

Scene Text Retrieval via Joint Text Detection and Similarity Learning

Hao Wang¹, Xiang Bai¹, Mingkun Yang¹, Shenggao Zhu², Jing Wang², Wenyu Liu¹

¹Huazhong University of Science and Technology, ²Huawei Cloud & AI

{wanghao4659,xbai,yangmingkun,liuwy}@hust.edu.cn, {zhushenggao,wangjing105}@huawei.com

Abstract

Scene text retrieval aims to localize and search all text instances from an image gallery, which are the same or similar to a given query text. Such a task is usually realized by matching a query text to the recognized words, outputted by an end-to-end scene text spotter. In this paper, we address this problem by directly learning a cross-modal similarity between a query text and each text instance from natural images. Specifically, we establish an end-to-end trainable network, jointly optimizing the procedures of scene text detection and cross-modal similarity learning. In this way, scene text retrieval can be simply performed by ranking the detected text instances with the learned similarity. Experiments on three benchmark datasets demonstrate our method consistently outperforms the state-of-the-art scene text spotting/retrieval approaches. In particular, the proposed framework of joint detection and similarity learning achieves significantly better performance than separated methods. Code is available at: <https://github.com/lanfeng4659/STR-TDSL>.

1. Introduction

In the past few years, scene text understanding has received rapidly growing interest from this community due to a large amount of everyday scene images that contain texts. Most previous approaches for scene text understanding focus on the topics of text detection, text recognition and word spotting, succeeding in many practical applications such as information security, intelligent transportation system [38, 29], geo-location, and visual search [3]. Different from the above topics, we study the task of scene text retrieval introduced by Mishra *et al.* [26], aiming to search all the text instances from a collection of natural images, which are the same or similar to a given text query. Different from a scene text reading system that must localize and recognize all words contained in scene images, scene text retrieval only looks for the text of interest, given by a user. Such a task is quite useful in many applications including product image retrieval, book search in library [37] and key

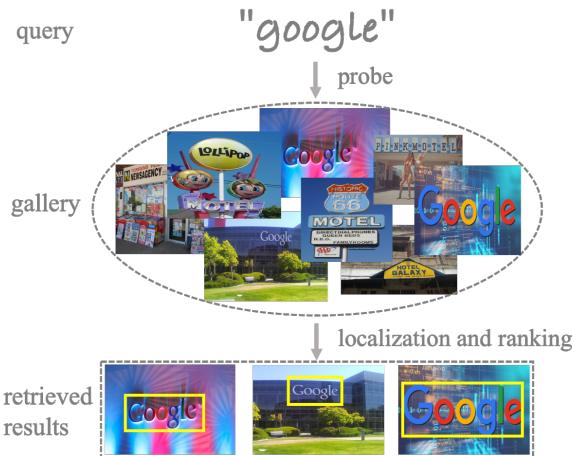


Figure 1. Given a query text “google”, the end-to-end scene text retrieval method aims to retrieve the images containing “google” from gallery, as well as their locations in the images.

frame extraction of videos [30].

As depicted in Fig. 1, the goal of scene text retrieval is to return all images that are likely to contain the query text, as well as the bounding boxes of such text. In this sense, scene text retrieval is a cross-modal retrieval/matching task that aims to close the semantic gap between a query text and each text proposal. Traditional text retrieval methods [2, 1, 5, 7] are often designed to handle cropped document text images. Since the well-cropped bounding boxes are not easy to obtain by a detector for natural images, these methods cannot be directly applied in scene text retrieval. Mishra *et al.* [26] first study text retrieval in scene images, which is cast as two separated sub-tasks: text detection and text image retrieval. However, the performance is limited, as their framework is designed based on handcraft features.

Another feasible solution to scene text retrieval is based on an end-to-end text recognition system, such as [14, 12]. Under this setting, the retrieval results are determined according to the occurrences of the given query word within the spotted words. As verified by [26, 6], such methods often achieve unsatisfactory retrieval performance, as the end-to-end text recognition system is optimized with a different

evaluation metric that requires high accuracy in terms of both detection and recognition. However, for a retrieval engine, more text proposals can be leveraged to reduce the performance loss brought by the missed detections. Gomez *et al.* [6] directly predict the corresponding Pyramidal Histogram Of Character [2] (PHOC) of each text instance for ranking. However, these methods don't explicitly learn the cross-modal similarity.

In this paper, we propose a new deep learning framework for scene text retrieval that combines the stages of text detection and cross-modal similarity learning. Our basic idea is motivated by the recent end-to-end text spotting methods [22, 24, 17, 12, 34, 18], which unify the two sub-tasks of text detection and text recognition with an end-to-end trainable network. Such methods usually achieve better spotting performance than those optimized separately, benefiting from feature sharing and joint optimization. Specifically, our network for scene text retrieval is optimized by two tasks: a text detection task that aims to predict bounding boxes of candidate text instances, and a similarity learning task for measuring the cross-modal similarity between a query text and each bounding box. With joint optimization, the text detection task and similarity learning task are complementary to each other. On the one hand, the detection task can pay more attention to the recall than the precision in order to avoid missing detections, as the feature matching process by the similarity learning task can effectively eliminate false alarms. On the other hand, the faithful cross-modal similarity provided by the similarity learning task can be used for looking for an accurate bounding box that contains a given text. In addition, both tasks share CNN features, significantly improving the inference speed.

Unlike general objects such as person and car, scene text is a kind of sequence-like object. In order to learn a proper similarity between a query text and a text proposal, we adopt the normalized edit distances as the supervision of the pairwise loss function, which has been widely used in string matching. However, learning such similarity is not easy, as the similarity values of all word pairs for training are not evenly distributed in the similarity space. For example, the number of word pairs with low similarity is much larger than those with high similarity. As a result, the network is prone to distinguish dissimilar word pairs but difficult to similar word pairs. To ease this problem, we propose a word augmentation method: a series of pseudowords that are very similar to a given query text, are randomly generated, which are fed into the network together with the query text during the training process.

