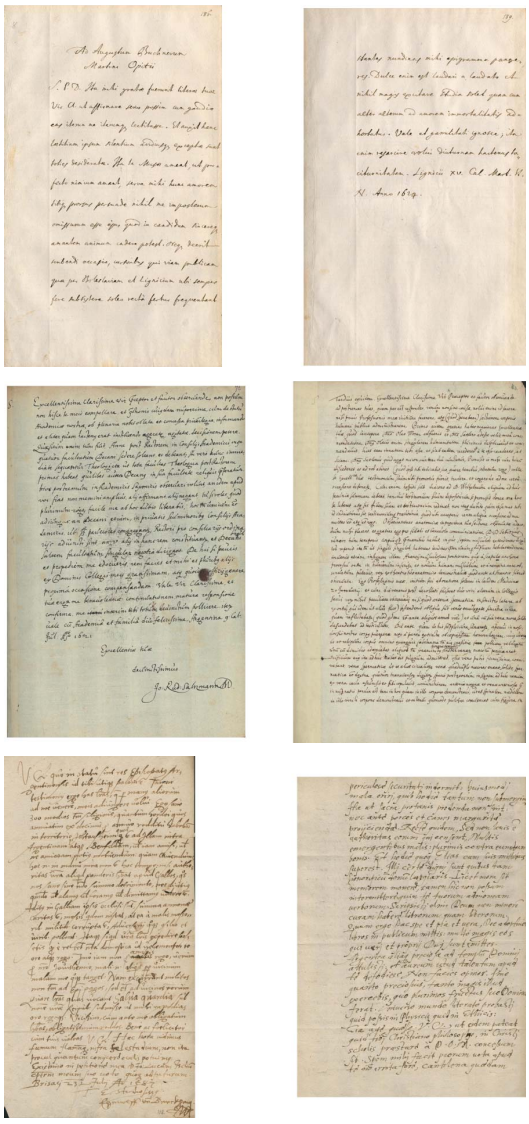

Fig. 5. Six pages of the test set. For the three images on the left no image of the same writer has been found within the first 10 pages of the ranking (referring to all methods submitted). The images on the right are pages written by the same writer.

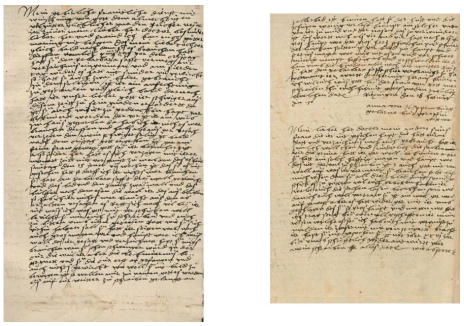

Fig. 6. For the left image all four other pages of the same writer in the dataset have been found by all participants. The right image is one of the other pages.

ACKNOWLEDGMENT

This work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 674943 (project READ). We would like to thank the Universitätsbibliothek Basel for their help.

REFERENCES

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

[2] 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)*, Aug 2013, pp. 1397–1401.

[3] F. Kleber, S. Fiel, M. Diem, and R. Sablatnig, "CVL-DataBase: An Off-Line Database for Writer Retrieval, Writer Identification and Word Spotting," in *2013 12th International Conference on Document Analysis and Recognition*, 2013, pp. 560–564.

[4] S. A. Mahmoud, I. Ahmad, W. G. Al-Khatib, M. Alshayeb, M. Tanvir Parvez, V. Märgner, and G. A. Fink, "KHATT: An Open Arabic Offline Handwritten Text Database," *Pattern Recognition*, vol. 47, no. 3, pp. 1096–1112, mar 2014.

[5] S. Fiel and R. Sablatnig, "Writer Identification and Retrieval Using a Convolutional Neural Network," in *Computer Analysis of Images and Patterns: 16th International Conference, CAIP 2015*. Cham: Springer International Publishing, 2015, pp. 26–37.

[6] Y. Tang and X. Wu, "Text-Independent Writer Identification via CNN Features and Joint Bayesian," in *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, Oct 2016, pp. 566–571.

[7] V. Christlein, D. Bernecker, F. Hönig, A. Maier, and E. Angelopoulou, "Writer Identification Using GMM Supervectors and Exemplar-SVMs," *Pattern Recognition*, vol. 63, pp. 258–267, 2017.

[8] A. Brink, R. Niels, R. van Batenburg, C. van den Heuvel, and L. Schomaker, "Towards robust writer verification by correcting unnatural slant," *Pattern Recognition Letters*, vol. 32, no. 3, pp. 449 – 457, 2011.

[9] A. Nicolaou, A. D. Bagdanov, M. Liwicki, and D. Karatzas, "Sparse radial sampling lbp for writer identification," in *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 2015, pp. 716–720.

[10] J. Wang, Y. song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu, "Learning fine-grained image similarity with deep ranking," *Computer Vision and Pattern Recognition*, 2014.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Computer Vision and Pattern Recognition*, 2015.

[12] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *Computer Vision and Pattern Recognition*, 2014.

[13] S. He and L. Schomaker, "Co-occurrence features for writer identification," in *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, Oct 2016, pp. 78–83.

[14] ——, "Beyond OCR: Multi-faceted understanding of handwritten document characteristics," *Pattern Recognition*, vol. 63, pp. 321 – 333, 2017.

[15] M. Bulacu and L. Schomaker, "Text-Independent Writer Identification and Verification Using Textural and Allographic Features," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 4, pp. 701–717, apr 2007.

[16] T. K. Hussein Mohammed, Volker Maergner and H. S. Stiehl, "Normalised Local Naive Bayes Nearest-Neighbour Classifier for Offline Writer Identification," in *2017 14th International Conference on Document Analysis and Recognition - accepted*, 2017.

[17] A. Gattal, C. Djeddi, Y. Chibani, and I. Siddiqi, "Isolated Handwritten Digit Recognition Using oBIFs and Background Features," in *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*, April 2016, pp. 305–310.

[18] A. J. Newell and L. D. Griffin, "Writer identification using oriented basic image features and the delta encoding," *Pattern Recogn.*, vol. 47, no. 6, pp. 2255–2265, Jun. 2014. [Online]. Available: https://doi.org/10.1016/j.patcog.2013.11.029

[19] V. Christlein, M. Gropp, S. Fiel, and A. Maier, "Unsupervised Feature Learning for Writer Identification and Writer Retrieval," in *2017 14th International Conference on Document Analysis and Recognition (ICDAR) - accepted*, 2017.