ICFHR2016 Handwritten Keyword Spotting Competition (H-KWS 2016)

Ioannis Pratikakis, Konstantinos Zagoris
Visual Computing Group
Department of Electrical & Computer Engineering
Democritus University of Thrace
Xanthi, Greece
{ipratika,kzagoris}@ee.duth.gr

Basilis Gatos
Institute of Informatics and Telecommunications
National Center for Scientific Research "Demokritos"
Athens, Greece
bgat@iit.demokritos.gr

Joan Puigcerver, Alejandro H. Toselli, Enrique Vidal
Pattern Recognition and Human Language Technology Research Center
Universitat Politècnica de València - Valencia, Spain
{joapuipe,ahector,evidal}@prhlt.upv.es

***Abstract*—The H-KWS 2016, organized in the context of the ICFHR 2016 conference aims at setting up an evaluation framework for benchmarking handwritten keyword spotting (KWS) examining both the Query by Example (QbE) and the Query by String (QbS) approaches. Both KWS approaches were hosted into two different tracks, which in turn were split into two distinct challenges, namely, a segmentation-based and a segmentation-free to accommodate different perspectives adopted by researchers in the KWS field. In addition, the competition aims to evaluate the submitted training-based methods under different amounts of training data. Four participants submitted at least one solution to one of the challenges, according to the capabilities and/or restrictions of their systems. The data used in the competition consisted of historical German and English documents with their own characteristics and complexities. This paper presents the details of the competition, including the data, evaluation metrics and results of the best run of each participating methods.**

***Keywords*-Query by String, Query by Example, Keyword Spotting, ICFHR'16 Contest, Evaluation Metrics**

I. Introduction

Handwritten keyword spotting is the task of detecting query words in handwritten document image collections without involving a traditional OCR step. Recently, handwritten word spotting has attracted the attention of the research community in the field of document image analysis and recognition since it has been proved a feasible solution for indexing and retrieval of handwritten documents in the case that OCR-based methods fail to deliver proper results.

This competition is a joint effort between the organizers of ICFHR 2014 H-KWS Competition [1] and the ICDAR2015 Competition on KWS[2] aiming to set up a common evaluation framework for benchmarking the two distinct variations for keyword spotting, namely the Query by Example (QbE) and the Query by String (QbS) case.

Clearly each of these variations of the KWS problem statement has its own degree of difficulty and application targets. For instance, QbS is mandatory for applications involving large-scale handwritten image indexing and search under the precision-recall trade-off model. In this case, given the scale, it can be very advantageous to use training-based KWS. Other kind of applications involve assisting human transcribers by allowing them to find words in a document which have a shape similar to a word or part of a word (perhaps one which the transcriber is not sure how to transcribe when it appears for the first time). In such applications, a training-free QbE system is more appropriate.

Although QbS and QbE address fundamental different problems they are both unified at the technical level since they may both either have dependencies of segmentation (segmentation-based) or not (segmentation-free) and they may both either involve training of data (supervised) or not (unsupervised). All alternatives will be examined in the proposed competition which makes it different compared to previously organized efforts.

The taxonomy and characteristics of the different tracks and challenges in the competition are shown below:

1) Track-I: Query by Example
   a) Challenge I.A: Segmentation-based
   b) Challenge I.B: Segmentation-free
2) Track-II: Query by String
   a) Challenge II.A: Segmentation-based
   b) Challenge II.B: Segmentation-free

Finally, unlike previous editions, the aim of this competition is twofold: to evaluate all the major KWS flavors using an unique evaluation protocol and assessment measures, and to compare the different participating methods under different amounts of training data and data from different languages. The purpose of the latter distinction is to clearly understand the data requirements of each method and their applicability to different languages.

II. Datasets

The proposed datasets consist of a series of documents from two different collections prepared in the European

2167-6445/16 $31.00 © 2016 IEEE
DOI 10.1109/ICFHR.2016.109

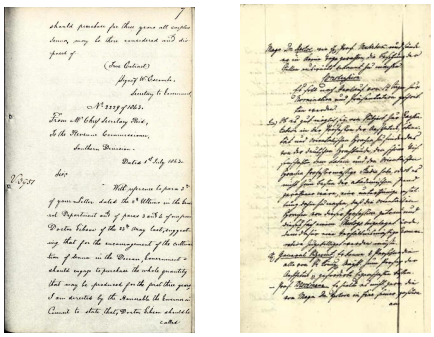

Figure 1: Examples of two document page images from the Botany (left) and Konzilsprotokolle (right) collections.

