The Pinkas Dataset

Berat Kurar Barakat*, Jihad El-Sana
Department of Computer Science
Ben-Gurion University of the Negev
berat@post.bgu.ac.il, el-sana@cs.bgu.ac.il

Irina Rabaev*
Software Engineering Department
Shamoon College of Engineering
irinar@ac.sce.ac.il

***Abstract*—In historical document image processing, datasets account for a significant part of any research, and are crucial for the diversity and abundance of experimental results, which contribute to the development of new algorithms to meet the new challenge. Moreover, they are very important for benchmarking processing algorithms. Numerous publicly available document image datasets of different languages have been emerged. However, current segmentation and recognition performances are nearly saturated with respect to the present publicly available datasets. As such, collecting and labelling historical document images is a burden on historical document image processing researchers. This paper introduces a public historical document image dataset, Pinkas dataset, with new challenges to open room for improvement and identify strengths and weaknesses of available processing algorithms. It is the first dataset in medieval handwritten Hebrew and fully labeled at word, line and page level by an expert of historical Hebrew manuscripts. Pinkas dataset contributes to the diversity of benchmarking standards. In this paper we present meta features of Pinkas dataset and apply recent word spotting algorithms to analyze the room for improvement in terms of performance.**
The full dataset is available for download at:
https://www.cs.bgu.ac.il/~berat

I. Introduction

Benchmark datasets with accompanied ground truth are of tremendous importance both for evaluation, analysis and comparison of algorithms and methods. Moreover, benchmark datasets contribute to the development of new algorithms to meet the new challenges.

In the recent decades a number of historical document datasets have been introduced. Among them are the datasets of Latin script, George Washington (GW) [1], Parzival [2], Saint Gall datasets [3], DIVA-HisDB [4], and datasets of Arabic script [5]–[7]. However, to the best of our knowledge, there is no publicly available annotated dataset of historical Hebrew documents.

In this paper we introduce Pinkas dataset of a medieval manuscript in Hebrew. It is the first dataset in medieval handwritten Hebrew which is fully annotated at word, line and page level by an expert of historical Hebrew manuscripts. The Pinkas dataset exhibits complex layouts and numerous degradation types: bleed-through, stains, uneven and faded ink, etc. (Figure 1 and Figure 2). The documents were written by different writers, who usually were not professional scribers. As a result, letter shapes are often vary inside one page or even one paragraph. This presents additional challenge for automatic processing of the documents.

The contribution of this paper is twofold. First, a 30 pages manuscript together with its ground truth at page, line and word levels is presented. The documents were annotated by a Hebrew paleographer. The presented Pinkas dataset will allow to transform research for analyzing Hebrew handwriting. Moreover, such dataset will also be useful to prove the generalizability and robustness of document processing algorithms. Second, experiments were run with three methods for word spotting, PHOCNet [8], siamese CNN [9], and exemplar SVM [10]. The results confirm challenging nature of the dataset and can serve as a benchmark for future studies.

The remainder of the paper is structured as follows. Section II reviews historical handwritten datasets developed during the last two decades. Section III gives a detailed insight on the Pinkas dataset, and Section IV overviews the annotation of the dataset. Section V defines an official partition for Pinkas dataset. Experimental results of word spotting methods and their comparative analysis are presented in Section VI. Finally, Section VII highlights conclusions and presents our future plans.

II. Review of existing datasets

Publicly available benchmark datasets provide a platform for evaluation and fair comparison of different methods. A number of historical document datasets supporting the evaluation of segmentation, word spotting and recognition tasks has been introduced in the recent decades.

DIVA-HisDB [4] consists of three medieval manuscripts with a total of 150 pages. The dataset provides a benchmark for layout analysis, text line segmentation, binarization and writer identification. The IAM Historical Document Database contains GW, Parzival and Saint Gall datasets of historical documents. The GW dataset [1] contains 20 pages from George Washington Papers collection, providing ground truth at page, text line and word levels. Parzival dataset [2] contains 47 pages of medieval German manuscript dated to 13th century. Saint Gall dataset [3] contains a 60 pages historical manuscript written in Latin by a single writer in the 9th century. Both Parzival and Saint Gall datasets provide ground truth at text line and word levels.

