

# Exploring Confidence Measures for Word Spotting in Heterogeneous Datasets

Fabian Wolf  
Department of Computer Science  
TU Dortmund University  
44227 Dortmund, Germany  
fabian.wolf@cs.tu-dortmund.de

Philipp Oberdiek  
Department of Computer Science  
TU Dortmund University  
44227 Dortmund, Germany  
philipp.oberdiek@cs.tu-dortmund.de

Gernot A. Fink  
Department of Computer Science  
TU Dortmund University  
44227 Dortmund, Germany  
gernot.fink@cs.tu-dortmund.de

**Abstract**—In recent years, convolutional neural networks (CNNs) took over the field of document analysis and they became the predominant model for word spotting. Especially attribute CNNs, which learn the mapping between a word image and an attribute representation, showed exceptional performances. The drawback of this approach is the overconfidence of neural networks when used out of their training distribution. In this paper, we explore different metrics for quantifying the confidence of a CNN in its predictions, specifically on the retrieval problem of word spotting. With these confidence measures, we limit the inability of a retrieval list to reject certain candidates. We investigate four different approaches that are either based on the network’s attribute estimations or make use of a surrogate model. Our approach also aims at answering the question for which part of a dataset the retrieval system gives reliable results. We further show that there exists a direct relation between the proposed confidence measures and the quality of an estimated attribute representation.

## I. INTRODUCTION

Word spotting is a powerful tool for exploring handwritten document collections. Machine learning based methods got increasingly popular and showed exceptional performances on numerous academic benchmarks [1]. It has been shown that attribute based word spotting systems are extremely robust to high variations in style and appearance of the documents [2]. While more and more sophisticated models emerge, they all share the assumption that representative training material is available. This is often the case for an academic benchmark but in a real world application the system is faced with an unknown dataset. Especially in historic and handwritten document collections, writing styles can change frequently. Degradation can drastically change the visual appearance of the documents over the collection making them highly heterogeneous. A limited amount of training material might be provided by some expert who annotated a small part of the dataset. Assessing if the training material is representative is practically impossible without manually evaluating the entire document collection.

The inability to reject certain candidates is one of the main drawbacks of a retrieval system. The result w.r.t. a given query is always an ordered list of instances. A common retrieval system does not provide any information on which parts of a dataset it is able to retrieve reliable results. State-

of-the-art approaches make use of CNNs that learn the mapping between a word image and an attribute embedding [2]. Attributes are semantic entities shared between multiple classes, which have been shown to be highly robust representations and allow for zero-shot learning [3]. Due to the overconfidence of neural networks, high attribute activations can be observed for whatever input is given to the network. Even an image of random noise would result in an attribute representation and the retrieval system would rank it w.r.t. the query. In order to extend the capabilities of a retrieval system, this work aims at finding a suitable confidence measure that allows to assess whether a CNN is able to estimate an accurate attribute representation.

We explore different confidence measures that will be evaluated according to the experimental setup depicted in Fig. 1. A *segmentation-based* word spotting system is trained on an annotated set of training material. In order to model a highly heterogeneous dataset, the test set is composed of an in-distribution (ID) part that is well represented by the training data and an out-of-distribution (OD) part. While not having any explicit relation to the training material, the OD set shall consist of a wide range of different samples, being more and less similar to the style of the training set. The word spotting system is used to generate a retrieval list for each query over the composed test set. Given a confidence measure, the generated retrieval lists are pruned by a simple thresholding method. All samples above a given confidence should have an accurate attribute estimation without necessarily being from the ID part. Samples with an associated confidence below the threshold are rejected. Thereby, inaccurate attribute estimations, which would lead to poor retrieval performance, are removed from the test set.

