A Handwritten French Dataset for Word Spotting - CFRAMUZ

Nikolaos Arvanitopoulos
School of Computer and Communication Sciences (IC)
Ecole Polytechnique Federale de Lausanne (EPFL)
Lausanne, Switzerland
nick.arvanitopoulos@epfl.ch

Gaspard Chevassus
School of Computer and Communication Sciences (IC)
Ecole Polytechnique Federale de Lausanne (EPFL)
Lausanne, Switzerland
chevassusgaspard@gmail.com

Daniele Maggetti
Faculty of Arts
University of Lausanne
Lausanne, Switzerland
Daniel.Maggetti@unil.ch

Sabine Süsstrunk
School of Computer and Communication Sciences (IC)
Ecole Polytechnique Federale de Lausanne (EPFL)
Lausanne, Switzerland
sabine.susstrunk@epf.ch

ABSTRACT

We present a new and freely available dataset, CFRAMUZ, for segmentation-free word spotting research. The dataset consists of seven novels with a total number of 64 pages and 18000 words written in french by the Swiss writer C.F. Ramuz. The novels cover the writer's whole period of life, therefore they show changes in the handwriting style. Together with the complete ground-truth of the dataset we provide an annotation tool. We provide evaluations of state-of-the-art word spotting approaches on this dataset. For completeness we also compare all the approaches on other commonly used datasets to demonstrate the new difficulties and challenges our new dataset introduces.

KEYWORDS

word-spotting, french dataset

ACM Reference Format:
Nikolaos Arvanitopoulos, Gaspard Chevassus, Daniele Maggetti, and Sabine Süsstrunk. 2017. A Handwritten French Dataset for Word Spotting - CFRAMUZ. In *Proceedings of The 4th International Workshop on Historical Document Imaging and Processing, Kyoto, Japan, November 10–11, 2017 (HIP2017),* 6 pages.
https://doi.org/10.1145/3151509.3151523

1 INTRODUCTION

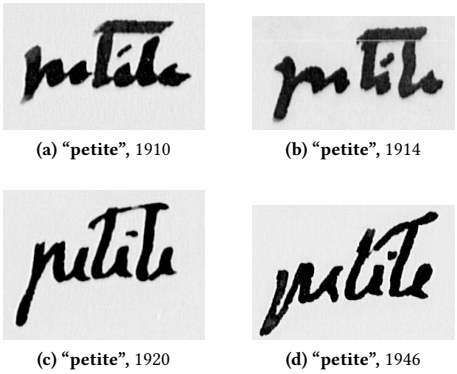
Word spotting is the problem of retrieving instances of a word given as query in a dataset of document pages. It has emerged as a more tractable alternative to word recognition for document indexing. Word spotting does not rely on word annotations, however these are needed to evaluate different techniques. The emergence of word spotting leads to an increased need for challenging datasets with word-level annotations in order to test the accuracy of new or existing approaches.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
HIP2017, November 10–11, 2017, Kyoto, Japan
© 2017 Association for Computing Machinery.
ACM ISBN 978-1-4503-5390-8/17/11…$15.00
https://doi.org/10.1145/3151509.3151523

There are several word spotting datasets available online. The IAM handwriting database [10] contains forms of unconstrained handwritten text written by 657 writers. It is used mainly for word recognition, however it contains box coordinates over words. The IFN/ENIT dataset [12] is a dataset in the Arabic language that can be used for word spotting, even though it targets mainly word recognition applications. Another dataset is the CVL-database [8] containing seven different handwritten texts (one German and six English texts) from 311 different writers. The dataset is suitable for writer retrieval, writer identification and word spotting.

Historical handwritten datasets exist in several languages. A recent historical dataset is the HADARA80P [11], which contains 80 pages from a historical Arabic manuscript together with complete ground-truth for segmentation-free word spotting. Historical datasets exist also in Latin [6] and German [7] and can be partially used for word spotting on line level. However, they do not contain comprehensive ground-truth on word level. One of the most popular historical word spotting datasets is the George Washington dataset [7, 9], which contains 20 pages from a collection of letters from George Washington [1]. It contains bounding boxes for 4894 words in total. The 5CofM dataset [2] contains scanned marriage licenses of the Barcelona Cathedral between 1451 and 1905. The ground-truth contains 50 pages from one volume written by the same writer.

