SEMANTIC-BASED SENTENCE RECOGNITION IN IMAGES USING BIMODAL DEEP LEARNING

Yi Zheng[⋆] *Qitong Wang*[†] *Margrit Betke*[⋆]

[⋆] Department of Computer Science, Boston University, MA, USA
[†] Department of Electrical and Computer Engineering, Virginia Tech, VA, USA

2021 IEEE International Conference on Image Processing (ICIP) | 978-1-6654-4115-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICIP42928.2021.9506688

ABSTRACT

The accuracy of computer vision systems that understand sentences in images with text can be improved when semantic information about the text is utilized. Nonetheless, the semantic coherence within a region of text in natural or document images is typically ignored by state-of-the-art systems, which identify isolated words or interpret text word by word. However, when analyzed together, seemingly isolated words may be easier to recognize. On this basis, we propose a novel "Semantic-based Sentence Recognition" (SSR) deep learning model that reads text in images with the help of understanding context. SSR consists of a Word Ordering and Grouping Algorithm (WOGA) to find sentences in images and a Sequence-to-Sequence Recognition Correction (SSRC) model to extract semantic information in these sentences to improve their recognition. We present experiments with three notably distinct datasets, two of which we created ourselves. They respectively contain scanned catalog images of interior designs and photographs of protesters with hand-written signs. Our results show that SSR statistically significantly outperforms baseline methods that use state-of-the-art single-word-recognition techniques. By successfully combining both computer vision and natural language processing methodologies, we reveal the important opportunity that bi-modal deep learning can provide in addressing a task which was previously considered a single-modality computer vision task.

Index Terms— Text recognition, bi-modal, deep learning, new labeled datasets

1. INTRODUCTION

Recognizing text in images is a research problem that has attracted significant interest due to its importance in document image analysis, image retrieval, scene understanding, and assistance to people with visual impairments. Early work focused on images of printed documents, which can be interpreted with traditional optical character recognition techniques. More recent work shifted to recognizing text in natural scene images with deep convolutional neural networks (CNNs) [1, 2]. Text recognizer takes input from the text detector in the form of a cropped image (i.e., bounding box or polygon) that contains the word. In this paper, we proposes a text recognizer called **SSR** for Semantic-based Sentence Recognition.

Although tremendous efforts have been devoted to improving the performance of single-word recognition models, being able to understand text in images automatically is still very challenging and remains an open problem, even if an accurate bounding box of the text is given. This also applies to the seemingly easy domain of document images that include photographs of natural scenes with text overlays. State-of-the-art methods, trained on existing datasets, treat every occurrence of text in an image as an isolated word region that needs to be interpreted individually. To create a research challenge for recognizing text in images holistically, instead of word-by-word, we provide two new labeled datasets called "Text-containing Protest Image Dataset" (**TPID**), and "Interior Design Dataset" (**IDD**), which we make publicly available [https://github.com/ivc-yz/SSR]. The labels are polygons around each word region and the words themselves (their character encoding). The datasets contain images of natural scenes with multi-word phrases, sentences, or paragraphs, a property that is rare in text image datasets. Existing datasets of outdoor scenes typically only contain single-word text, e.g., on traffic signs, street signs, or store name signs on building facades (only exception we are aware of is the BDI dataset [3], which we also include in our experiments).

The innovative insight that our paper offers is that images with word groups contain semantic information that should not be ignored but exploited. The ability of a model to read text should improve when semantic information is available. In this paper, we show how a deep learning model can be designed and trained to take advantage of semantic information in order to recognize multi-word text in images.

2. METHODOLOGY

2.1. Overview

Our Semantic-based Sentence Recognition (SSR) system, illustrated in Figure 1, consists of four components that collectively recognize text in images by understanding context.

The first two components of our SSR framework have been shown to work well for single-word recognition. The first component, the rectifier, processes each image region that contains a single word by cropping it from the original image, relying on the input coordinates of its bounding polygon (blue outlines in the input image in Fig. 1). A perspective transformation algorithm then rectifies these word regions, converting quadrilateral subimages into axis-aligned subimages (see the word images surrounded by the dashed red line in Fig. 1). For single-word recognition (SWR), we use an existing text recognition model (any state-of-the-art single-word recognition model can be applied here.

The third and fourth components of our SSR framework are our innovation: WOGA uses the original polygon coordinates of each rectified word region to produce phrases, sentences, or paragraphs (see green dashed line in Fig. 1). A sequence-to-sequence deep network, called SSRC, is then trained to solve the task of correcting the words in these phrases, sentences, or paragraphs (dotted blue line in Fig. 1).