## II. RELATED WORK

### A. Word Spotting

Word spotting describes the task of retrieving a subset of word images from a document collection that are relevant w.r.t. a query. In contrast to methods aiming at directly transcribing a document, word spotting systems have been shown to be extremely robust. This makes word spotting a highly suitable technique, especially when confronted with

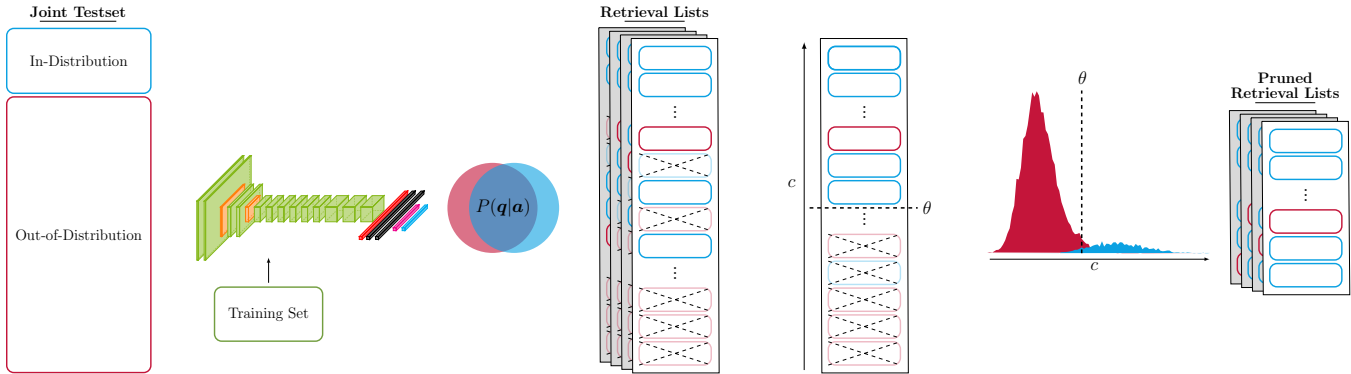


Figure 1: The figure visualizes the experimental setup. A *segmentation-based* word spotting system is trained on a training set and evaluated on a composed test set including ID and OD samples. For each query, a retrieval list is obtained by evaluating a probabilistic retrieval model. All samples are associated to a confidence measure. Inaccurate estimations are removed by thresholding.

handwritten historic documents that often show high variability and suffer from degradation effects. For an extensive overview of word spotting methods, see [1].

In general, most approaches make several assumptions w.r.t. the document collection and the query protocol. *Segmentation-based* methods (c.f. e.g. [2], [3]) require a previous segmentation of document pages into individual word images, which is in general not an easy to solve problem. The *segmentation-free* approach does not pose this requirement, but aims at solving the retrieval and segmentation problem jointly. Considering the query provided by the user, two different protocols are distinguished. In case of *query-by-example* (QbE), the query is provided as a word image. *Query-by-string* (QbS) allows for string based representations as queries.

In [3], the concept of attribute-based learning was introduced to the field of word spotting. Attributes are entities that are shared between different classes. With respect to word images, a specific word can be considered as a class while its characters can be interpreted as attributes. Taking spatial relations between characters into account, the *Pyramidal Histogram of Characters* (PHOC) is derived. In [3], the mapping between attribute embedding and word images is learned with a set of support vector machines. This allows to map word images and strings in a common subspace where the retrieval problem can be solved by comparing distances between attribute vectors. Inspired by the success of CNNs, [2] used a neural network to learn the attribute embedding. This approach outperformed all previous methods by a large margin and still defines the state-of-the-art for *segmentation-based* word spotting. In [4], a probabilistic retrieval model (PRM) is proposed. While cosine similarity and euclidean distances do not provide a robust distance metric in high dimensional spaces, the PRM gives a probabilistic description of similarity between query and estimated attribute vector. Even though the attribute CNN approach has shown excellent performance

on numerous commonly used academic benchmarks, this comes at the cost of requiring training material. Works such as [5] and [6] try to alleviate the data problem by transfer learning and incorporating synthetic data, but still the necessity of representative training data is inherent to any machine learning based approach.