The contribution in this work is three-fold. First, we present a new end-to-end trainable network for joint optimizing scene text detection and cross-modal similarity learning, which is able to efficiently search text instances from natural images that contain a query text. Second, we

propose a word augmentation method by generating similar pseudowords as input queries, which enhances the discriminatory power of the learned similarity. Third, we collect and annotate a new dataset for Chinese scene text retrieval, consisting of 23 pre-defined query words and 1667 Chinese scene text images. This dataset is adopted to verify the effectiveness of text retrieval methods over non-Latin scripts.

2. Related work

Traditional text retrieval methods [1, 2, 31, 36] are proposed to retrieve cropped document text images. A popular pipeline of these methods is to represent both text string and word image and then calculate the distance between their representations. In [2], PHOC is first proposed to represent text string for retrieval. Then, Sudholt *et al.* [31] and Wilkinson *et al.* [36] respectively propose to predict PHOC and DCToW from word image using neural networks. Without carefully designing a hand-crafted representation of text, Gomez *et al.* [7] propose to learn the Levenshtein edit distance [16] between text string and word image. The learned edit distances are directly used to rank word images. However, the above methods consider perfectly cropped word images as inputs, rather than more universal and challenging scene text images.

Mishra *et al.* [26] first introduce the task of scene text retrieval, and adopt two separate steps for character detection and classification. Then, images are ranked by scores, *i.e.*, the probability of query text exists in the images. Ghosh *et al.* [4] propose a scene text retrieval method consisting of a selective search algorithm to generate text regions and a SVM-based classifier to predict corresponding PHOCs. Nevertheless, these methods ignore the relevance and complementarity between text detection and text retrieval. Moreover, it is inefficient to separate scene text retrieval into two sub-tasks. To integrate text detection and text retrieval in a unified network, Gomez *et al.* [6] introduce the first end-to-end trainable network for scene text retrieval, where all text instances are represented with PHOCs. Specifically, the method simultaneously predicts text proposals and corresponding PHOCs. Then, images in the gallery are ranked according to the distance between the PHOC of a query word and the predicted PHOC of each detected text proposal.

A few text spotting methods [12, 14] are also adopted for scene text retrieval. These methods first detect and recognize all possible words in each scene image. Then, the probability that the image contains the given query word is represented as the occurrences of the query word within those spotted words. Unlike retrieving text instances by matching a query word to the spotted words, our method directly measures cross-modal similarity between them.

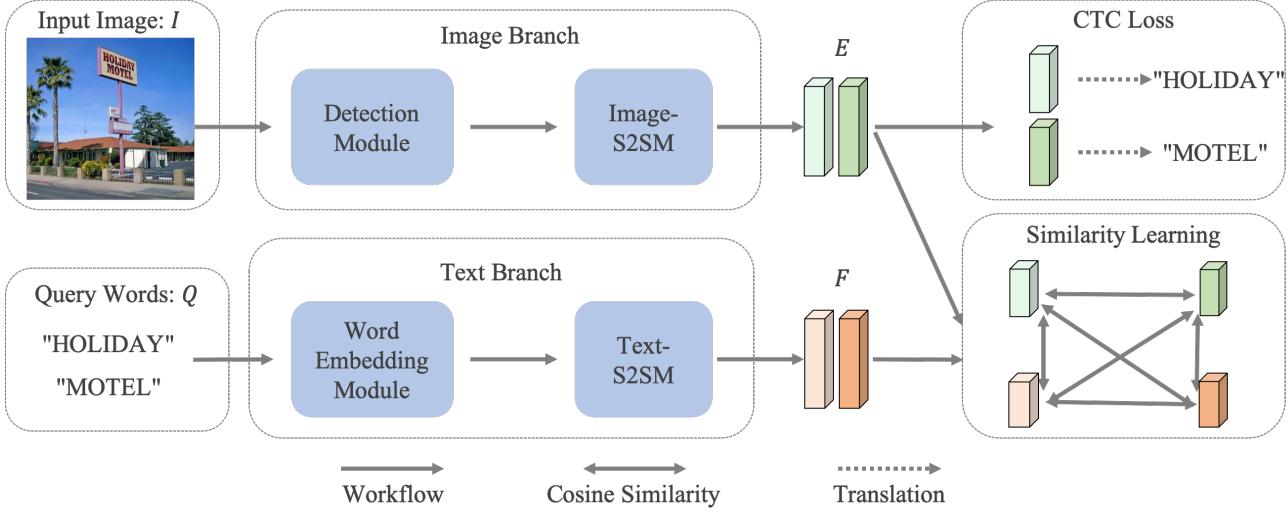


Figure 2. Illustration of our proposed framework. Given an image, text regions are detected and the corresponding features are extracted for the subsequent Image-S2SM. Meanwhile, features F of all query words Q are obtained through the word embedding module and the Text-S2SM. The whole network is jointly optimized by text detection task, text translation task and similarity learning task.

3. Methodology

As illustrated in Fig. 2, our network consists of two branches, *i.e.*, image branch for extracting features E of all possible text proposals and text branch that converts query words $Q = \{q_i\}_{i=1}^N$ into features F . Then, the pairwise similarity between F and E is calculated for ranking.

3.1. Image Branch

The image branch aims at extracting features of all possible text proposals. Unlike general objects, scene text usually appears in the form of a character sequence. Therefore, the sequence-to-sequence module (Image-S2SM), whose structure is detailed in Tab. 1, is used to enhance the contextual information of each text proposal. As shown in Fig. 2, there are two modules in the image branch, including the text detection module and Image-S2SM. To simplify the detection pipeline, the text detection module is based on an anchor-free detector [32], where the backbone is FPN [21] equipped with ResNet-50 [11].