VML database [5] consists of five books with a total of 680 pages written in Arabic by five writers during the 11th to 15th

*These authors contributed equally to this work.

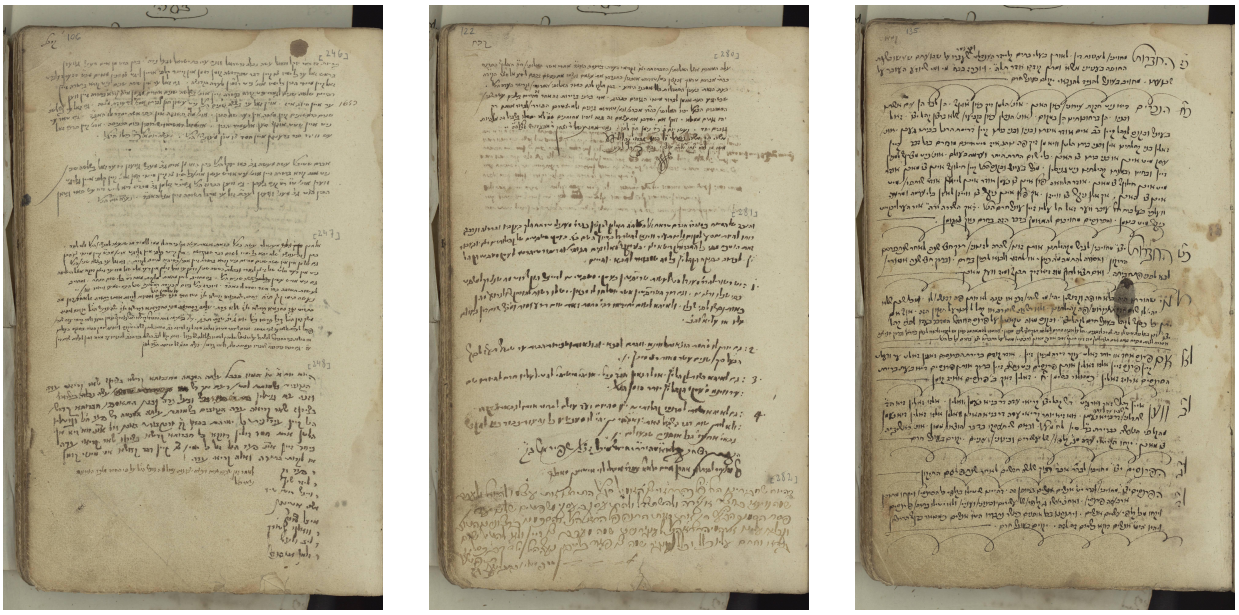


Fig. 1. Sample document images from the Pinkas dataset. Paragraphs are separated by drawings or by space. Some paragraphs are assigned by a number which is written in a spatial proximity to them.

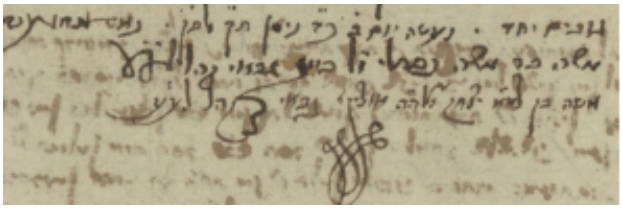
centuries. The ground truth contains transcriptions at word level. WAHD [6] is a database only for writer identification of Arabic historical documents, whereas [7] is a text line segmentation dataset for challenging handwritten documents.

Pinkas dataset is related to the both, segmentation and word spotting datasets. It has two main distinctions though, in comparison to other datasets. First, it is the first historical dataset in handwritten Hebrew. Second, the Pinkas contains more heterogeneous document images, since it was written by numerous scribers.