project READ[1]: the *Alvermann Konzilsprotokolle* and the *Botany* in British India collections. The former, in good preservation state, belongs to the University Archives Greifswald and involves around 18 000 pages. This collection contains fair copies of the minutes, written during the formal meetings held by the central administration between the years 1794-1797. The documents belong to the University Archives and were digitized and provided by the University Library in Greifswald. Transcripts were provided by the University Archives (Dirk Alvermann). On the other hand, the *Botany*[2] in British India is from the India Office Records and provided by the British Library. This collection covers the following topics: botanical gardens; botanical collecting; useful plants (economic and medicinal). Fig.1 shows an example page from each dataset.

For each collection, several training set partitions were released sequentially in order to evaluate the competing systems under different amounts of available training data. For each partition, the set of page images and two XML files, containing the word-level and line-level transcription and segmentation, were given. However, only three pages from the first training partition of each dataset were manually segmented at a word-level. The word-level bounding boxes of the remaining training pages were obtained by means of Viterbi forced alignment using the line-level segmentation, which was performed manually by human operators.

Each test dataset comprises 20 pages wherein the bounding boxes of all words were manually obtained.

The query set of each dataset is provided in UTF-8 plain text format (QbS) and word image queries (QbE) of various length and frequency. 150 and 200 different words were manually selected for the *Botany* and the *Konzilsprotokolle* datasets, respectively. Fig. 2 shows the frequency and the query length distribution for each query set.

All data used in the competion, including transcriptions and evaluation ground-truth for KWS, was released after the competition and it is available through the competition's webpage[3].

[1]http://read.transkribus.eu

[2]http://www.bl.uk/reshelp/findhelpregion/asia/india/indiaofficerecords/botanymat.html

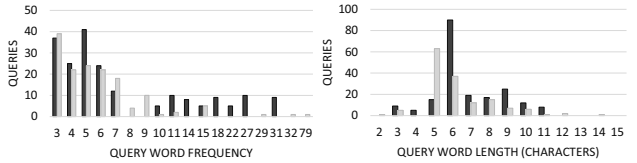

Figure 2: Query word statistics for Botany (light) and Konzilsprotokolle (dark): frequency (left) and length (right).

Table I: Details on the number of pages, lines and words on each partition of the training data and the test data.

			Botany	Konzilsprotokolle
Train	I	Pages	10	10
		Lines	275	263
		Words	1 684	1 849
	II	Pages	30	30
		Lines	676	824
		Words	3 611	5 968
	III	Pages	114	45
		Lines	2 961	1 235
		Words	16 686	9 102
Test		Pages	20	20
		Lines	607	524
		Words	3 318	3 891

III. Evaluation metrics

Mean average precision (mAP) was used to evaluate the solution of each participant. For each query in the set Q, its (interpolated) average precision was computed using the (interpolated) precision-at-top-k, $\pi(k)$ ($\hat{\pi}(k)$), and the recall-at-top-k, $\rho(k)$. Eq. 1 defines the (interpolated) precision and recall scores of the top-k results, using the set of all relevant items R, and the set of top-k results in the solution $S(k)$; specifies the interpolated average precision expression, AP, where $\Delta\rho(k)$ is the difference in recall between items k and $k-1$; and defines the mAP metric, from the AP of each query q, AP(q).

$$\pi(k) = \frac{R \cap S(k)}{S(k)};\;\; \rho(k) = \frac{R \cap S(k)}{R};\;\; \hat{\pi}(k) = \max_{j:\rho(j)\geq\rho(k)} \pi(j)$$

$$\text{AP} = \sum_{k=1}^{n} \hat{\pi}(k) \cdot \Delta\rho(k); \qquad \text{mAP} = \frac{\sum_{q\in Q} \text{AP}(q)}{|Q|} \tag{1}$$

In segmentation-free challenge, a detected bounding box may not match exactly with the references. Thus, a detected box is considered a correct match when the relative overlapping area with a reference box is greater than or equal to 0.5, and has the same label as the reference. The relative overlapping area is computed as the intersection over union (IoU) areas, as follows:

$$\text{IoU} = \frac{|A \cap B|}{|A \cup B|} \tag{2}$$

[3]https://www.prhlt.upv.es/contests/icfhr2016-kws/data.html

Table II: Penalization applied on the mAP obtained by the solutions submitted on each period of the competition.