Given an image, the detection module in the image branch first yields K text proposals. The corresponding region of interest (RoI) features $P = \{p_i\}_{i=1}^K$ are extracted via RoI-Align [10] and are fed into the subsequent Image-S2SM to generate the features $E \in \mathbb{R}^{K \times T \times C}$. T and C stand for the width and channel of RoI features.

3.2. Text Branch

Different from images, query words are a set of text strings that are unable to be directly forwarded by the neural network. So a word embedding module is employed to convey query words into features. Similar to Image-S2SM, the sequence-to-sequence module (Text-S2SM), whose struc-

ture is detailed in Tab. 1, is also used in this branch.

Word Embedding Module consists of an embedding layer and a bilinear interpolation operator. Specifically, given query words Q , the word q_i can be considered as a character sequence $(y_1, \dots, y_{|q_i|})$, where $|q_i|$ is the number of characters in q_i and y_j is the one-hot representation of the j -th character of q_i . The embedding layer first converts each character y_j into $2C$ dimensional feature, producing an embedded feature sequence for each word. Then, each feature sequence is concatenated and interpolated as a fixed-length feature $\hat{f}_i \in \mathbb{R}^{T \times 2C}$. Finally, all features $\{\hat{f}_i\}_{i=1}^N$ of Q are stacked as output features $\hat{F} \in \mathbb{R}^{N \times T \times 2C}$.

After the word embedding module for query words, the obtained features \hat{F} is projected to $F \in \mathbb{R}^{N \times T \times C}$ by Text-S2SM. Then, both features E and F are fed for the subsequent similarity learning task.

3.3. Similarity learning

After the features of text proposals and query words, $E \in \mathbb{R}^{K \times T \times C}$ and $F \in \mathbb{R}^{N \times T \times C}$, are extracted, the pairwise similarity between query words Q and features of text proposals P can be formulated as similarity matrix $\hat{S}(Q, P) \in \mathbb{R}^{N \times K}$. Here, the value of $\hat{S}_{i,j}(Q, P)$ is equal to the cosine similarity between features F_i and E_j , which is formulated via

$$\hat{S}_{i,j}(Q, P) = \frac{\tanh(V(F_i)) \tanh(V(E_j))^T}{\|\tanh(V(F_i))\| * \|\tanh(V(E_j))\|}. \quad (1)$$

V stands for the operator that reshapes a two-dimensional matrix into one-dimensional vector.

During training, the predicted similarity matrix $\hat{S}(Q, P)$ is supervised by the target similarity matrix $S(Q, P)$. Each target similarity, $S_{i,j}(Q, P)$ is the normalized edit distance

	Layers	Parameters (kernel, stride)	Output Size
Image-S2SM	conv layer $\times 2$	(3,(2,1))	$(N, 2C, h, T)$
	average on h	-	$(N, 2C, 1, T)$
	BiLSTM	-	(N, T, C)
Text-S2SM	conv layer	(1,(1,1))	$(N, 2C, 1, T)$
	BiLSTM	-	(N, T, C)

Table 1. Architectures of Image-S2SM and Text-S2SM. BiLSTM stands for bidirectional LSTM [13] layer.

between the corresponding word pairs (q_i, q_j) , which is defined as Eq. 2. *Distance* is the Levenshtein edit distance [16], $|q_i|$ denotes the character number of q_i .

$$S_{i,j}(Q, P) = 1 - \frac{\text{Distance}(q_i, q_j)}{\max(|q_i|, |q_j|)}. \quad (2)$$

Besides $\hat{S}(Q, P)$, both $\hat{S}(P, P) \in \mathbb{R}^{K \times K}$ and $\hat{S}(Q, Q) \in \mathbb{R}^{N \times N}$ are also calculated for assistant training. During inference, the similarity between q_i and an input image is equal to the maximum value of $\hat{S}_i(Q, P)$, which is used for ranking.

To close the semantic gap between visual feature E and textual feature F , the connectionist temporal classification [8] (CTC) loss is adopted for aligning visual feature to a text string. Particularly, for each E_i , a classifier consisting of a fully connected layer and a softmax layer is supervised by the associated sequence label of q_i .

3.4. Word augmentation strategy

We observe that the similarity learning task suffers from imbalanced distribution in the similarity space, as most random word pairs are dissimilar. As shown in Fig. 3 (a), for query words, the proportion of word pairs with low similarity is much larger than those with high similarity. As a result, the learned network tends to distinguish dissimilar word pairs but difficult to handle similar word pairs.

To alleviate this effect, a word augmentation strategy is proposed to randomly generate pseudowords that are very similar to the given query words. Then, both original query words and pseudowords are fed into the network during the training process. As illustrated in Fig. 3 (b), with such an augmentation strategy, the proportion of word pairs with high similarity is greatly enlarged, which eases the problem brought by uneven similarity distribution.

As illustrated in Fig. 4, we define four types of character edit operators, *i.e.*, inserting a random character behind the current character, deleting the current character, replacing the current character with a random one, and keeping the current character. For each character in the query word q_i , an edit operator is randomly chosen. Then, a pseudoword \tilde{q}_i similar with q_i is generated. Obviously, the higher the ratio of keeping the current character is, the more similar

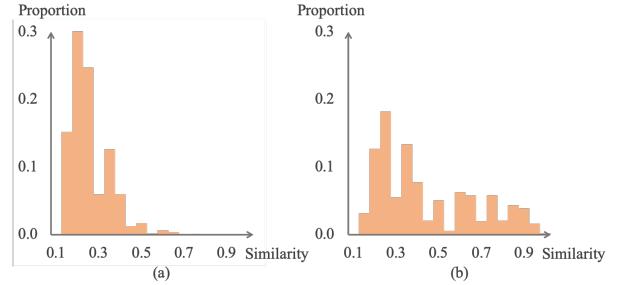


Figure 3. Similarity distribution of word pairs from query words Q in (a) and augmented query words \tilde{Q} in (b) on SynthText-900k.