### B. Uncertainty

The task of uncertainty estimation of neural networks and OD detection has recently been an active field. Estimating the uncertainty of neural networks by applying dropout during test time was analysed by [7]. However, this method is computationally expensive, as one has to make a large number of forward passes through the network. Different types of uncertainty were analysed by [8]. They distinguish model capacity uncertainty, intrinsic data uncertainty and open set uncertainty.

Approaches by [9] and [10] rely on additional surrogate models. A baseline for the use of confidence measures for OD detection in a classification scenario was proposed by [9]. The use of the maximum softmax entry as a confidence measure yielded good results. The interpretability however is a major drawback, as most neural networks are overconfident in their decisions. This leads to high confidence values for most of the OD samples. They also suggested the use of a multi headed neural network utilizing an auxiliary decoder together with an 'abnormality module', which increased the separability between ID and OD examples significantly. Multi headed neural networks were also used by [10] who added a second branch in parallel with the fully connected layer of a neural network. During training, they used a joint loss function based on an interpolation approach between network prediction and supplied label. The interpolation factor could then be interpreted as a confidence measure. This yielded good performances on classification tasks but showed a high regularization effect, which made it necessary to introduce additional hyperparameters.

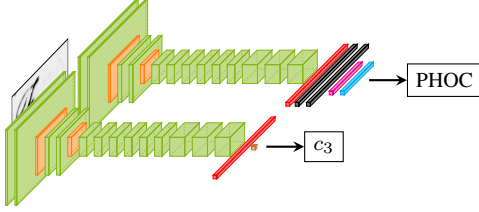


Figure 2: Task independent metaclassifier.

### III. METHOD

With all proposed confidence measures, we aim at quantifying the quality of a predicted attribute vector. The assumption is that data points that are dissimilar from the training distribution have an inaccurate attribute prediction, which results in a wrong position in the retrieval list.

#### A. Word Spotting

Our baseline word spotting system is based on the design of [2]. The attribute embedding is a 4-level PHOC representation of partitions 1, 2, 4 and 8 based on the lower case Latin alphabet plus digits. We employ a TPP-PHOCNet to estimate a PHOC vector  $\hat{\mathbf{a}} \in (0, 1)^{540}$  for a word image.

In analogy to [4] the retrieval list for the query  $q$  with the corresponding attribute embedding  ${}^q\mathbf{a}$  is ranked according to the posterior probability

$$p({}^q\mathbf{a}|\mathbf{x}) = \prod_{i=1}^{540} \hat{a}_i^{q a_i} \cdot (1 - \hat{a}_i)^{(1-q a_i)}$$

which serves as a similarity measure. For a given annotation  $t$ , the quality of an estimated PHOC vector can be quantified by evaluating the probabilistic model w.r.t. the ground truth attribute embedding  ${}^t\mathbf{a}$ . Therefore,  $p_t|\mathbf{x} = p({}^t\mathbf{a}|\mathbf{x})$  describes the probability of  ${}^t\mathbf{a}$  being the embedding of  $\mathbf{x}$ .

#### B. Sigmoid Activation and Test Dropout

Considering sigmoid activation as a pseudo probability, we derive the confidence measure  $c_1$  by taking the mean over the activation of all active attributes. An attribute is considered active in case of  $\hat{a}_i > 0.5$ .

The second confidence measure is based on dropout at test time. We apply dropout with a probability of 0.5 to all but the last fully connected layer. Each sample is passed through the network 100 times with both dropout layers being active. The variance of the estimations for each attribute is determined. The confidence measure  $c_2$  is then obtained by averaging over all attribute variances. Opposed to the other confidence measures, a high confidence corresponds to a small value of  $c_2$ .