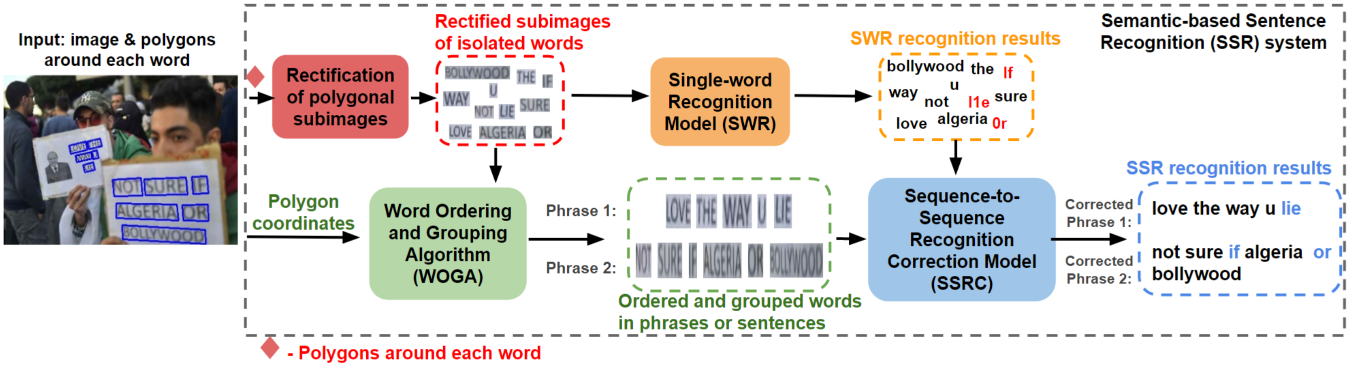
Input: image & polygons around each word
Rectified subimages of isolated words
Rectification of polygonal subimages
BOLLYWOOD THE IF U WAY NOT LIE SURE LOVE ALGERIA OR
Single-word Recognition Model (SWR)
SWR recognition results
bollywood the If u way not l1e sure love algeria 0r
Semantic-based Sentence Recognition (SSR) system
Polygon coordinates
Word Ordering and Grouping Algorithm (WOGA)
Phrase 1:
Phrase 2:
LOVE THE WAY U LIE
NOT SURE IF ALGERIA OR BOLLYWOOD
Ordered and grouped words in phrases or sentences
Sequence-to-Sequence Recognition Correction Model (SSRC)
SSR recognition results
Corrected Phrase 1:
love the way u lie
Corrected Phrase 2:
not sure if algeria or bollywood
- Polygons around each word

Fig. 1: Semantic-based Sentence Recognition (SSR) Architecture. In the example, the phrases on two protest signs are recognized by ordering and grouping the words on each sign and correcting the initially misidentified words *lie, if,* and *or*. In our experiments, we compare the accuracy of the SWR recognition results (orange) with the accuracy of the SSRC-produced SSR recognition results (blue).

Algorithm 1 Word Ordering and Grouping Algorithm

```
Input: Polygon coordinates and rectified subimages
// Ordering words:
Create directed graph G with costs on vertices & edges
Repeat
    Set the flow number to 1.
    Solve the min-cost flow problem which yields $S_i$ for a flow path $i$. Trace the flow path $i$ to produce the list of nodes $L_i$.
    Use the order of the nodes in $L_i$ to order the word regions that correspond to the nodes in $L_i$.
    Combine the word regions into region $R_i$.
    Delete the nodes on the list $L_i$ from $G$
Until $G$ is empty
// Grouping regions of words:
Repeat
    Randomly pick $R_i$ and find all $R_j$'s that satisfy conditions $(C_1)$ and $(C_2)$ and group them into the same phrase.
    Remove the regions included in that phrase from further consideration.
Until all regions are grouped
Output: Ordered words grouped in phrases
```

2.2. Word Ordering and Grouping Algorithm

WOGA arranges isolated words in an image into the correct logical order and then groups them into phrases or sentences that belong together. The pseudo code of WOGA is shown in Algorithm 1.

WOGA is inspired by a min-cost flow method [4], which has been successfully applied to the text detection problem [5]. In our case, isolated word regions correspond to vertices of a graph, and a group of ordered word regions corresponds to a flow in that graph. WOGA determines a cost between neighboring word regions and the probability of choosing a word region to be the starting and ending point of a sentence.

