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 Süustrunk
School of Computer and Communication Sciences (IC)
Ecole Polytechnique Federale de Lausanne (EPFL)
Lausanne, Switzerland
sabine.susstrunk@epfl.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:
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
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.

The rest of the paper is organized as follows. In Section 2 we describe in detail the dataset acquisition 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.

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.
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.

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

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

<table>
<thead>
<tr>
<th>Word</th>
<th># occurrences</th>
</tr>
</thead>
<tbody>
<tr>
<td>.</td>
<td>1424</td>
</tr>
<tr>
<td>et</td>
<td>555</td>
</tr>
<tr>
<td>de</td>
<td>536</td>
</tr>
<tr>
<td>.</td>
<td>527</td>
</tr>
<tr>
<td>il</td>
<td>458</td>
</tr>
<tr>
<td>un</td>
<td>202</td>
</tr>
<tr>
<td>plus</td>
<td>128</td>
</tr>
<tr>
<td>tout</td>
<td>115</td>
</tr>
<tr>
<td>une</td>
<td>115</td>
</tr>
<tr>
<td>est</td>
<td>107</td>
</tr>
</tbody>
</table>

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
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].

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”. On the second line we query a more difficult word “étaint”, with retrieval results “étaint”, “tiraient” and “s’étaient”, respectively (Figs. 5f, 5g, 5h).

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
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 Byron (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.

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

![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.](image-url)

![Figure 7: Two pages from the GW and LB datasets, respectively.](image-url)
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 throughout 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


<table>
<thead>
<tr>
<th>Method</th>
<th>Dataset</th>
<th>GW</th>
<th>LB</th>
<th>CFRAMUZ</th>
</tr>
</thead>
<tbody>
<tr>
<td>EAWS</td>
<td>96.86</td>
<td>99.68</td>
<td>88.07</td>
<td></td>
</tr>
<tr>
<td>EEWS</td>
<td>50.92</td>
<td>83.60</td>
<td>29.20</td>
<td></td>
</tr>
<tr>
<td>BoVWWS</td>
<td>40.91</td>
<td>93.47</td>
<td>50.47</td>
<td></td>
</tr>
<tr>
<td>FKWS</td>
<td>36.30</td>
<td>83.44</td>
<td>46.05</td>
<td></td>
</tr>
</tbody>
</table>

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