#### C. Task Independent Metaclassifier

A task independent (TI) metaclassifier is an additional surrogate model, which has no relation of the task learned by the main model. It receives the same input and classifies

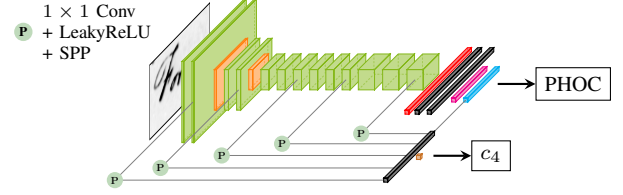


Figure 3: Task dependent metaclassifier using deep features.

in ID and OD. For training, one only needs the training set of the main model and some data of a different distribution. This could be anything from gaussian-/normal-noise over synthetically generated to real world samples. This makes it rather easy to obtain the OD data. The proposed independent metaclassifier is shown in Fig. 2. Its architecture is a replication of the PHOCNet, but with a different MLP part. The original fully connected layers and the sigmoid output are replaced by a single neuron with sigmoid activation.

Let  $X$  be the training set of word images for our task with a set of known PHOC vector representations  $A = \{\mathbf{a} \in \{0, 1\}^{540}\}$  and  $O$  a set of word images, which are sampled from a different distribution than  $X$ , without a known transcription or PHOC vector. The PHOCNet is trained on  $X$  with labels  $A$  to approximate the distribution  $p(\mathbf{a}|\mathbf{x})$  whereas the metaclassifier is trained on  $X \cup O$  to approximate the distribution  $p(d|\mathbf{x})$ , giving the probability that  $\mathbf{x}$  belongs to  $X$ . The label set  $D$  for the training of the metaclassifier can be obtained as

$$D = \{d(\mathbf{x})|\mathbf{x} \in X \cup O\} \text{ with } d(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbf{x} \in X \\ 0 & \text{else} \end{cases}.$$

The surrogate model is trained with binary crossentropy loss and hyperparameters as described in Sec. IV-B. The confidence measure  $c_3$  results from the penultimate layer of the metaclassifier.

#### D. Task Dependent Metaclassifier

A task dependent (TD) metaclassifier receives the featuremaps learned by the PHOCNet as input. Using the already learned representations, it allows to reduce the number of additional parameters and produces a confidence measure, which is semantically related to the main task. Defining with  $f_i := f_i(\mathbf{x}, \mathbf{w})$  the output of the PHOCNet up until the  $i$ -th layer w.r.t. the weights  $\mathbf{w}$ , the task dependent metaclassifier learns to approximate the distribution  $p(d|f_{s_1}, \dots, f_{s_l})$ ,  $s_1, \dots, s_l \in \{1, \dots, L\}$  with  $L$  being the number of layers (only counting layers that have trainable weights). Here, we choose  $f_2, f_4, f_7, f_{10}, f_{13}$  and  $f_{16}$ , which are passed through a  $1 \times 1$  convolution with one feature map followed by a leaky ReLU activation and spatial pyramid pooling (SPP) up to level 4. The resulting feature vectors and the penultimate layer of the PHOCNet