To the best of our knowledge, the only dataset available for the french language is the Rimes dataset [5], which was created to evaluate systems of recognition and indexing of handwritten letters sent by postal mail or fax. Contrary to the non-historical Rimes dataset, our proposed dataset, CFRAMUZ, is based on original historical handwritten text from the beginning of the 20-th century composed in an uncontrolled environment. This property makes it the first historical dataset based on the french language. The texts are written by one author, C.F. Ramuz, and span his entire period of life. On this dataset we observe a significant change in the handwriting style of the author after a specific time period. In Fig. 1 we show an example of the french word "petite". We observe that from 1910 to 1914 the handwriting style of the writer is similar (Figs. 1a, 1b). However, from 1920 the writer changes his style significantly (Figs. 1c, 1d). This significant change in the handwriting style can benefit research that evaluates the handwriting of an individual across time.



(a) “petite”, 1910 (b) “petite”, 1914

(c) “petite”, 1920 (d) “petite”, 1946

Figure 1: Illustration of the different handwriting styles across the dataset. The word “petite” written in the first style in Figs. 1a, 1b and the same word written in the second style in Figs. 1c, 1d.

The dataset contains seven novels written by the author, containing 64 pages with 18027 words in total. The number of unique words is 2998. The ground-truth contains annotated words with bounding boxes and separate files with one-to-one page transcriptions. Together with the dataset we provide an annotation tool that enables ground truth creation. The dataset together with the annotation tool is available online[1].

The rest of the paper is organized as follows. In Section 2 we describe in detail the dataset acquistion process and the ground-truth creation. In Section 3 we provide extensive evaluations of state-of-the-art word spotting approaches on our dataset. For completeness, we provide evaluations on several other commonly used word spotting datasets. Finally, in Section 4 we conclude our work.

2 THE C.F. RAMUZ DATASET

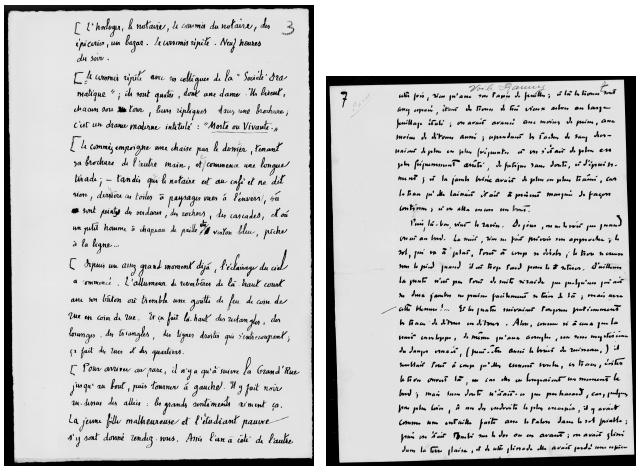
2.1 The dataset

The CFRAMUZ dataset consists of seven novels written by the french-speaking Swiss writer Charles Ferdinand Ramuz (1878-1947). We chose the novels so that they span his entire life of work, from 1910 to 1946. Even though the novels were written by the same writer, we observe a significant change in his handwriting style (see Fig. 1).

C.F. Ramuz was born in the Canton of Vaud and educated in the University of Lausanne. He and an artistic impression of his works appear on the present 200 Swiss franc note. He died in Pully, Switzerland. A complete compilation of all the works of C.F. Ramuz can be found in *Œuvres Complètes* [14]. In Table 1 we show detailed statistics for each novel of the dataset.

In Table 2 we show statistics of the most frequent words in the dataset. In Table 2a we show the top five most frequent words, including punctuation symbols. We see that the most frequent words are prepositions, pronouns and conjunctions. In Table 2b we show the top five most frequent words that are either nouns or verbs. In our dataset, counts, articles and common verbs in third-person (e.g., est, avait, a) are the most frequent.

[1]http://ivrl.epfl.ch/research/handwriting_recognition



(a) Anti-poétique, page 3 (b) La mort du grand Favre, page 7

Figure 2: Two pages from different novels of the CFRAMUZ dataset.

Figure 3: A screenshot from the annotation tool.

2.2 Acquisition

All the works of C.F. Ramuz are scanned in micro-film. From these scans we selected seven novels and transferred them to uncompressed TIFF grayscale images. Two pages from different novels can be seen in Fig. 2. We selected novels of high image quality and simple layout, so that they are suitable for segmentation-free word spotting methods.