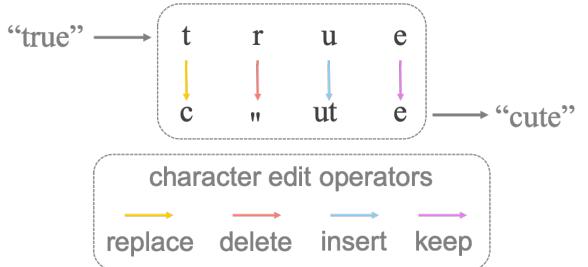


Figure 4. The process of word augmentation strategy. Given an input word “true”, this strategy outputs a word “cute”.

the pseudoword is with the input word. In our experimental setting, the sampling ratio among four character edit operators is set to 1 : 1 : 1 : 5. The algorithm description of the word augmentation strategy is summarized in Appendix A. Finally, query words $\tilde{Q} = \{\tilde{q}_i\}_{i=1}^{2N}$ consisting of original words Q and their associated pseudowords $\hat{Q} = \{\hat{q}_i\}_{i=1}^N$, are fed into text branch for training. In this case, the predicted pairwise similarity matrices are $\hat{S}(\tilde{Q}, P) \in \mathbb{R}^{2N \times K}$, $\hat{S}(P, P) \in \mathbb{R}^{K \times K}$ and $\hat{S}(\tilde{Q}, \tilde{Q}) \in \mathbb{R}^{2N \times 2N}$.

3.5. Loss functions

The objective function consists of three parts, which is defined as follows,

$$L = L_d + L_s + L_c, \quad (3)$$

where L_d is the detection loss in [32]. L_c is CTC loss for text translation task. L_s is the cross-modal similarity learning loss, which is calculated with Smooth-L1 loss L_r for regression. The loss function L_s is formulated as

$$\begin{aligned} L_s &= \frac{1}{K} \sum_i^K \max(L_r(\hat{S}_i(P, P), S_i(P, P))) \\ &+ \frac{1}{2N} \sum_i^{2N} \max(L_r(\hat{S}_i(\tilde{Q}, P), S_i(\tilde{Q}, P))) \\ &+ \frac{1}{2N} \sum_i^{2N} \max(L_r(\hat{S}_i(\tilde{Q}, \tilde{Q}), S_i(\tilde{Q}, \tilde{Q}))), \end{aligned} \quad (4)$$

where \hat{S} and S are the predicted similarity matrix and its associated target similarity matrix. $2N$ and K are the number of query words after augmentation and the number of text instances respectively.

4. Experiments

First, we introduce the datasets used in our experiments and the new dataset created by us. Then, the implementation details are given. Third, we evaluate our method and make comparisons with the state-of-the-art scene text retrieval/spotting approaches. Last, we provide the ablation studies and validate a potential application of our method.

4.1. Datasets

Street View Text dataset (SVT) [35] has 349 images collected from Google Street View. This dataset is divided into a train set of 100 images and a test set of 249 images. 427 annotated words on the test set are used as query text.

IIT Scene Text Retrieval dataset (STR) [26] consists of 50 query words and 10,000 images. It is a challenging dataset due to the variation of fonts, styles and view points.

Coco-Text Retrieval dataset (CTR) is a subset of Coco-Text [33]. We select 500 annotated words from Coco-Text as queries. Then, 7196 images in Coco-Text containing such query words are used to form this dataset.

SynthText-900k dataset [6] is composed of about 925,000 synthetic text images, generated by a synthetic engine [9] with slight modifications.

Multi-lingual Scene Text 5k dataset (MLT-5k) is a subset of MLT [27], which consists of about 5000 images containing text in English.

All text instances from these datasets are in English. In our experiments, SVT, STR and CTR are the testing datasets, while SynthText-900k and MLT-5k are only used for training the proposed model.

In order to validate our method’s effectiveness on non-Latin scripts, we create a new dataset for Chinese scene text retrieval, named as **Chinese Street View Text Retrieval** dataset (CSVTR¹). This dataset consists of 23 pre-defined query words in Chinese and 1667 Chinese scene text images collected from the Google image search engine. Each image is annotated with its corresponding query word among the 23 pre-defined Chinese query words. As shown in Fig. 5, most query words are the names of business places, such as *Seven Days Inn*, *China Construction Bank*, *Wallace*, and *Yunghe Soy Milk*.

4.2. Implementation details

The whole training process includes two steps: pre-trained on SynthText-900k and fine-tuned on MLT-5k. In

¹<https://github.com/lanfeng4659/STR-TDSL>



Figure 5. Examples from our CSVTR dataset.

the pre-training stage, we set the mini-batch to 64, and input images are resized to 640×640 . In the fine-tuning stage, data augmentation is applied. Specifically, the longer sides of images are randomly resized from 640 pixels to 1920 pixels. Next, images are randomly rotated in a certain angle range of $[-15^\circ, 15^\circ]$. Then, the heights of images are rescaled with a ratio from 0.8 to 1.2 while keeping their widths unchanged. Finally, 640×640 image patches are randomly cropped from the images. The mini-batch of images is kept to 16.

We optimize our model using SGD with a weight decay of 0.0001 and a momentum of 0.9. In the pre-training stage, we train our model for 90k iterations, with an initial learning rate of 0.01. Then the learning rate is decayed to a tenth at the 30kth iteration and the 60kth iteration, respectively. In the fine-tuning stage, the initial learning rate is set to 0.001, and then is decreased to 0.0001 at the 40kth iteration. The fine-tuning process is terminated at the 80kth iteration. For testing, the longer sides of input images are resized to 1280 while the aspect ratios are kept unchanged.

All experiments are conducted on a regular workstation with NVIDIA Titan Xp GPUs with Pytorch. The model is trained in parallel on four GPUs and tested on a single GPU.