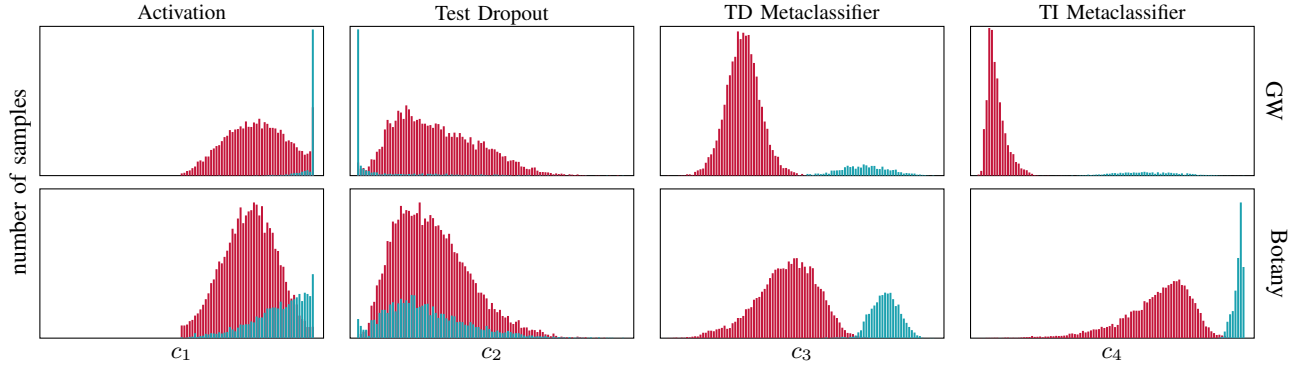


Figure 4: Distribution of confidences of the ID (blue) and OD (red) set for all proposed methods.

Dataset	Train	Test	Historic	Writers
GW [11]	PHOCNet	ID	yes	1
Botany [12]	PHOCNet	ID	yes	1
IAM [13]	-	OD	no	657
HWSynth [5]	MC	-	no	0

Table I: Datasets used in this work. Each dataset is either used to train the PHOCNet, the metaclassifier (MC) or models ID or OD.

are then concatenated and fed through a single neuron with sigmoid activation. First, the PHOCNet weights are trained as described in Sec. IV-B. Afterwards, we freeze the weights of the attribute CNN and train the metaclassifier with binary crossentropy loss on  $X \cup O$ , label set  $D$  and hyperparameters as described in Sec. IV-B. The confidence measure  $c_4$  is then taken from the penultimate layer of the metaclassifier.

#### IV. EXPERIMENTS

See Fig. 1 for an overview of our experimental setup. We train a *segmentation-based* word spotting system on a designated training set. The different confidence measures are then evaluated on a composed test set with the aim to distinguish between ID and OD and to prune inaccurate attribute vector estimations from the resulting retrieval lists.

##### A. Datasets

We use four different publicly available datasets. The George Washington (GW) and Botany dataset are well known to the word spotting community and both have a distinctive style. Our baseline word spotting system is trained on the training partition in case of Botany or follows the common four-fold cross validation approach of GW, c.f. [3]. The respective test partitions are used as ID sets. Note that for the Botany test set, annotations are only available for those samples which are relevant w.r.t. a query following the standard protocol [12]. The IAM database contains a wide range of different writing styles, which makes it a suitable choice as the OD set. Our choice of datasets is further motivated by the fact that our experiments require overlapping lexica, since word images relevant to a query shall exist in the ID and OD set. For training

the metaclassifiers, we use the synthetic dataset HWSynth to model OD samples. See Tab. I for an overview of the different datasets.

##### B. Training Setup

For all our experiments, we train the TPP-PHOCNet for 100 000 iterations with an initial learning rate of  $10^{-4}$ , which is divided by 10 after 70 000 iterations. We use Adam optimization with a mini-batch size of 10, hyperparameters  $\beta_1 = 0.9$  and  $\beta_2 = 0.999$  and a weight decay of  $5 \cdot 10^{-5}$ . Analogue to Sec. III-B, we apply two dropout layers during training. Furthermore, a simple data augmentation strategy is used as we apply a random affine transformation to all input images at training time.

The metaclassifiers are trained for 25 000 iterations with an initial learning rate of  $10^{-2}$ , which is divided by 10 after 10 000, 15 000 and 20 000 iterations. For optimization, we use the same Adam optimizer as for the TPP-PHOCNet with a weight decay of  $5 \cdot 10^{-4}$ .

##### C. Results and Discussion

Our experiments investigate the following questions:

- Are the proposed confidence measures suitable for separating ID and OD examples?
- Does a relation between confidence and the *quality* of an estimated attribute vector exist?
- Is it possible to prune the resulting retrieval lists by *thresholding* in order to compromise between the reliability of the results and the coverage of the dataset?