All word region candidates are first sorted according to their horizontal coordinates based on the assumption that sentences usually start from the left to the right. First, for each word region A and its corresponding vertex v_A, the directed edge from v_A to v_B, the vertex corresponding to word region B, is restricted by the three constraints (the symbols used are defined in Figure 2(a), and T_H, T_O, and T_A are fixed thresholds):

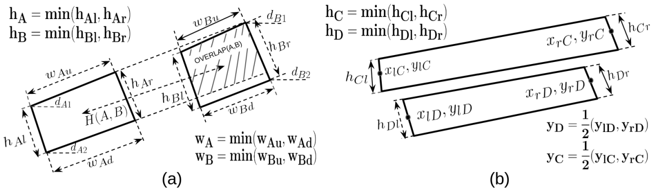
$h_A = \min(h_{Al}, h_{Ar})$
$h_B = \min(h_{Bl}, h_{Br})$
w_{Au}
h_{Ar}
d_{A1}
$H(A,B)$
h_{Al}
d_{A2}
w_{Ad}
w_{Bu}
d_{B1}
OVERLAP(A,B)
h_{Br}
h_{Bl}
d_{B2}
w_{Bd}
$w_A = \min(w_{Au}, w_{Ad})$
$w_B = \min(w_{Bu}, w_{Bd})$
(a)
$h_C = \min(h_{Cl}, h_{Cr})$
$h_D = \min(h_{Dl}, h_{Dr})$
x_rC, y_rC
h_{Cr}
h_{Cl}
x_lC, y_lC
x_rD, y_rD
h_{Dr}
h_{Dl}
x_lD, y_lD
$y_D = \frac{1}{2}(y_{lD}, y_{rD})$
$y_C = \frac{1}{2}(y_{lC}, y_{rC})$
(b)

Fig. 2: Definitions of symbols: (a) A and B are two word regions. (b) C and D are two regions of ordered and combined word regions.

(1) The horizontal distance between A and B should satisfy the condition $(H(A,B) - w_A/2 - w_B/2)/min(h_A, h_B) \leq T_H$.

(2) The orientation similarity between A and B should satisfy *OVERLAP(A,B)/AREA(B)* $\geq T_O$ and $\min(abs(d_{A1}-d_{B1}), abs(d_{A2}-d_{B2})) \leq T_A$.

(3) The size difference of A and B should satisfy the condition $abs(h_A - h_B)/\min(h_A, h_B)$.

To each directed edge, we assign the neighbor cost

$$C_n = \gamma\, D(A,B) + (1-\gamma)\, S(A,B), \quad (1)$$

where γ is a weighing factor, $D(A,B) = H(A,B)/\frac{1}{2}(w_A + w_B)$ is the Euclidean distance between the centers of word regions A and B as normalized by the mean of their window weights, and $S(A,B) = abs(h_A - h_B)/\min(h_A, h_B)$ is the size difference of A and B. Each vertex is defined as a source and a sink respectively, and the other vertices are connected to both, where the edge connecting with the source has an entry cost C_{en}, and the edge connecting with the sink has an exit cost C_{ex}. The entry cost is defined as

$$C_{en}(A) = \max_i (P(v_i \rightarrow v_A)), \quad (2)$$

where P represents the probability of any vertex v_i that could reach v_A in the directed graph. If no v_i reaches v_A, the chance of a flow starting at v_A is large, which is consistent with a small entry cost at v_A. In this case, $C_{en}(A)$ is set to 0. The exit cost $C_{ex}(A)$ can be similarly defined except that v_i ranges over all the vertices that could be reached by v_A. The total cost C_{total} of a flow in graph G can thus be defined. Minimization of C_{total} can be efficiently solved by the min-cost flow algorithm, yielding

$$C_{\text{total}} = \sum_i C_{en}(i)\, f_{en,i} + \sum_{i,j} C_n(i,j)\, f_{i,j} + \sum_i C_{ex}(i)\, f_{i,ex}, \quad (3)$$

Table 1: Examples of inputs to our SSRC.

Original sequence	SSRC input sequence
sitting room	$\langle GO\rangle$ ‘ s i t t i n g ‘ r o o m ‘ $\langle END\rangle$
black or yellow-red	$\langle GO\rangle$ ‘ b l a c k ‘ o r ‘ y e l l o w - r e d ‘ $\langle END\rangle$

where $C_n(i, j)$ is a neighbor cost between v_i and v_j, $C_{en}(i)$ and $C_{ex}(i)$ are the respective entry and exit costs of vertex i, and variables $f_{i,j}$, $f_{en,i}$ and $f_{i,ex}$ should be either 0 or 1 to enforce that each vertex belongs to at most one flow and they are determined while solving the min-cost flow problem. As a result, the min-cost flow "prefers" a region of ordered word regions that have similar sizes and are close to each other (smaller C_n) and a word region that has high probability to be entry or exit (smaller C_{en} and C_{ex}). After finding all regions of ordered word regions, WOGA groups them into the same region by checking if they satisfy the following conditions:

(C_1) $x_{lD} < x_{rC}$ and $x_{lC} < x_{rD}$.