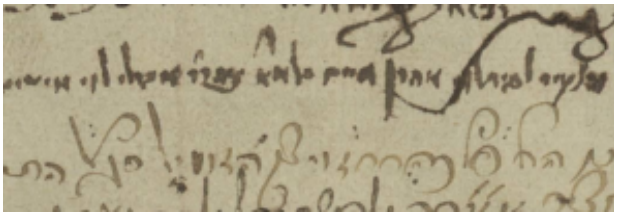
III. Description of the Pinkas Dataset

Pinkas dataset is created from a historical Hebrew manuscript that contains records of Jewish communities in Europe in the early modern period (c. 1500-1800). The Hebrew term for such records is pinkas and its plural is pinkassim. These records register the ways in which the Jewish community organized its social, economic, religious, cultural and even family life. They contain the results of annual elections for the executive board and sub-committees, appointments of rabbis and doctors, marriage, death and burial registers, expenses and incomes, memory of historical events, and even communal conflicts. Obviously, pinkassim are an invaluable source for learning about Jewish life, culture and development, and in some cases even for tracing life paths of individual members of the community. The presented dataset contains manuscripts from the records of Frankfurt community. Frankfurt Jewish community was one of the biggest and important Jewish communities in Germany, and providing a searchable database of these manuscripts would undoubtedly lead to a breakthrough in the research.

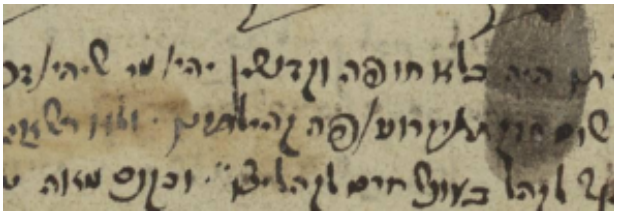
The dataset consists of 30 pages digitized by full color digital images in JPG format with high resolution. The pages



(a) Bleed through



(b) Faded and uneven ink



(c) Stains

Fig. 2. Some degradation examples from the Pinkas dataset.

exhibit numerous degradations, complex layout and different handwritings. They are written in a mixture of Medieval Hebrew and did not follow any fixed spelling rules. Any word was written to the ear as it seemed convenient or correct to the scriber. One can often see two or three variations of the spelling of the same word within one line. The rules of

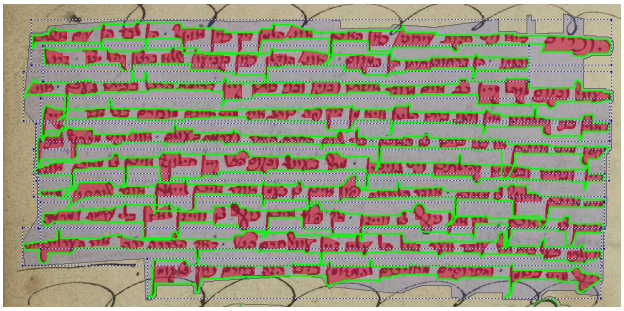


Fig. 3. Main text (purple), line (green) and word (red) segmentation levels.

grammar were also not scrupulously respected. In addition, usually, the writers were not professional scribers. This adds additional challenge, since very often the same letters are written in different shapes. Only a few number of scholars in the world specialize in reading these texts.

Figure 1 and Figure 2 present samples from the dataset. We can notice severe bleed through, stains and uneven ink even on the same page. Some of the pages contain decorated letters and drawings to separate paragraphs, others use space for separation. Sometimes marginal notes interleave with the main text. There are a variety of handwritings, variable line curvatures, and different letter sizes and shapes. As we can see, this manuscript is very challenging both from computer and human processing points of view.

IV. Annotation of the Pinkas Dataset

The manuscript was annotated at page, text line and word levels using semi-automated tools of Aletheia system [11]. Figure 3 illustrates the segmentation: main text is shown in purple, lines in green and words in red. The ground truth is available in PAGE [12] format. After an initial annotation, all pages were corrected precisely by a Hebrew paleographer.