Dates	Training avail.	Penalized mAP
June 14–21	I	$\frac{\text{mAP}}{1.0}$
June 22–25	I + II	$\frac{\text{mAP}}{1.5}$
June 26–29	I + II + III	$\frac{\text{mAP}}{2.0}$

In order to assess the performance of each system under different amounts of training data, the mAP obtained on each challenge was penalized depending on the amount of training data available at the time of the submission. The penalty factors were obtained based on the results from the ICDAR2015 Competition on KWS[2], comparing the best performing training-based system and the best knowledge-based system. Taking into account these and the differences in the total amount of training data available in the two competitions, the penalty factors shown in Tab. II were applied on each period of the competition.

On each submission, a XML filke was requested for each dataset and the mAP for that submission was computed as the average mAP over the two datasets.

Following these rules, the score of a participant U in a given track was computed as follows. First, for each submission S, having access to the training data T (available in the corresponding period) on the challenge A, the average mAP over the two datasets (D_1 and D_2) was computed:

$$\begin{aligned}\text{mAP}(U,A,T,S) &= 0.5 \cdot \text{mAP}(U,A,T,S,D_1) + \\ &\quad 0.5 \cdot \text{mAP}(U,A,T,S,D_2)\end{aligned}$$

Then, the penalty factor $P(T)$ for the training data T was applied to obtain a penalized mAP:

$$\text{PmAP}(U,A,T,S) = \frac{\text{mAP}(U,A,T,S)}{P(T)}$$

Only the least penalized submission is considered for each challenge, A, as the final score for the given user, U:

$$\text{PmAP}(U,A) = \max_{T,S} \text{PmAP}(U,A,T,S)$$

Finally, the score of the user in the given track combines the penalized mAP obtained in the two challenges as follows, in order to give extra credits to those teams that were able to participate in both challenges, without penalizing excessively those participants that decided to send solutions to only one of the two challenges in each track.

$$\text{Score}(U) = \max_{A} \text{PmAP}(U,A) + 0.2 \cdot \min_{A} \text{PmAP}(U,A)$$

The participants could check the mAP, penalized mAP and their final score on each track using the same web-based interface created for submitting results to the contest. In order to avoid that they overfit on the test set, the submissions were restricted to one every 2 hours. Moreover, two software implementations were given beforehand: one used to compute the mAP of a particular submission[4] and another that computes the final track score of the participants and ranks them based on that score[5].

IV. Participants

Nine teams registered in the competition, from which four submitted at least one solution to the automatic evaluation system. Four teams participated in the track I: Query-by-Example and three in the track II: Query-by-String. In this section, the best performing systems of each participant team, according to the rules described in Sect. III, are described.

Computer Vision Center (CVCDAG), Universitat Autònoma de Barcelona, Spain – Track I.A, I.B, II.A. *(Suman Kumar Ghosh, Ernest Valveny, Marçal Rossinyol)* Word images are first encoded into feature vectors using Fisher Vector. Then, these feature vectors are used together with pyramidal histogram of characters labels (PHOC) [3] to learn SVM-based attribute models. PHOC encodes if a particular character appears in a particular spatial region of the string. The basic representation is just a binary histogram of characters, encoding which characters appear in the string. In order to add more discriminative power new levels are added to this histogram in a pyramidal way. At each level of the pyramid the word is further split and a new histogram of characters is added for each new division to account for characters at different parts of the word. At the end, 5 levels are used leading to a word representation of 604 dimensions. Then using learned SVM attributes from images and their corresponding text labels, a common subspace is learned to make the comparison between binary embedding and real valued attribute trivial. To learn this common subspace Canonical Correlation analysis is performed. For the segmentation-free challenge, a sliding window based approach similar to described in [4] was used.

Pattern Recognition Group (PRG), TU Dortmund University, Germany – Track I.A, I.B, II.A, II.B. *(Sebastian Sudholt, Leonard Rothacker, Gernot A. Fink)*. The method used in the word spotting competition is the recently invented PHOCNet, which is under review for ICFHR2016[6]. The PHOCNet is a 19-layer Convolutional Neural Network, specifically designed for learning document image attributes. For the competition, the exact same setup as is described in the paper was used: a Convolutional Neural Network (ConvNet) was trained with the PHOC[3] representation for each word image. The same training parameters as described in the preprint were used for the competition. Afterwards, the PHOCNet can predict the PHOC for a given

[4]https://www.prhlt.upv.es/contests/icfhr2016-kws/software/icfhr16kws_evaluation_toolkit.zip