2.3 Ground-truth

The novels were annotated and transcribed by literature experts in the works of C.F. Ramuz. The original images were cropped so that they did not contain black borders. The word segmentation was done by the experts using the dedicated annotation tool. Fig. 3 shows a screenshot of the annotation tool used in the ground-truth creation process.

The annotation tool enables the user to create ground-truth data. Features, such as insertion, deletion and modification of word rectangles exist to help the user in her work. Detailed documentation and user manual are available together with the software.

Novels	Year	# Pages	# Words	# Classes
Le petit enterrement	1910	9	2525	686
La Mort du grand Favre	1910	10	2941	807
Mousse	1910	9	2875	793
L'epine dans le doigt	1914	7	1625	535
Adieu à beaucoup de personnages	1914	11	3341	1012
Anti-Poétique	1920	9	2302	716
La cloche qui sonne toute seule	1946	10	2418	712
Style1	[1910 – 1914]	46	13307	2415
Style2	[1920 – 1946]	19	4720	1199
Total		64	18027	2998

Table 1: The novels contained in the CFRAMUZ dataset together with their properties. By classes we denote the number of unique words in each dataset.

Word	# occurrences
,	1424
et	555
de	536
.	527
il	458

(a) Top-five occurred words in the whole dataset.

Word	# occurrences
un	202
plus	128
tout	115
une	115
est	107

(b) Top-five occurred words, excluding prepositions and pronouns.

Table 2: Statistics of the most common words in the dataset.

For each page of the dataset we provide a one-to-one transcription in a text file. The word spotting ground-truth of each page is represented as text and XML files. Each line of the ground-truth file contains the properties of a word in the document page:

- Unique ID for each word
- (x, y) coordinates of the upper left corner of the word rectangle
- width and height of the word rectangle
- line number of the word
- word number in the current line
- UTF-8 word transcription

The first line of each file contains the path of the corresponding document image. This is done in case the user wants to edit the ground-truth with the provided annotation tool in an intermediate stage of the ground-truth creation process. Using the tool, the user can directly load the ground-truth file and the tool will automatically superimpose the ground-truth on top of the file which is denoted on the path.

3 WORD SPOTTING EVALUATION

In this section, we describe the methods used for the experimental evaluation on the CFRAMUZ dataset. We give details on the evaluation process together with results of the methods on other commonly used handwritten word spotting datasets.

3.1 Methods

We use four common word spotting algorithms for our experimental evaluation: *Word Spotting with Embedded Attributes (EAWS)* [4], *Efficient Exemplar Word Spotting (EEWS)* [3], *Bag-of-Visual-Words Word Spotting (BoVWWS)* [15] and *Fisher Kernels Word Spotting (FKWS)* [13]. Let us note here that a direct comparison of segmentation-free and segmentation-based methods may not be precise or even fair, because segmentation-free word spotting is a more difficult problem than segmentation-based word spotting. However, we present the different methods on the same graphs to provide a unified view of their relative performances.

In the following subsections we give a short description of the above mentioned state-of-the-art methods. It is important to note here that there are additional word spotting methods that have shown state-of-the-art results in word-spotting [16, 17]. However, an extensive review and evaluation of state-of-the-art techniques is out of the scope of this paper and is left for future research. In this work we introduce a new dataset that enables the interested researcher to make this type of comparison.

3.1.1 Word Spotting and Recognition with Embedded Attributes (EAWS). In [4] the authors use the notion of embedded attributes. In this word spotting approach words and strings can be compared in a common vectorial subspace. Word labels and word images are embedded in a common subspace. Then word spotting and recognition consist of a simple nearest neighbor problem. Labels and word images are embedded with pyramidal histogram of characters (PHOC) in a d-dimensional space. Words and character images are encoded using Fisher Vectors and these feature vectors are used together with the PHOC labels to learn SVM-based attribute models.