Page level segmentation determines the main text, side text, signature-marks and dates. Main text and side text regions are based on spatial features. We can see examples of main and side text segmentation in Figure 4a. Decorated letters and words are also considered as a part of a main text, since their size only slightly exceed the size of the main text. An example of decorated word is presented in Figure 4a (the top right corner). Some of the paragraphs are enumerated, i.e. a number is assigned to each paragraph (Figure 1). These enumerations are kept together with the corresponding main text region, as they logically belong to it and are written in the spatial proximity to the main text. Signatures are segmented as a separate page level class (Figure 4b). Paragraph separator drawings, as those present in the rightmost document of Figure 1, are discarded at this stage. In the future work, these features could be included. The stated date (in Arabic digits), which sometime appear at the margins of a page, are annotated as an additional page level class. Details of region class distributions are summarized in Table I.

Text line level segmentation determines the text lines within the main and side text areas. Text line annotations contain all pixels of the text line, including overlapping text line parts, as shown in Figure 5a.

Word level segmentation determines the words of a text line. Words are separated based on their spatial, semantic and verbatim features within a context. Very often only a professional transcriber is able to recognize the boundaries of each word, since in many cases there is no space separation between them, as exemplified in Figure 5b.

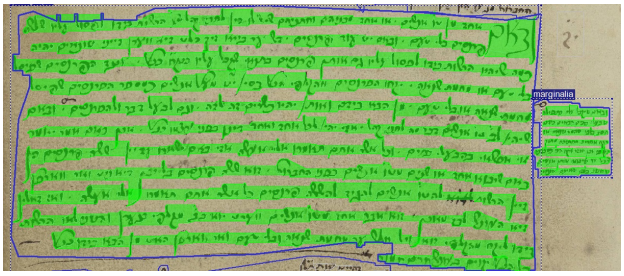
The total amount of text lines, words and word classes is summarized in Table II.

TABLE I
Pinkas: Number of regions per category.

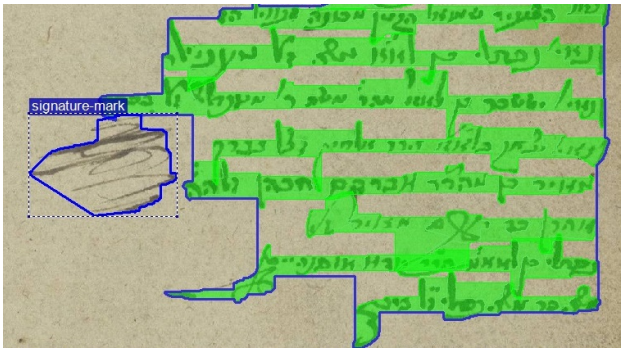
Main text	Side text	Signature-marks	Dates
108	7	13	11

TABLE II
Pinkas: Total amount of lines, words and word classes.

Lines	Words	Word classes
1013	13744	3387



(a) Main and side texts; a decorated word at the top right corner of the main text.



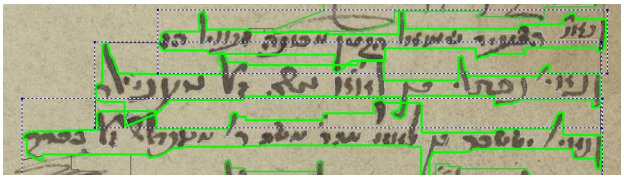
(b) Signature mark.

Fig. 4. Page level segmentation classes.

V. Train and Test Split

Lack of official partition of a dataset leads to incomparable results as is common with GW dataset [10]. Therefore, for the sake of comparable and fair results we define an official train and test set partition.

In the real scenario of historical document image analysis, a scholar would annotate the words of a manuscript in their appearance order. Hence, he would annotate page after page



(a) Segmentation of overlapping lines

(b) Word segmentation. Note that there is no space separation between the fifth and sixth words.

Fig. 5. Text line and word segmentations.

without considering whether all possible word classes exist in the train set or not. On the contrary, uniform or nearly uniform data division is crucial to the performance of machine learning algorithms [13]. Therefore, historical document image analysis literature tends to partition the data nearly uniformly in word level [8], [10], [14]–[16] with some exceptions [9], [17]. Uniform partitioning is optimal from machine learning point of view but is not a realistic case in historical document image analysis field.