[5]https://www.prhlt.upv.es/contests/icfhr2016-kws/software/ranker_toolkit.py

[6]Preprint available at https://arxiv.org/abs/1604.00187

word image without having to rescale or crop the image. Note that, in contrast to [3], character unigrams are used as attributes instead of bigrams. For the segmentation-based tasks, the word images are processed by the ConvNet and the predicted PHOC representation was then compared to the query PHOC. The query PHOC can either be another PHOC prediction from the ConvNet (QbE) or the PHOC extracted from the query string (PHOC). Similarity between queries is measured by the Bray-Curtis dissimilarity[5]. For the segmentation-free task, a sliding window over the document images was used to extract the PHOC for each window position. For efficiency reasons, the ConvNet output for 6 patch sizes was pre-calculated by clustering the training word image sizes. For QbE, each query is then mapped to its closest pre-computed patch size and retrieval is performed with this size (QbE). For QbS, the training word image with minimal Bray-Curtis dissimilarity was used as the query PHOC for retrieval.

Visual Information and Interaction (QTOB), Uppsala University, Sweeden – Track I.A, II.A. *(Anders Brun, Fredrik Wahlberg, Kalyan Ram, Tomas Wilkinson).* A ConvNet is trained to extract an image representation by using a triplet network approach [6] whereby a descriptor is learned for a word image by trying to predict whether or not words are belonging to the class. For the ConvNet architecture, the 34-layer ResNet from [7] is used. Then, a fully connected network is used to learn an embedding from the image representation space to a word embedding space by minimizing the cosine distance between the embedded images and their corresponding string representations. Once word images are embedded, either query-by-example or query-by-string word spotting can be performed in the word embedding space by means of the cosine distance. To embed text strings into a high dimensional space, a novel encoding based on the Discrete Cosine Transform (DCT). Essentially, it applies the DCT to a one-hot encoding of the word and keeps the first N components. In this competition, $N = 3$ components were used, which results in a feature vectors of around 150 dimensions, depending on the size of the alphabet of the dataset. Data augmentation was used to increase the size of the training set by applying simple geometric and morphological operations to the already existing training data. All the models were trained on an NVIDIA GTX Titan using Torch[8].

Tel Aviv University (TAU), Israel – Track I.A, I.B *(Adi Silberpfennig, Lior Wolf, Nachum Dershowitz).* This approach is based on previous work [9], [10] and has been used in previous work [11]. Whereas, originally, it was used for a segmentation - free KWS scenario [10], here it is used for both QbE challenges. In the segmentation - free case, a first step to extract word candidates from the document pages is carried out. The images are binarized and connected components are computed, filtering out too small or too big components. These steps are also done in the segmentation - based challenge, in order to filter noise from the provided word images. Additionally, in the segmentation - based case, a margin of a fixed size is added around the original images. Each word image patch (including query images) is resized to a patch of fixed size (168×72). The regular patch is divided into non-overlapping cells of 8×8 pixels, from which 31 HOG descriptors and 58 LBP descriptors are extracted and concatenated into a single vector which is normalized to have norm 1. Vectors from all cells are concatenated into a single vector of $16\,821$ elements. A matrix M consisting of the vector representations for $K = 900$ random candidates is then considered. Each vector $\mathbf{v}$ is transformed into a vector $\mathbf{u}$ by means of a linear projection, $\mathbf{u} = M \cdot \mathbf{v}$. Then, $\mathbf{u}$ is randomly split into fixed groups of size $L = 3$ and max-pooling is performed in each group, reducing the dimensionality of $\mathbf{u}$ to 300 elements. All hyper-parameters were tuned using the training data. The vectors obtained from the query images are then compared, by means of L2 distance, to the word vectors from the document images, and ranked according to the L2 distance. In order to eliminate overlapping windows, only candidates with the highest rank, out of all candidate targets that contain the same connected component as their largest component, are considered. In order to improve the performance of the method, each query image is considered more than once, by shifting the original image a fixed number of pixels in the four directions. Then, the maximum scores from all the images are selected. Finally, a re-ranking procedure is employed considering only the top 100 results for each query and re-ranking the results using the cosine distance of the HOG+LBP vectors.

V. Results

Table III shows best penalized mAP results obtained by each team on each dataset and challenge, the average penalized mAP for each challenge and the team's final score used to compute the ranking. Due to the penalization factors used, all the submissions in this table were uploaded during period I of the contest, i.e. using only at most 10 pages of training data. Thus, the penalized mAP in these cases is the raw mAP obtained by each submission.