**Separability:** According to our experimental setup, we assume that the word spotting system is giving reliable results on the ID sets. This does not hold true for the OD part, as no related training material was used. The proposed confidence measures are supposed to quantify this assumption and they should yield higher confidences for the ID parts. Fig. 4 shows the distributions of the different confidence measures. On GW, high confidences can be observed w.r.t. to the attribute estimations of the ID samples and the confidence measures derived from sigmoid activation or via test dropout. In case of Botany, this observation does not hold true.

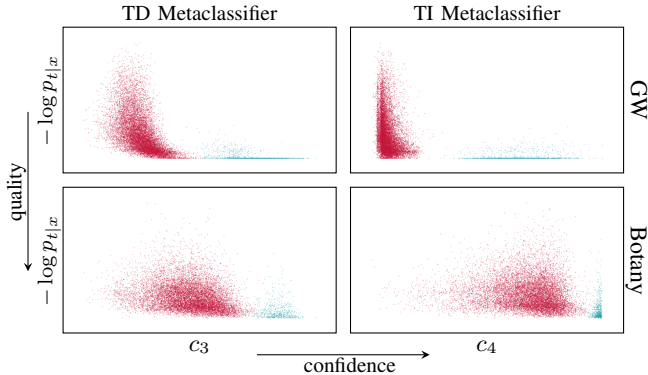


Figure 5: Distribution of ID and OD samples over the negative logarithmic posterior  $p_{t|x}$  and the confidence measure.

For sigmoid activation, slightly higher confidences can be observed for the ID set. The distributions of confidences resulting from test dropout do not allow to easily distinguish between ID and OD. In contrast, both meta-classifiers are almost perfectly able to distinguish between ID and OD samples on both datasets. The confidences for ID samples are almost exclusively higher than those of the OD samples.

**Relation to Attribute Vector Quality:** As discussed in Sec. III-A, the quality of an attribute vector estimation w.r.t. the ground truth transcription can be described by  $p_{t|x}$ . Even though the meta-classifiers are almost perfectly able to distinguish ID and OD, that is not exactly what we aim for in terms of measuring confidence. Due to the multi writer characteristic of the OD set, a wide range of writing styles is represented. Given the training material, a confidence measure should quantify how well an attribute vector can be estimated for a sample. In case of the OD set, this should correspond to the quality of an estimated attribute vector measured by the posterior  $p_{t|x}$ .

Fig. 5 shows the distribution of ID and OD samples over the confidence measure and the posterior  $p_{t|x}$ , which quantifies the quality of the estimation. Almost all ID samples of GW have a high probability  $p_{t|x}$  and can be assumed to be precise estimations of their ground truth embeddings. This is not the case for the Botany dataset, where samples with a high confidence are not concentrated at high posteriors. Despite a high mean average precision (mAP) on Botany according to the standard protocol, this observation indicates that the task of estimating an attribute vector is solved poorly compared to the GW dataset. Comparing the task dependent and independent meta-classifiers, one can observe a correlation between confidence and quality especially for task-dependency. This is far less observable for the task independent meta-classifier. While the deep features used by the task dependent meta-classifier provide a semantic relation to the task of estimating an attribute vector, the task independent confidence measure is solely based on visual similarity. This might explain the observed characteristics.