Consequently, we choose to partition the dataset at page level, 80% for training and 20% for testing. Words in the first 24 pages are in the train set and words in the following 6 pages are in the test set. Table III presents the number of classes and samples in the train and test partitions of the Pinkas dataset in comparison to first cross-validation split of GW dataset [15]. Notice that 34% of the classes in GW test set are out of vocabulary (OOV) whereas 49% of the classes in Pinkas test set are OOV. This shows that dealing with nonuniform data split should be inherent to the historical document image processing algorithms.

TABLE III
STATISTICS OF PINKAS DATASET TRAIN AND TEST PARTITION IN COMPARISON TO GW DATASET.

	Train		Test		
Dataset	**Classes**	**Samples**	**Classes**	**Samples**	**OOV**
Pinkas	3117	10397	1251	3278	603
GW	966	3645	471	1215	160

VI. WORD SPOTTING EXPERIMENTS

We applied two supervised segmentation based, and one unsupervised segmentation free word spotting methods for evaluating the difficulty of recognizing the words in the Pinkas dataset. Segmentation based methods are siamese CNN and PHOCNet, segmentation free method is exemplar SVM. In all the experiments we used the official partition described in section V. Table IV summarizes the baseline results on Pinkas dataset.

A. Siamese CNN

Siamese CNN [18] contains two branches that share the same CNN architecture and the same weights. The input is a pair of word images and the output is a similarity rank of the input pair. We first based the branches of the siamese CNN on the architecture of Shi et al. [19]. Then through experiments we tune the hyperparameters to fit our task. Figure 6 shows the final architecture.

Each CNN branch has seven convolutional layers. Dotted lines indicate identical weights. Numbers in parentheses are number of filters, filter size and stride. All convolutional and fully connected layers are followed by ReLU activation functions except fc2 which feeds into a sigmoid function using binary cross entropy loss.

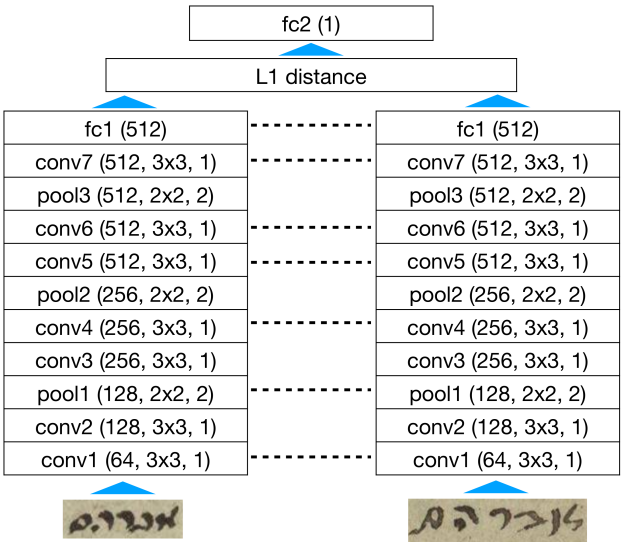


Fig. 6. Siamese CNN for word spotting. Dotted lines stand for shared weights, conv stands for convolutional layer, fc stands for fully connected layer and pool is a max pooling layer.

Number of all possible same word image pairs is $75,258$ and different word image pairs is $44,561,818$. The difficulty with such a large dataset is that it is impossible to train the algorithm on all same and different pairs. We first balance the train set using augmentation. Then from each class 100 same pairs and an equal number of different pairs are generated. After a certain point of learning most pairs are correctly classified and using them no longer improved the performance. Hence, following the idea in [20] we mine the pairs with the largest loss and recreate the different pairs from the word classes in these hard negative pairs. Pairs with large loss correspond to word classes that are hard to discriminate. We continue training using the new hard pairs and reached to mAP value of 61.5%.

TABLE IV
MAP RESULTS OF WORD SPOTTING METHODS ON THE PINKAS DATASET.