4.3. Experimental results

The experimental results of previous state-of-the-art methods and our method are summarized in Tab. 2. Note that two most recent state-of-the-art scene text spotting approaches, namely Mask TextSpotter v3 [19] and ABCNet [23] are also adopted in the comparisons. Specifically, edit distances between a query word and the spotted words from scene images are used for text-based image retrieval. The retrieval results by Mask TextSpotter v3 and ABCNet are obtained with their officially released models^{2,3}.

²<https://github.com/MhLiao/MaskTextSpotterV3>

³https://github.com/Yuliang-Liu/bezier_curve_text_spotting

Method	SVT	STR	CTR	FPS
Mishra <i>et al.</i> [26]	56.24	42.70	-	0.10
Jaderberg <i>et al.</i> [14]	86.30	66.50	-	0.30
He <i>et al.</i> [12] (dictionary)	80.54	66.95	-	2.35
He <i>et al.</i> [12] (PHOC)	57.61	46.34	-	2.35
Gomez <i>et al.</i> [6]	83.74	69.83	41.05*	43.50
Gomez <i>et al.</i> [6] (MS)	85.18	71.37	51.42*	16.10
Mafla <i>et al.</i> [25]	85.74	71.67	-	42.20
ABCNet [23]	82.43*	67.25*	-	17.50
Mask TextSpotter v3 [19]	84.54*	74.48*	55.54*	2.40
Ours	89.38	77.09	66.45	12.00
Ours (MS)	91.69	80.16	69.87	3.80

Table 2. Performance comparisons (mAP scores) with the state-of-the-art text retrieval/spotting methods on SVT, STR and CTR. MS means multi-scale testing. * represents the result is obtained with the officially released code from authors. Results with (dictionary) are filtered using a pre-defined word dictionary. Results with (PHOC) denote that recognized words are transformed to PHOC for ranking.

4.3.1 Comparisons with state-of-the-art methods

We can observe that the proposed method outperforms other methods by a significant margin over all datasets. The method [14] exhibits the best performance among previous methods on SVT, but still performs worse than our method by 3.08%. Compared with other methods, Mask TextSpotter v3 performs best on STR and CTR, yet our method still obtains 2.61% and 10.91% improvements. Moreover, our method runs much faster than Mask Textspotter v3 with single scale testing. Due to the equipment with a fast detector, YOLO [28], these PHOC-based methods [6, 25] achieve the fastest inference speed. However, our method obtains significantly superior performance than PHOC-based methods among all datasets. The consistent improvements over both PHOC-based methods and end-to-end text spotters further demonstrate that it is essential to learn a cross-modal similarity for scene text retrieval.

To further boost the performance, we use multi-scale testing by combining the outputs under multiple resolutions of test images. In this experiment, longer sides of input images are resized to 960, 1280 and 1600. Compared with single scale testing, the performance is improved by 2.31%, 3.07% and 3.42% on SVT, STR and CTR.

Some qualitative results of our method are shown in Fig. 6. We can find that the proposed method can accurately localize and retrieve text instances from an image gallery with the given query word.

4.3.2 Chinese scene text retrieval

To further confirm the generality of our method over non-Latin scripts, we conduct experiments on the CSVTR dataset for Chinese scene text retrieval. Our method is compared with the recent end-to-end scene text retrieval method [25] with its officially released code⁴. Follow-

Method	Fea. Dim.	Char. Num.	mAP	FPS
Mafla <i>et al.</i> [25]	14266	1019	4.79	1.70
Ours	1920	1019	60.23	12.00
Ours [†]	1920	3755	50.12	12.00

Table 3. Feature dimension, retrieval performance and inference time comparisons on CSVTR dataset. *Fea. Dim.* and *Char. Num.* denote feature dimension and the number of character types.

ing [9], we first generate 70k synthesized images with horizontal text instances in Chinese and their corresponding annotations. Then, the models of our method and [25] are trained with this synthesized dataset and evaluated on the CSVTR dataset.

According to the definition of PHOC [2], the dimension of PHOC dramatically increases as the number of character types grows. Due to the limitation of GPU memory, only around 1500 Chinese character types are allowed for PHOC in [25] during training. In order to successfully apply the PHOC-based retrieval method to Chinese scene text retrieval, we combine 1000 most common Chinese characters and characters in CSVTR to construct the character set of 1019 categories. In Tab. 3, our method achieves 60.23% mAP under the setting of 1019 character types. However, the method in [25] only performs 4.79% mAP, and its inference speed decreases to just 1.7 FPS, while our method remains 12 FPS. Moreover, our method could support all 3755 character categories appearing in the training dataset and achieve 50.12% mAP. These results reveal that our method is more robust and can be easily generalized to non-Latin scripts. Some qualitative results of retrieving Chinese text instances are shown in Appendix B.

4.4 Ablation study

In this section, we first provide elaborate ablation studies to further verify the effectiveness of the proposed word augmentation strategy (WAS) and CTC loss. The results are

⁴<https://github.com/AndresPMD/Pytorch-yolo-phoc>



Figure 6. Retrieval results on STR. Only top 8 retrieved images are shown for the query word “coffee”.

Method	SVT	STR	CTR
Baseline	88.05	73.70	62.97
+ CTC	88.62	75.54	64.08
+ WAS	88.68	75.76	64.59
+ WAS + CTC	89.38	77.09	66.45
Separated	82.88	70.83	61.16

Table 4. The ablation studies on WAS and CTC. “+ Component” denotes adding “Component” to the Baseline method. Comparisons with the method learning in separated manner are also given. The numbers in the table stand for mAP scores.

reported in Tab. 4. The *Baseline* is the basic method based on similarity learning without WAS and CTC loss. Then we discuss the influences of assistant training by S(P, P) and S(Q, Q), and compare our cross-modal similarity learning method with the PHOC-based method on the same detection network in Tab. 5.