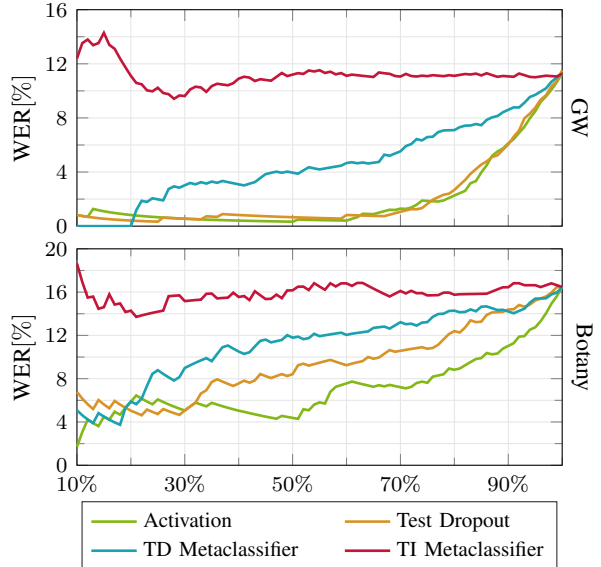


Figure 6: Cumulative word error rates.

To further investigate the relation between confidence measures and attribute vector quality, we conduct the following experiment. The given word spotting system can easily be extended to perform lexicon-based word recognition analogue to the approach of [14]. Let  $\mathbb{L}$  be the lexicon obtained from all available training and test set transcriptions with corresponding attribute representation  ${}^l\mathbf{a}$ . Based on the attribute representation estimated for a word image  $\mathbf{x}$  the recognition result  $s$  is given by

$$s = \operatorname{argmax}_{l \in \mathbb{L}} p({}^l\mathbf{a}|\mathbf{x}).$$

Note that we do not expect the proposed system to give state-of-the-art recognition results, as we believe that for this task sequential models are superior to holistic attribute based approaches. Nevertheless, evaluating word error rates gives a performance measure that describes if the most probable string of the lexicon is equal to the ground truth transcription. This experiment allows to evaluate the probability  $p_{t|x}$  for a given attribute vector estimation, with a commonly known and easier to interpret performance measure.

We first sort the ID set w.r.t. confidence. Then we determine the word error rate (WER) for the most confident  $x\%$  of the dataset. Fig. 6 shows the WER over different portions of the ID datasets. Even though our system yields a WER over the entire GW test set of 11.52%, it drops to below 2% for the most confident 70% using sigmoid activations or test dropout as confidence measure. All confidence measures, despite the task independent meta-classifier, achieve significantly lower WER on the more confident parts of the test sets. This further supports our conclusion that those confidence measures, semantically related to the task, also quantify the quality of an estimated attribute vector.



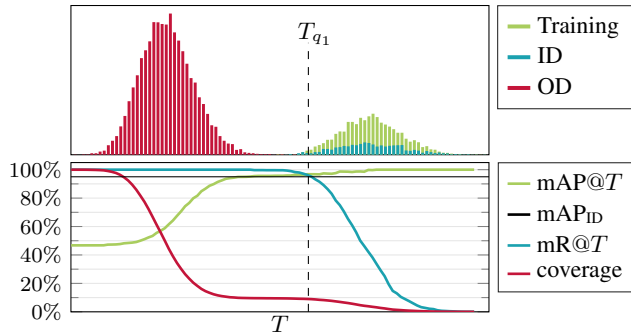


Figure 7: Thresholding characteristics of a task dependent meta-classifier on a test set composed of GW and the IAM database.

**Thresholding:** Fig. 7 visualizes the distribution of confidences for the training, ID and OD set, given a task dependent meta-classifier w.r.t. the GW dataset. If only ID samples are considered, the word spotting system produces the baseline performance  $mAP_{ID}$  for QbS word spotting. All unique strings of the ID set serve as queries ones. For the composed test set, we consider the case where all samples below a certain confidence are removed from the retrieval lists. At each possible threshold  $T$ , the  $mAP@T$  gives the QbS mAP over the pruned dataset. Additionally, Fig. 7 reports the  $mR@T$ , which is the mean recall over the pruned dataset. The coverage describes the percentage of the joint test set lying above the threshold  $T$ .