Dataset	**Siamese CNN**	**PHOCNet**	**PHOCNet One hot**	**Exemplar SVM**
Pinkas	**61.5**	56.6	53.3	1.5

B. PHOCNet

PHOCNet [8] is a CNN that is trained with the Pyramidal Histogram of Characters (PHOC) [21] as labels of the word images. It is a state of the art word spotting method and authors published its source code. Using authors' published code, we trained PHOCNet using 4 levels $[2, 3, 4, 5]$ PHOC representation of unigrams. There are 43 unigrams, which lead to final phoc size of 602. Training with this configuration for $80,000$ iterations reached to mAP value of 56.6%.

At its first level, PHOC label embedding is not word discriminative [21]. Words such as "listen" and "silent" share the same representation. Therefore, commonly its pyramid version is used. Pyramid version can represent most of the character order information but not fully. We run an experiment with one hot encoding of the words. One hot encoding is commonly used for representing biological sequences [22] and can fully represent the character order information. It assumes a maximum possible word length, which is 13 in Pinkas dataset, and encodes characters as binary sparse vectors (Figure 7). Training with this configuration reached to mAP value of 53.3%.

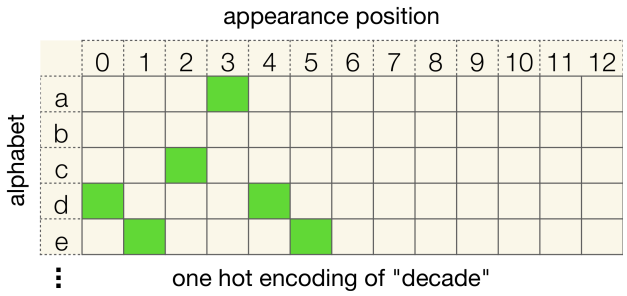


Fig. 7. One hot encoding.

C. Exemplar SVM

Exemplar SVM [10] is an unsupervised segmentation free method for word spotting in document images. Documents are represented with a grid of HOG descriptors, and a sliding window approach is used to locate the document regions that are most similar to the query. Using the published source code by its authors, this method achieved a mAP value of 1.5%.

VII. Conclusion and Future Work

Research in historical document analysis is a challenging problem. Benchmark datasets lie at the heart of development, assessment and comparison of the algorithms. This paper introduces the Pinkas dataset, a historical Hebrew handwritten dataset, which provides numerous challenges for historical document image analysis. It is the first dataset in medieval handwritten Hebrew and fully labeled at word, line and page level.

The dataset contains 30 historical document images together with their page, line and word levels segmentation ground truths. Moreover an official train and test set partition is defined and three word spotting methods are used to set the baselines. Results show that there is a big room for improvement and confirm the challenging nature of the dataset. The Pinkas dataset contributes to the diversity of benchmarking standards and is available for download at: https://www.cs.bgu.ac.il/~berat

In future, we plan to run baseline experiments for page segmentation and text line segmentation of the Pinkas dataset. We also plan to extend the dataset and publish further baseline results.

https://github.com/ssudholt/phocnet
http://almazan.github.io/ews/

Acknowledgment

Authors would like to thank Daria Vasyutinsky Shapira for proof reading and Gunes Cevik for segmenting the dataset. This research was supported in part by Frankel Center for Computer Science at Ben-Gurion University of the Negev.

References

[1] T. M. Rath and R. Manmatha, "Word spotting for historical documents," *International Journal of Document Analysis and Recognition (IJDAR)*, vol. 9, no. 2-4, pp. 139–152, 2007.

[2] A. Fischer, M. Wuthrich, M. Liwicki, V. Frinken, H. Bunke, G. Viehhauser, and M. Stolz, "Automatic transcription of handwritten medieval documents," in *2009 15th International Conference on Virtual Systems and Multimedia*. IEEE, 2009, pp. 137–142.

[3] A. Fischer, V. Frinken, A. Fornés, and H. Bunke, "Transcription alignment of latin manuscripts using hidden markov models," in *Proceedings of the 2011 Workshop on Historical Document Imaging and Processing*. ACM, 2011, pp. 29–36.