3.1.2 Efficient Exemplar Word Spotting (EEWS). In [3], image documents are divided into cells of equal size and represented by HOG histograms. Queries are represented analogously using cells of the same size in pixels. Then a similarity measure between the

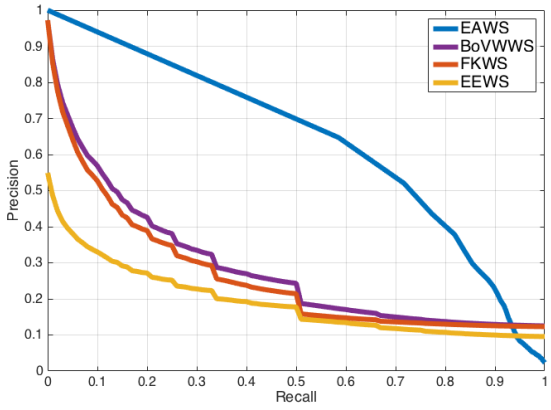


Figure 4: Precision-Recall curves of the state-of-the-art on the CFRAMUZ dataset. EAWS [4] is the most accurate method by a significant margin.

document region and the query using dot product is applied to calculate the scores of document regions and produce a ranking result.

3.1.3 Bag-of-Visual-Words Word Spotting (BoVWWS). In [15], the input image documents are segmented into sub-images using standard segmentation techniques, and then are represented by a sequence of SIFT vectors of 128 dimensions. Then the SIFT vectors of the entire dataset are gathered together and partitioned into a certain number of clusters by K-means. For each word image, the occurrence counts of the SIFT vectors relative to each cluster is calculated. This occurrence vector represents the Bag-of-Visual-Words (BoVW) for the word image.The query image is represented in the same way. Finally the distances between the BoVW of the word images and the query image are computed using cosine similarity.

3.1.4 Fisher Kernels Word Spotting (FKWS). In [13], similar to BoVWWS word spotting, the input image documents are segmented into sub-word images by standard segmentation techniques, and are represented by sequences of SIFT vectors of 128 dimensions. The SIFT vectors of the entire documents are gathered together to learn a Gaussian mixture model of a certain number of clusters. The fisher vectors encode the SIFT vectors of the word images relative to the means, covariances and prior probabilities of the Gaussian Mixture Model. The query image is also represented in the same way as the input word images, and the fisher vector for the query image is computed. Finally, the distances between the fisher vectors of each word image and the query image is computed, and the retrieved result can be obtained by sorting the distances.

3.2 Experimental Results

In this subsection we provide extensive experimental comparisons of the state-of-the-art methods on our dataset, as well as the commonly used datasets George Washington (GW) [7] and Lord Byron (LB) [15].

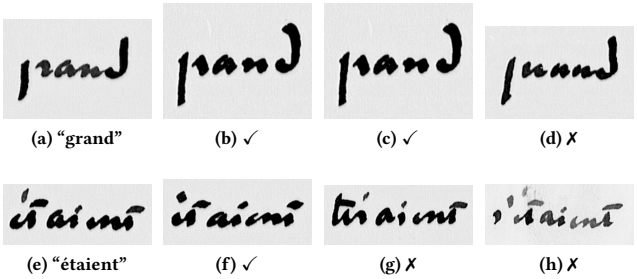


Figure 5: EAWS retrieval results on two queries. On the first line we query the word “grand” and obtain correct results except for Fig. 5d with the similar word “quand”. On the second line we query a more difficult word “étaient”, with retrieval results “étaient”, “tiraient” and “s’étaient”, respectively (Figs. 5f, 5g, 5h).

3.2.1 Evaluation on CFRAMUZ. We randomly split the dataset into 60% training, 20% validation and 20% test set. As queries we used all the word examples in the form of image snippets that belong to the test dataset. The partition setup and sample indices are provided together with the dataset. In Fig. 4 we show precision-recall curves for the compared algorithms on the CFRAMUZ dataset. The best performing method is EAWS [4]. We observe that in the case of EEWS [3] the precision-recall curve does not start from 1. This is due to the fact that this method is segmentation-free and in some query cases (e.g., “.”,“,”,“:”, etc.) the precision is not 1, because the algorithm is not able to find all relevant repetitions of the query. This leads to a significant drop in the accuracy of the algorithm, because these types of queries are very common in our dataset.