In order to provide with a deeper analysis of the participating methods, Table IV shows the mAP of the best submission of each team, on each dataset and challenge, without applying any penalization regarding the amount of training data used. We should stress, however, that only Table III was used to rank the teams and choose the winners of each track following the stated competition rules.

VI. Conclusions

This competition on Keyword Spotting entailed several innovations with respect to most previous similar competitions and its aims were ambitious in many directions. First, the rules were established so as to allow a fully

Table III: Best penalized Mean Average Precision (PmAP) results in each track: (a) Query by Example and (b) Query by String. Each row shows the best results of each participant, on each dataset and challenge. The average PmAP is also shown for each challenge, and the last column shows the team's final score according to the competition rules described in Sect. III. The best result for each challenge and dataset is highlighted in bold.

(a) Track I: Query by Example

	Segm. based			Segm. free			
Team	Botany	Konzil.	Average	Botany	Konzil.	Average	Final Score
CVCDAG	**75.77**	77.91	**76.84**	0.21	0.0	0.10	**76.86**
PRG	46.61	**88.14**	67.38	15.89	52.20	34.05	74.18
TAU	50.64	71.11	60.87	**37.48**	**61.78**	**49.63**	70.80
QTOB	54.95	82.15	68.55	—	—	—	68.55

(b) Track II: Query by String

	Segm. based			Segm. free			
Team	Botany	Konzil.	Average	Botany	Konzil.	Average	Final Score
PRG	36.47	**76.93**	56.70	**11.80**	**48.41**	**30.10**	**62.72**
CVCDAG	**65.69**	55.27	**60.48**	—	—	—	60.48
QTOB	3.40	12.19	7.79	—	—	—	7.79

Table IV: Best Mean Average Precision (mAP) results obtained by each team, on each dataset and challenge, without any penalization applied regarding to the amount of training data used. The period on which each submission was done it is also shown. Subtable (a) shows the Query by Example track and (b) Query by String. The best result for each challenge and dataset is highlighted in bold.

(a) Track I: Query by Example

	Segm. based				Segm. free			
Team	Botany	Konzil.	Average	Period	Botany	Konzil.	Average	Period
CVCDAG	75.77	77.91	76.84	I	0.42	0.0	0.21	III
PRG	**89.69**	**96.05**	**92.87**	III	15.89	52.20	34.05	I
TAU	50.64	71.11	60.87	I	**37.48**	**61.78**	**49.63**	I
QTOB	54.95	82.15	68.55	I	—	—	—	—

(b) Track II: Query by String

	Segm. based				Segm. free			
Team	Botany	Konzil.	Average	Period	Botany	Konzil.	Average	Period
PRG	**74.47**	**94.20**	**84.34**	III	**11.80**	**48.41**	**30.10**	I
CVCDAG	65.69	82.91	74.30	II	—	—	—	—
QTOB	3.40	12.19	7.79	I	—	—	—	—

homogeneous assessment of all the major flavors of KWS. To this end a unique evaluation protocol and assessment measures were defined. However, we also wanted to take into account the amount of previous information required by the different KWS approaches so that methods which rely on the least amount of information became better scored in the final evaluation ranking. Therefore, to better benchmark the capabilities of methods based on training data, we established penalty factors roughly proportional to the amount of training data used to obtain each result.

An analysis of the results show that the penalty factors adopted were strongly affecting systems based on training data. For practical applications of KWS to indexing moderately large text image collections, asking for 40 annotated pages for training is not really significant and even 154 annotated pages (as provided over all periods for the *Botany* dataset) are perfectly affordable when indexing large collections of, say, tens or hundreds of thousands of images. Systems which significantly rely on training, such as those of PRG, do take great advantage of the available training material. For instance, in the QbE segmentation-based challenge, they achieve more than 15 percent better mAP than the winner system of that track (CVDAG). Clearly, such an overwhelming superiority had made PRG the winner of both Tracks I and II, should the training data penalties had been just a little less severe. This is a lesson learn from this competition which should be carefully taken into account in future similar events.