[4] F. Simistira, M. Seuret, N. Eichenberger, A. Garz, M. Liwicki, and R. Ingold, "Diva-hisdb: A precisely annotated large dataset of challenging medieval manuscripts," in *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*. IEEE, 2016, pp. 471–476.

[5] M. Kassis, A. Abdalhaleem, A. Droby, R. Alaasam, and J. El-Sana, "Vml-hd: The historical arabic documents dataset for recognition systems," in *2017 1st International Workshop on Arabic Script Analysis and Recognition (ASAR)*. IEEE, 2017, pp. 11–14.

[6] A. Abdelhaleem, A. Droby, A. Asi, M. Kassis, R. Al Asam, and J. Elsanaa, "Wahd: a database for writer identification of arabic historical documents," in *2017 1st International Workshop on Arabic Script Analysis and Recognition (ASAR)*. IEEE, 2017, pp. 64–68.

[7] B. Kurar Barakat, A. Droby, M. Kassis, and J. El-Sana, "Text line segmentation for challenging handwritten document images using fully convolutional network," in *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*. IEEE, 2018, pp. 374–379.

[8] S. Sudholt and G. A. Fink, "Phocnet: A deep convolutional neural network for word spotting in handwritten documents," in *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*. IEEE, 2016, pp. 277–282.

[9] B. Kurar Barakat, R. Alasam, and J. El-Sana, "Word spotting using convolutional siamese network," in *2018 13th IAPR International Workshop on Document Analysis Systems (DAS)*. IEEE, 2018, pp. 229–234.

[10] J. Almazán, A. Gordo, A. Fornés, and E. Valveny, "Efficient exemplar word spotting." in *Bmvc*, vol. 1, no. 2, 2012, p. 3.

[11] C. Clausner, S. Pletschacher, and A. Antonacopoulos, "Aletheia-an advanced document layout and text ground-truthing system for production environments," in *2011 International Conference on Document Analysis and Recognition*. IEEE, 2011, pp. 48–52.

[12] S. Pletschacher and A. Antonacopoulos, "The page (page analysis and ground-truth elements) format framework," in *2010 20th International Conference on Pattern Recognition*. IEEE, 2010, pp. 257–260.

[13] P. S. Crowther and R. J. Cox, "A method for optimal division of data sets for use in neural networks," in *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*. Springer, 2005, pp. 1–7.

[14] J. A. Rodríguez-Serrano and F. Perronnin, "A model-based sequence similarity with application to handwritten word spotting," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 11, pp. 2108–2120, 2012.

[15] J. Almazán, A. Gordo, A. Fornés, and E. Valveny, "Handwritten word spotting with corrected attributes," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 1017–1024.

[16] T. Wilkinson and A. Brun, "Semantic and verbatim word spotting using deep neural networks," in *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*. IEEE, 2016, pp. 307–312.

[17] D. Aldavert, M. Rusinol, R. Toledo, and J. Lladós, "Integrating visual and textual cues for query-by-string word spotting," in *2013 12th International Conference on Document Analysis and Recognition*. IEEE, 2013, pp. 511–515.

[18] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature verification using a" siamese" time delay neural network," in *Advances in neural information processing systems*, 1994, pp. 737–744.

[19] B. Shi, X. Bai, and C. Yao, "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 11, pp. 2298–2304, 2016.

[20] E. Simo-Serra, E. Trulls, L. Ferraz, I. Kokkinos, P. Fua, and F. Moreno-Noguer, "Discriminative learning of deep convolutional feature point descriptors," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 118–126.

[21] J. Almazán, A. Gordo, A. Fornés, and E. Valveny, "Word spotting and recognition with embedded attributes," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 12, pp. 2552–2566, 2014.

[22] C. Angermueller, T. Pärnamaa, L. Parts, and O. Stegle, "Deep learning for computational biology," *Molecular systems biology*, vol. 12, no. 7, p. 878, 2016.