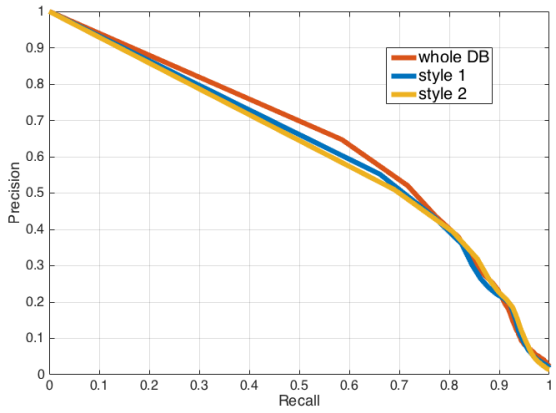
In Fig. 5 we show some qualitative results of EAWS [4] with two different query words, on the complete dataset. Using as query the word “grand” (Fig. 5a) the first two retrieval results are correct (Figs. 5b, 5c), however the third result is the incorrect word “quand” (Fig. 5d). With the word “étaient” (Fig. 5e) the retrieval results are less robust due to existence of many words of similar orthography but different meaning in the dataset. The second and third retrieval results (Figs. 5g, 5h) correspond to the words “tiraient” and “s’étaient”, respectively.

3.2.2 Per-Style Evaluation. In this subsection we split the CFRAMUZ dataset in two groups according to the different handwriting styles and we perform the following experiments:

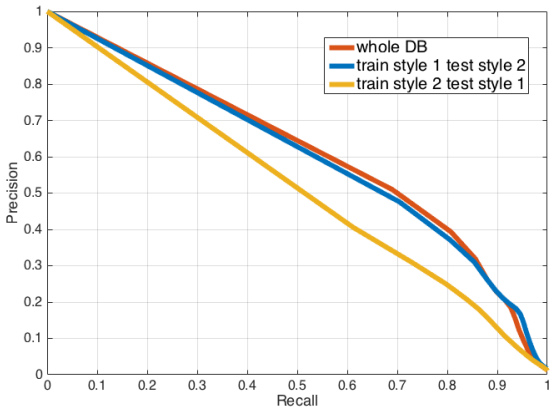
- Training and testing on each style separately.
- Training on style 1 and testing on style 2.
- Training on style 2 and testing on style 1.

The novels that belong to each style are shown in Table 1. For the training and testing on each style separately we use a random split of 60% training, 20% validation and 20% test set. For the different style training procedures we split the data examples that belong to one of the styles into 80% training and 20% validation sets. As queries we used all the word examples from the other style. The specific split for each setup is provided together with the dataset.

We perform these experiments to evaluate the difficulty of each handwriting style. For the experiments we used the best performing



(a) Train and test on each style separately



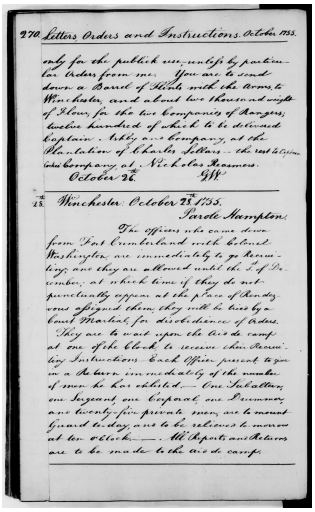
(b) Train and test on different styles

Figure 6: Comparison of EAWS on different styles of the CFRAMUZ dataset. In Fig. 6a we show the accuracy of the algorithm in each style separately. Due to the smaller amount of data in each dataset, the accuracy of the algorithm slightly drops compared to a complete training. In Fig. 6b we train the algorithm on style 1 and test on style 2, and vice versa. We observe that by training on style 2 the algorithm is not able to generalize well on the rest of the data. However, by training only on style 1 the accuracy of the algorithm is almost equivalent as if using the whole dataset for training. Style 1 is more complete with more complex word variations than style 2. By training on style 1, the learning algorithm automatically adapts to the variations of style 2.

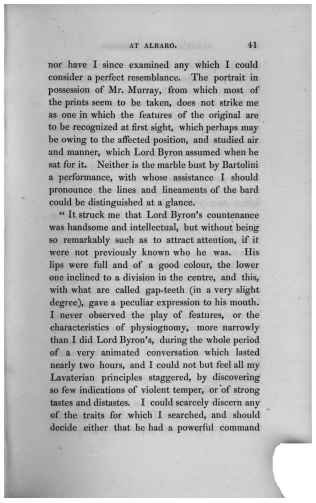
method EAWS [4]. The Precision-Recall curves for the different experiments are shown in Fig. 6. In Fig. 6a we compare the accuracy of EAWS by training in each handwriting style separately. Despite the smaller datasets, we do not observe a significant drop in the accuracy of the algorithm compared to a training experiment on the whole dataset. In Fig. 6b we train EAWS [4] on one handwriting style and test on the other. We observe that by training only on the handwriting style 2 the algorithm is not able to generalize well. The handwriting style contains less data with few variations that are not representative of the complete dataset. On the other hand, by training on handwriting style 1 the algorithm is able to generalize even though it was never trained with data from style 2. Style 1 contains more data examples per word and larger variety. This is an indication that style 1 is more challenging than style 2. The word variations in style 1 are a super-set of the variations in style 2. Therefore, by adapting to style 1, the learning algorithm automatically adapts to style 2.