On the other hand, in this competition we did not apply any penalty related to the amount of information entailed by the word segmentation which is available to all the segmentation-based challenges. Real KWS applications of

moderate size can hardly rely on word segmentation; either because automatic word segmentation is very prone to segmentation errors and because manually producing or amending word segmentation bounding boxes is exceedingly expensive to produce even a few tens of annotated page images. Our choice of not penalizing (segmentation-based) systems which need word bounding boxes is clearly inconsistent with our general aim of taking into account the amount of previous information required by the different KWS approaches. Therefore, if segmentation-based and segmentation-free challenges are to be uniformly considered in future KWS competitions, significant penalty factors should be applied to the results of segmentation-based systems. Alternatively, segmentation-based challenges should not be explicitly considered; that is, segmentation-based systems should have to provide by themselves for an automatic segmentation of the given unsegmented test images.

There is another point in which we did not fully comply with our aim of unifiorm evaluation. For the sake of homogeneity, the very same query words were actually used to evaluate QbS and QbE systems. However, in the QbE track for many of these words, several (up to ten) query images per query word were manually selected and used as query examples. Obviously, in the QbS track, there is no point in repeating several times the same query. Therefore, in the QbS track all the query words have the same impact on the overal mAP result. Conversely, in the QbE track, those words with more examples have the greatest impact. According to the statisitiscs of Fig. 2, about 90% of the query images correspond to long words, with 5 or more characters, which are easier to spot by both QbS and QbE systems. But, because of the repeated examples, these good spots have a higher (positive) impact on the QbE systems than on the QbS ones. These subtle, but important evaluation condition differences probably explain the unexpected fact that all the systems achieved better results in Track I (QbE) than in Track II (QbS).

Therefore, in future competitions, we strongly suggest query words to be *randomly* selected so as to approach as much as possible realistic conditions of practical use – and the amount of query words should be very much larger. Using a few thousands of not manually chosen words should probably be enough to ensure the required degree of variability and realistic difficulty.

ACKNOWLEDGMENT

This work was partially supported by the Spanish MEC under FPU grant FPU13/06281, by the Generalitat Valenciana under the Prometeo/2009/014 project grant ALMAMATER, and through the EU projects: HIMANIS (JPICH programme, Spanish grant Ref. PCIN-2015-068) and READ (Horizon-2020 programme, grant Ref. 674943).

REFERENCES

[1] I. Pratikakis, K. Zagoris, B. Gatos, G. Louloudis, and N. Stamatopoulos, "ICFHR 2014 Competition on Handwritten Keyword Spotting (H-KWS 2014)," in *Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on*, Sept 2014, pp. 814–819.

[2] J. Puigcerver, A. H. Toselli, and E. Vidal, "ICDAR2015 Competition on Keyword Spotting for Handwritten Documents," in *Document Analysis and Recognition (ICDAR), 2015 13th International Conference on*, Aug 2015, pp. 1176–1180.

[3] 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, Dec 2014.

[4] S. K. Ghosh and E. Valveny, *A Sliding Window Framework for Word Spotting Based on Word Attributes*. Cham: Springer International Publishing, 2015, pp. 652–661. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-19390-8_73

[5] S. Sudholt and G. A. Fink, *A Modified Isomap Approach to Manifold Learning in Word Spotting*. Cham: Springer International Publishing, 2015, pp. 529–539. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-24947-6_44

[6] V. Balntas, E. Johns, L. Tang, and K. Mikolajczyk, "PN-Net: Conjoined Triple Deep Network for Learning Local Image Descriptors," Tech. Rep., 2016. [Online]. Available: http://arxiv.org/abs/1601.05030

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," Tech. Rep., 2015. [Online]. Available: http://arxiv.org/abs/1512.03385

[8] R. Collobert, K. Kavukcuoglu, and C. Farabet, "Torch7: A Matlab-like Environment for Machine Learning," in *BigLearn, NIPS Workshop*, 2011.

[9] Q. Liao, J. Z. Leibo, Y. Mroueh, and T. A. Poggio, "Can a biologically-plausible hierarchy effectively replace face detection, alignment, and recognition pipelines?" *CoRR*, vol. abs/1311.4082, 2013. [Online]. Available: http://arxiv.org/abs/1311.4082

[10] A. Kovalchuk, L. Wolf, and N. Dershowitz, "A Simple and Fast Word Spotting Method," in *Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on*, Sept 2014, pp. 3–8.

[11] A. Silberpfennig, L. Wolf, N. Dershowitz, S. Bhagesh, and B. B. Chaudhuri, "Improving OCR for an under-resourced script using unsupervised word-spotting," in *Document Analysis and Recognition (ICDAR), 2015 13th International Conference on*, Aug 2015, pp. 706–710.