WAS. As we discussed before, WAS is important for cross-modal similarity learning. Compared with the *Baseline* method, WAS respectively increases the mAP by 0.63%, 2.06% and 1.62% on SVT, STR and CTR datasets. The results reveal the importance of WAS, as it can enlarge the proportion of similar word pairs to prevent the model from suffering from imbalanced sample distribution in the training stage. Such an augmentation strategy facilitates the retrieval model to better distinguish similar words.

CTC loss. Learning cross-modal similarity is of great importance in our method. CTC loss is adopted to align the cross-modal features of query words and text instances. From Tab. 4, we can observe that CTC loss also obtains improvements of 0.57%, 1.84% and 1.11% on SVT, STR and CTR, compared with the *Baseline* method. As shown

Method	SVT	STR	CTR
Detection+PHOC	80.87	60.00	42.89
Ours [‡]	85.28	66.50	51.79
Ours	87.78	70.18	53.90

Table 5. Ablation study with PHOC. All models are trained with SynthText-900k. [‡] denotes training without S(P, P) and S(Q, Q). The numbers in the table stand for mAP scores.

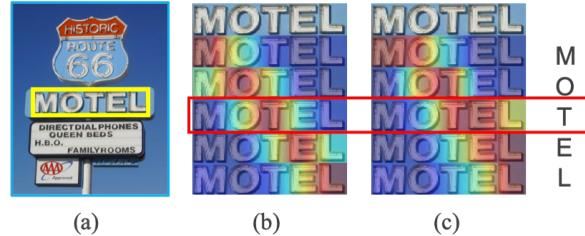


Figure 7. (a): The retrieved image with the query word “MOTEL”. The visualization of similarity between each character of “MOTEL” with text image when the network is trained with CTC loss (b) or without CTC loss (c).

in Fig. 7, such a process assists the image branch to focus on text regions and thus extract purified visual features, which facilitates the cross-modal similarity measure. Therefore, similarity learning can benefit from aligning the cross-modal features via the text recognition task.

After combining the two components, the performance is further improved on all datasets, which demonstrates WAS and CTC loss are complementary for text retrieval.

Ours vs. Separated. In general, scene text retrieval methods in an end-to-end manner could achieve better performance than methods with separated detection and re-

trieval, due to feature sharing and joint optimization. To confirm this assumption, we separate our scene text retrieval pipeline into scene text detection and text image retrieval. The text detection network is first trained. Then, text images are cropped according to annotated bounding boxes to train the scene text retrieval network. For a fair comparison, the scene text retrieval network consists of a ResNet-50 backbone and a similarity learning network. And the similarity learning network is composed of a word embedding module, Image-S2SM and Text-S2SM. During inference, text images are cropped according to the detection results from the text detection network, and then fed to the text image retrieval network for ranking. The results are shown in Tab. 4. Compared with the separated retrieval system, the end-to-end retrieval system can respectively improve the performance by 6.50%, 6.26% and 5.29% on SVT, STR and CTR, demonstrating its effectiveness.

Ours vs. PHOC. For a fair comparison with PHOC-based methods [6, 25] on the same text detector, we drop out the text branch and add a classifier behind Image-S2SM to predict the PHOC vector. The better results achieved in Tab. 5 demonstrate the effectiveness of the cross-modal similarity method. The reason behind might be PHOC embedding does not explicitly learn a faithful similarity between a word and a word image.

The impact of $S(P, P)$ and $S(Q, Q)$. Both $S(P, P)$ and $S(Q, Q)$ are only utilized for training. As shown in Tab. 5, the performance degrades 2.5% mAP on average among all datasets if they are removed.

4.5. Application to weakly supervised text annotation

In this section, we extend our retrieval method to a novel application. Nowadays, there are a large number of images with corresponding descriptions on the Internet. Given the associated words, our method is able to precisely localize them in images, which is essentially a weakly supervised text annotation task. Specifically, for each word appearing in the image, its region annotation can be formulated as the bounding box of text proposal that is most similar to the word. We evaluate the annotation quality using the scene text detection protocol in [15].

To demonstrate its effectiveness, a recent text detector DB [20] and an end-to-end text spotter Mask TextSpotter v3 [19] are adopted for comparisons. Firstly, the models of DB, Mask TextSpotter v3 and our proposed method are consistently trained with the SynthText-900k. Then, the detected results by DB, the spotted text boxes by Mask TextSpotter v3, and the retrieved text proposals by our method are used for evaluation on IC13 [15] and CTR. As shown in Fig. 8, our method outperforms Mask TextSpotter v3 with improvements of over 10.0%. Furthermore, compared with the text detector DB, our method achieves sig-

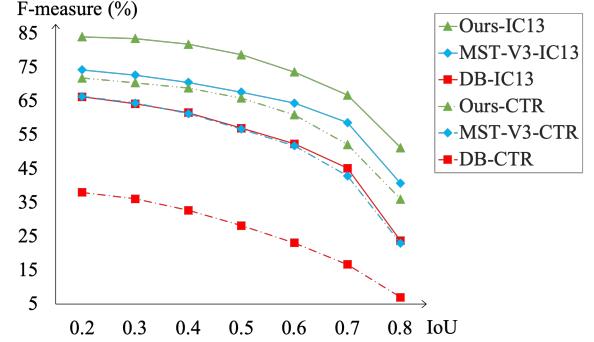


Figure 8. Comparison of text detection results on IC13 and CTR. MST-V3 denotes Mask TextSpotter v3.

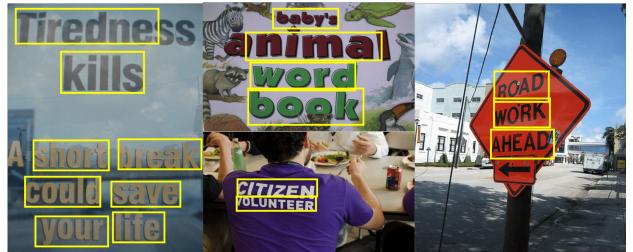


Figure 9. The bounding boxes annotated by our method on IC13 and CTR.

nificant improvements of over 20.0% and 35.0% on IC13 and CTR. Some qualitative results of the annotated text instances on IC13 and CTR are shown in Fig. 9. We can observe that the bounding boxes of most text instances are annotated accurately.