3.3 Evaluation on other datasets

In this section we compare the results of the previously presented algorithms on the George Washington (GW) [7], Lord Bryon (LB) [15] and on our dataset. The LB dataset consists of 20 printed pages from a book written in 1825 with a total of 4988 words and 1569 word classes. The GW dataset consists of 20 handwritten pages with a total of 4894 words and 1471 word classes. For both datasets, in the case of segmentation-based methods we used the online available experimental setup of EAWS [4][2]. In the case of segmentation-free methods we used the online available experimental setup of EEWS [3] [3]. Two sample images of the two datasets are shown in Fig. 7.



(a) GW



(b) LB

Figure 7: Two pages from the GW and LB datasets, respectively.

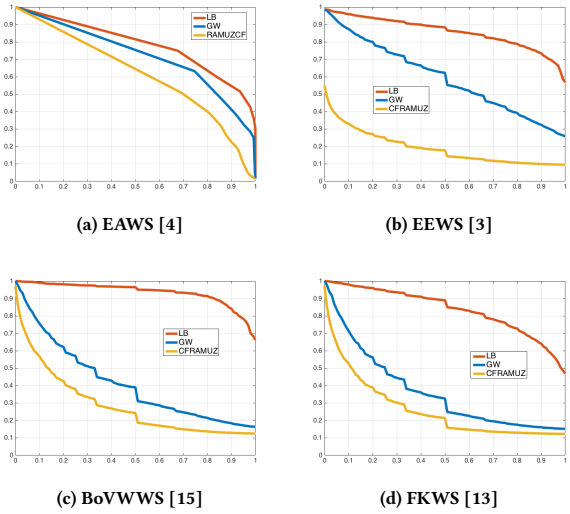
In Fig. 8 we show the precision-recall curves of all the state-of-the-art methods on all datasets. CFRAMUZ is the most challenging dataset. This can be explained by the particularities of the French

[2]http://almazan.github.io/watts/

[3]http://almazan.github.io/ews/

Method	Dataset		
	GW	LB	CFRAMUZ
EAWS	**96.86**	**99.68**	**88.07**
EEWS	50.92	83.60	29.20
BoVWWS	41.09	93.47	50.47
FKWS	36.30	83.44	46.05

Table 3: mean Average Precision (mAP) results of all the tested algorithms on all dataset. EAWS is the better method on all datasets.



(a) EAWS [4] **(b) EEWS [3]**

(c) BoVWWS [15] **(d) FKWS [13]**

Figure 8: Comparison of all the methods on the three handwritten datasets. CFRAMUZ is the most challenging dataset.

language, which gives more variability to our dataset: French contains many groups of words with similar visual features but with different meanings. This characteristic of the language poses several challenges to algorithms that depend heavily on off-the-shelf visual descriptors for image representation. However, more sophisticated descriptors, such as PHOC used in EAWS [4] are partially able to overcome this problem, by taking into account labeled information.

In Table 3 we summarize the mean Average Precision (mAP) results of all the tested methods on all the datasets. As mentioned before, the EAWS [4] algorithm is the better algorithm by a significant margin in all tested datasets. Our dataset is the most challenging one for EAWS [4] and EEWS [3]. The GW dataset is the hardest for the feature-based approaches BoVWWS [15] and FKWS [13]. The LB dataset is the easiest one for all methods, due to the fact that it contains printed text.

4 CONCLUSION

We provide a novel and freely available handwritten dataset for segmentation-free word spotting applications in the French language. To the best of our knowledge, it is the first french historical dataset for word-spotting. The dataset contains works from a single writer through-out his entire life, while exhibiting a significant change of the handwriting style. We present the whole data acquisition and ground-truth creation process. Together with the dataset and its complete ground-truth we provide a simple and intuitive annotation tool for ground-truth creation. Extensive experimental results show that, due to the particularities of the french language, our dataset poses new challenges to state-of-the-art algorithms compared to commonly used English handwritten datasets. Our dataset can benefit research that evaluates handwriting styles of an individual across time, therefore we believe it is a valuable contribution to the community.