Starting with the entire joint test set, the  $mAP@T$  is significantly below the  $mAP_{ID}$ , since poor attribute estimations are not ranked properly. Increasing the threshold leads to removing inaccurate estimations, improving the  $mAP@T$  at the cost of a lower coverage. The  $mAP@T$  approaches the baseline performance as most of the OD samples are removed from the retrieval lists. A further increase of the threshold leads to pruning ID samples, which lowers the  $mR@T$ . As we only consider queries that occur at least once, the  $mAP@T$  further improves as only increasingly confident parts of the ID set are considered. The threshold can be considered a parameter allowing to compromise between the quality of retrieval results and the coverage of the test set. It is also possible to estimate a threshold based on the training distribution. As depicted in Fig. 7 top, using the one percent quantile of the training distribution  $T_{q_1}$  as a threshold, the ID set is almost exactly separated from the OD set, reproducing the baseline performance on the pruned test set.

## V. CONCLUSIONS

In this work, we proposed four different confidence measures for a word spotting system and showed that those semantically related to the task provide an estimation of the attribute vector quality. We conclude that a task dependent meta-classifier is a suitable model to distinguish ID and OD samples while quantifying quality. This allows to identify parts of a dataset for which reliable results can be obtained.

## REFERENCES

- [1] A. P. Giotis, G. Sfikas, B. Gatos, and C. Nikou, "A survey of document image word spotting techniques," *Pattern Recognition*, vol. 68, pp. 310–332, 2017.
- [2] S. Sudholt and G. A. Fink, "Attribute CNNs for word spotting in handwritten documents," *Int. Journal on Document Analysis and Recognition*, vol. 21, no. 3, pp. 199–218, 2018.
- [3] J. Almazán, A. Gordo, A. Fornés, and E. Valveny, "Word spotting and recognition with embedded attributes," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 36, no. 12, pp. 2552–2566, 2014.
- [4] E. Rusakov, L. Rothacker, H. Mo, and G. A. Fink, "A probabilistic retrieval model for word spotting based on direct attribute prediction," in *Proc. of the Int. Conf. on Frontiers in Handwriting Recognition*, 2018, pp. 38–43.
- [5] P. Krishnan and C. V. Jawahar, "Matching handwritten document images," in *Proc. of the European Conf. on Computer Vision*, 2016, pp. 766–782.
- [6] N. Gurjar, S. Sudholt, and G. A. Fink, "Learning deep representations for word spotting under weak supervision," in *Proc. of the Int. Workshop on Document Analysis Systems*, 2018, pp. 7–12.
- [7] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *Proc. of the Int. Conf. on Machine Learning*, 2016.
- [8] R. E. Harang and E. M. Rudd, "Principled uncertainty estimation for deep neural networks," *CoRR*, vol. abs/1810.12278, 2018.
- [9] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," in *Int. Conf. on Learning Representations*, 2017.
- [10] T. DeVries and G. W. Taylor, "Learning confidence for out-of-distribution detection in neural networks," *arXiv preprint arXiv:1802.04865*, 2018.
- [11] T. M. Rath and R. Manmatha, "Word spotting for historical documents," *Int. Journal on Document Analysis and Recognition*, vol. 9, no. 2-4, pp. 139–152, 2007.
- [12] I. Pratikakis, K. Zagoris, B. Gatos, J. Puigcerver, A. H. Toselli, and E. Vidal, "ICFHR2016 handwritten keyword spotting competition (H-KWS 2016)," in *Proc. of the Int. Conf. on Frontiers in Handwriting Recognition*, 2016, pp. 613–618.
- [13] U. Marti and H. Bunke, "The IAM-database: an english sentence database for offline handwriting recognition," *Int. Journal on Document Analysis and Recognition*, vol. 5, no. 1, pp. 39–46, 2002.
- [14] A. Poznanski and L. Wolf, "CNN-N-Gram for handwriting word recognition," in *Proc. IEEE Comp. Soc. Conf. on Computer Vision and Pattern Recognition*, 2016.