5. Conclusion

In this paper, we have presented an end-to-end trainable framework combining scene text detection and pairwise similarity learning, which could search the text instances that are the same or similar with a given query text from natural images. The experiments demonstrate that the proposed method consistently outperforms the state-of-the-art retrieval/spotting methods on three benchmark datasets. Besides, it has been shown that our framework could cope with Chinese scene text retrieval, which is more challenging for existing methods. In the future, we would like to extend this method in dealing with more complicated cases, such as multi-oriented or curve text.

References

- [1] David Aldavert, Marçal Rusiñol, Ricardo Toledo, and Josep Lladós. Integrating visual and textual cues for query-by-string word spotting. In *ICDAR*, 2013. 1, 2
- [2] Jon Almazán, Albert Gordo, Alicia Fornés, and Ernest Valveny. Word spotting and recognition with embedded attributes. *TPAMI*, 36(12):2552–2566, 2014. 1, 2, 6

- [3] Xiang Bai, Mingkun Yang, Pengyuan Lyu, Yongchao Xu, and Jiebo Luo. Integrating scene text and visual appearance for fine-grained image classification. *IEEE Access*, 6:66322–66335, 2018. 1
- [4] Suman K. Ghosh, Lluís Gómez, Dimosthenis Karatzas, and Ernest Valveny. Efficient indexing for query by string text retrieval. In *ICDAR*, 2015. 2
- [5] Suman K. Ghosh and Ernest Valveny. Query by string word spotting based on character bi-gram indexing. In *ICDAR*, 2015. 1
- [6] Lluís Gómez, Andrés Mafla, Marçal Rusiñol, and Dimosthenis Karatzas. Single shot scene text retrieval. In *ECCV*, 2018. 1, 2, 5, 6, 8
- [7] Lluís Gómez, Marçal Rusiñol, and Dimosthenis Karatzas. LSDE: levenshtein space deep embedding for query-by-string word spotting. In *ICDAR*, 2017. 1, 2
- [8] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *ICML*, 2006. 4
- [9] Ankush Gupta, Andrea Vedaldi, and Andrew Zisserman. Synthetic data for text localisation in natural images. In *CVPR*, 2016. 5, 6
- [10] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 3
- [11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3
- [12] Tong He, Zhi Tian, Weilin Huang, Chunhua Shen, Yu Qiao, and Changming Sun. An end-to-end text spotter with explicit alignment and attention. In *CVPR*, 2018. 1, 2, 6
- [13] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 4
- [14] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Reading text in the wild with convolutional neural networks. *IJCV*, 116(1):1–20, 2016. 1, 2, 6
- [15] Dimosthenis Karatzas, Faisal Shafait, Seiichi Uchida, Masakazu Iwamura, Lluis Gomez Bigorda, Sergi Robles Mestre, Joan Mas, David Fernandez Mota, Jon Almazan Almazan, and Lluis Pere De Las Heras. Icdar 2013 robust reading competition. In *ICDAR*, 2013. 8
- [16] Vladimir Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklad*, volume 10, pages 707–710, 1966. 2, 4
- [17] Hui Li, Peng Wang, and Chunhua Shen. Towards end-to-end text spotting with convolutional recurrent neural networks. In *ICCV*, 2017. 2
- [18] Minghui Liao, Pengyuan Lyu, Minghang He, Cong Yao, Wenhao Wu, and Xiang Bai. Mask textspotter: An end-to-end trainable neural network for spotting text with arbitrary shapes. *TPAMI*, 2019. 2
- [19] Minghui Liao, Guan Pang, Jing Huang, Tal Hassner, and Xiang Bai. Mask textspotter v3: Segmentation proposal network for robust scene text spotting. In *ECCV*, 2020. 5, 6, 8
- [20] Minghui Liao, Zhaoyi Wan, Cong Yao, Kai Chen, and Xiang Bai. Real-time scene text detection with differentiable binarization. In *AAAI*, 2020. 8
- [21] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 3
- [22] Xuebo Liu, Ding Liang, Shi Yan, Dagui Chen, Yu Qiao, and Junjie Yan. Fots: Fast oriented text spotting with a unified network. In *CVPR*, 2018. 2
- [23] Yuliang Liu, Hao Chen, Chunhua Shen, Tong He, Lianwen Jin, and Liangwei Wang. Abcnet: Real-time scene text spotting with adaptive bezier-curve network. In *CVPR*, 2020. 5, 6
- [24] Pengyuan Lyu, Minghui Liao, Cong Yao, Wenhao Wu, and Xiang Bai. Mask textspotter: An end-to-end trainable neural network for spotting text with arbitrary shapes. In *ECCV*, 2018. 2
- [25] Andrés Mafla, Rubén Tito, Sounak Dey, Lluís Gómez, Marçal Rusiñol, Ernest Valveny, and Dimosthenis Karatzas. Real-time lexicon-free scene text retrieval. *Pattern Recognition*, 110:107656, 2020. 6, 8
- [26] Anand Mishra, Kartikey Alahari, and C. V. Jawahar. Image retrieval using textual cues. In *ICCV*, 2013. 1, 2, 5, 6
- [27] Nibal Nayef, Yash Patel, Michal Busta, Pinaki Nath Chowdhury, Dimosthenis Karatzas, Wafa Khelif, Jiri Matas, Umapada Pal, Jean-Christophe Burie, Cheng-lin Liu, et al. Icdar2019 robust reading challenge on multi-lingual scene text detection and recognition—rrc-mlt-2019. In *ICDAR*, 2019. 5
- [28] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *CVPR*, 2017. 6
- [29] Xuejian Rong, Chucai Yi, and Yingli Tian. Recognizing text-based traffic guide panels with cascaded localization network. In *ECCV*, 2016. 1
- [30] Hao Song, Hongzhen Wang, Shan Huang, Pei Xu, Shen Huang, and Qi Ju. Text siamese network for video textual key frame detection. In *ICDAR*, 2019. 1
- [31] Sebastian Sudholt and Gernot A Fink. Phocnet: A deep convolutional neural network for word spotting in handwritten documents. In *ICFHR*, 2016. 2
- [32] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *ICCV*, 2019. 3, 4
- [33] Andreas Veit, Tomas Matera, Lukas Neumann, Jiri Matas, and Serge Belongie. Coco-text: Dataset and benchmark for text detection and recognition in natural images. *arXiv preprint arXiv:1601.07140*, 2016. 5
- [34] Hao Wang, Pu Lu, Hui Zhang, Mingkun Yang, Xiang Bai, Yongchao Xu, Mengchao He, Yongpan Wang, and Wenyu Liu. All you need is boundary: Toward arbitrary-shaped text spotting. In *AAAI*, pages 12160–12167, 2020. 2
- [35] Kai Wang, Boris Babenko, and Serge J. Belongie. End-to-end scene text recognition. In *ICCV*, 2011. 5
- [36] Tomas Wilkinson and Anders Brun. Semantic and verbatim word spotting using deep neural networks. In *ICFHR*, 2016. 2
- [37] Xiao Yang, Dafang He, Wenyi Huang, Alexander Ororbia, Zihan Zhou, Daniel Kifer, and C. Lee Giles. Smart library: Identifying books on library shelves using supervised deep learning for scene text reading. In *JCDL*, 2017. 1