REFERENCES

[1] 1741–1799. George Washington Papers at the Library of Congress from 1741-1799. (1741–1799), 270–279,300–209 pages. Letterbook 1.
[2] J. Almazán, D. Fernández, A. Fornés, J. Lladós, and E. Valveny. 2012. A Coarse-to-Fine Approach for Handwritten Word Spotting in Large Scale Historical Documents Collection. In *2012 International Conference on Frontiers in Handwriting Recognition*. 455–460.
[3] Jon Almazán, Albert Gordo, Alicia Fornés, and Ernest Valveny. 2012. Efficient Exemplar Word Spotting. In *Proceedings of the British Machine Vision Conference*. 67.1–67.11.
[4] J. Almazán, A. Gordo, A. Fornés, and E. Valveny. 2014. Word Spotting and Recognition with Embedded Attributes. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36, 12 (Dec 2014), 2552–2566.
[5] Emmanuel Augustin, Jean-marie Brodin, Matthieu Carré, Edouard Geoffrois, Emmanuèle Grosicki, and Françoise Prêteux. 2006. RIMES evaluation campaign for handwritten mail processing. In *Proc. of the Workshop on Frontiers in Handwriting Recognition*.
[6] Andreas Fischer, Volkmar Frinken, Alicia Fornés, and Horst Bunke. 2011. Transcription Alignment of Latin Manuscripts Using Hidden Markov Models. In *Proceedings of the 2011 Workshop on Historical Document Imaging and Processing (HIP '11)*. 29–36.
[7] Andreas Fischer, Andreas Keller, Volkmar Frinken, and Horst Bunke. 2012. Lexicon-free handwritten word spotting using character HMMs. *Pattern Recognition Letters* 33, 7 (2012), 934–942.
[8] F. Kleber, S. Fiel, M. Diem, and R. Sablatnig. 2013. CVL-DataBase: An Off-Line Database for Writer Retrieval, Writer Identification and Word Spotting. In *2013 12th International Conference on Document Analysis and Recognition*. 560–564.
[9] V. Lavrenko, T. M. Rath, and R. Manmatha. 2004. Holistic word recognition for handwritten historical documents. In *First International Workshop on Document Image Analysis for Libraries, 2004. Proceedings*. 278–287.
[10] U.-V. Marti and H. Bunke. 2002. The IAM-database: an English sentence database for offline handwriting recognition. *International Journal on Document Analysis and Recognition* 5, 1 (2002), 39–46.
[11] W. Pantke, M. Dennhardt, D. Fecker, V. Märgner, and T. Fingscheidt. 2014. An Historical Handwritten Arabic Dataset for Segmentation-Free Word Spotting - HADARA80P. In *14th International Conference on Frontiers in Handwriting Recognition*. 15–20.
[12] Mario Pechwitz, Samia Snoussi Maddouri, Volker Märgner, Noureddine Ellouze, and Hamid Amiri. 2002. IFN/ENIT - database of handwritten Arabic words. In *In Proc. of CIFED*. 129–136.
[13] F. Perronnin and J. A. Rodriguez-Serrano. 2009. Fisher Kernels for Handwritten Word-spotting. In *2009 10th International Conference on Document Analysis and Recognition*. 106–110.
[14] Charles Ferdinand Ramuz. [n. d.]. *Œuvres Complètes*. Editions Slatkine.
[15] M. Rusinol, D. Aldavert, R. Toledo, and J. Llados. 2011. Browsing Heterogeneous Document Collections by a Segmentation-Free Word Spotting Method. In *2011 International Conference on Document Analysis and Recognition*. 63–67.
[16] S. Sudholt and G. A. Fink. 2016. PHOCNet: A Deep Convolutional Neural Network for Word Spotting in Handwritten Documents. In *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*. 277–282.
[17] Z. Zhong, W. Pan, L. Jin, H. Mouchère, and C. Viard-Gaudin. 2016. SpottingNet: Learning the Similarity of Word Images with Convolutional Neural Network for Word Spotting in Handwritten Historical Documents. In *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*. 295–300.