- [38] Yingying Zhu, Minghui Liao, Mingkun Yang, and Wenyu Liu. Cascaded segmentation-detection networks for text-based traffic sign detection. *TITS*, 19(1):209–219, 2018. 1

Appendix

A. The details of word augmentation strategy

The algorithm below describes the details of the word augmentation strategy. Given an input word M , this algorithm outputs a word W . MD_4 stands for multinomial distribution on the variables that indicate the four types of character edit operators. And p_i is the probability of sampling the i th character edit operator. In our experimental setting, the sampling ratio among four character edit operators is set to $1 : 1 : 1 : 5$. l is the number of character types. n is the character number of the input word M . j is a random number in the range $[0, l - 1]$.

Algorithm 1 Word Augmentation Strategy

Require:

Character set: $B = \{b_0, b_1, \dots, b_{l-1}\}$.

Multinomial distribution: $MD_4(4 : p_1, \dots, p_4)$.

Input: Word $M = [m_0, m_1, \dots, m_{n-1}]$, $m_i \in B$.

Ensure:

Output: Word $W = []$.

1: Generate random sequence $A = (a_0, \dots, a_{n-1})$,

$A \sim MD_4(4 : p_1, \dots, p_4)$,

$a_i \in \{\text{insert}, \text{delete}, \text{replace}, \text{keep}\}$.

2: **for** $i=0$; $i < n$; $i++$ **do**

3: **if** $a_i = \text{insert}$ **then**

4: $W.\text{append}(m_i)$

5: $W.\text{append}(b_j)$

6: **end if**

7: **if** $a_i = \text{delete}$ **then**

8: continue

9: **end if**

10: **if** $a_i = \text{replace}$ **then**

11: $W.\text{append}(b_j)$

12: **end if**

13: **if** $a_i = \text{keep}$ **then**

14: $W.\text{append}(m_i)$

15: **end if**

16: **end for**

B. Examples of Chinese scene text retrieval results

Some qualitative results of Chinese scene text retrieval are shown in Fig. 10. We can observe that our method not only obtains the correct ranking lists but also localizes the bounding boxes of text instances.

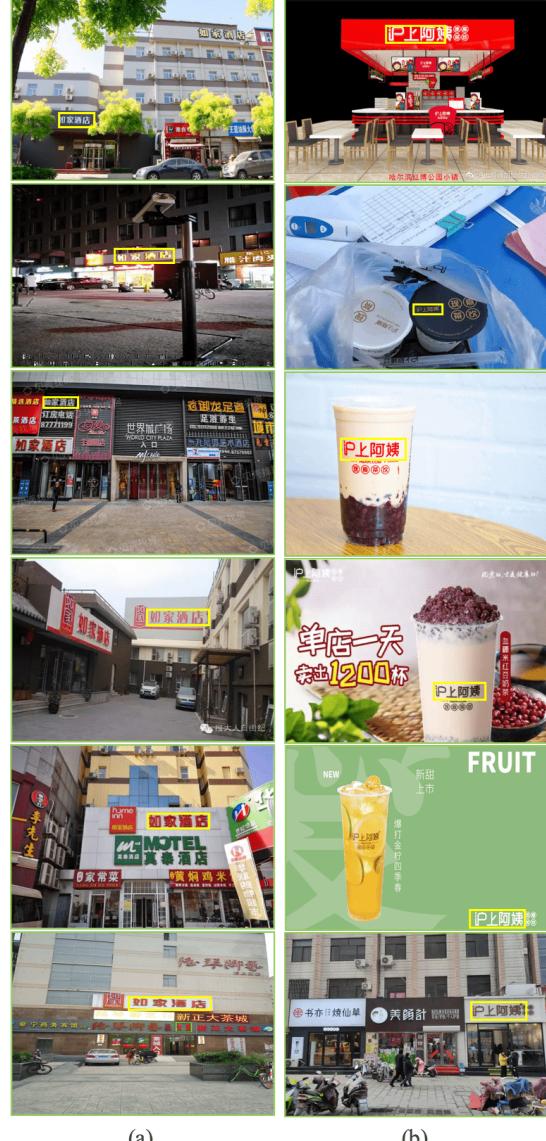


Figure 10. Retrieval results on CSVTR. Only top 6 retrieved images are shown for query words “Home Inn” in (a) and “Auntea Jenny” in (b).