(C_2) $abs(y_C - y_D)/min(h_C, h_D) < T_Y$.

The symbols used in the conditions above are illustrated in Fig 2(b). Extensive tests on the training datasets show that by setting T_H, T_O, T_A and T_Y to 0.7, 0.5, 10, and 2, respectively, WOGA groups word regions correctly. We set γ empirically to 0.7 so that the distance cost will penalize more than the size difference cost. Sample results of WOGA are shown in Fig. 3.

2.3. Sequence-to-Sequence Recognition Correction (SSRC) Model

Inspired by a sequence-to-sequence-based approach [6] that solves the NLP task of correcting spelling errors, we propose an attention-based[7] sequence-to-sequence recognition correction (SSRC) model, which can generates a "focus range" to indicate which parts of the input sequence should be focused on. SSRC, whose encoder and decoder are four-layer Bidirectional LSTM, can output a variable-length information sequence from a variable-length input sequence.

Before training our model, we need to transform the input sequences into a form that our model can understand. Some examples are shown in Table 1. We treat each character as one word, then split sentences into words using the grave accent (‘) and split each word into characters using the space symbol. The beginning of the sequence is marked as $\langle GO\rangle$, and the end of the sequence is marked as $\langle END\rangle$.

In this framework, The input sequence $\mathbf{x}$ is transferred into a context vector sequence c by a bidirectional LSTM-based encoder for the purpose of making the model to more effective in combining "front and back memory information." The encoder process [8] can be defined as

$$\overrightarrow{h_t} = f(x_t, \overrightarrow{h_{t-1}}) \text{ and } \overleftarrow{h_t} = f(x_t, \overleftarrow{h_{t-1}}), \tag{4}$$

where $\{\overrightarrow{h_1}, ..., \overrightarrow{h_t}, \overleftarrow{h_1}, ..., \overleftarrow{h_t}\} \in \mathbb{R}^{2t}$ are the encoder hidden states at time t, where t is the length of the input sequence of the SSRC model.

The decoder is trained to predict the next character y_t, given the context vector sequence c and all the previously predicted characters $y_1, y_2, ..., y_{t-1}$. The probability for the output sequence $\mathbf{y}$ can be defined as

$$P(\mathbf{y}) = \prod_{t=1}^{T} P(y_t | y_1, ..., y_{t-1}, c), \tag{5}$$

where the conditional probability can be defined as

$$P(y_t | y_1, ..., y_{t-1}, \mathbf{x}) = g(y_{t-1}, s_t, c_t), \tag{6}$$

with $s_t = f(s_{t-1}, y_{t-1}, c_t)$ denoting the decoder hidden state at time t.

In the attention-based sequence-to-sequence model [7, 9], the mapping from each context vector c_i to the encoder hidden state $\{\overrightarrow{h_1}, ..., \overrightarrow{h_t}, \overleftarrow{h_1}, ..., \overleftarrow{h_t}\}$ is computed as

$$c_i = \sum_{j=1}^{t} \alpha_{ij} \overleftarrow{h_j} + \sum_{j=1}^{t} \alpha_{ij} \overrightarrow{h_j}, \tag{7}$$

where the weight α_{ij} for each h_j can be computed by

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})}, \tag{8}$$

and where $e_{ij} = a(s_{i-1}, h_j)$ scores the match between input character j and output character i.

During the training process, the training loss is computed on the output of the decoder at the character level. At the each time step t, the implemented loss function is the cross-entropy loss per time step that is relevant to the previous time step [6], which is computed as

$$\text{Loss}(x, y) = -\sum_{t=1}^{T} \log(P(y_t | x, y_{t-1}, y_{t-2}, ..., y_1)). \tag{9}$$

3. EXPERIMENTAL METHODOLOGY

We evaluated the proposed SSR method on three datasets and compared its performance to the performance of two baseline SWR methods. State-of-the-art single-word recognition models use standard benchmark datasets such as [10, 11, 12, 13, 14, 15, 16, 17]. It is important to note that the text in these image datasets is typically a single word or several isolated words that are not relevant to each other semantically. So the text in the image cannot express unified semantic information. SSR is specifically designed to use semantic information to improve the prediction of any text recognition baseline module. In order to test the benefit of using semantic information in SSR, we work with newly created datasets, which we describe in detail next.

3.1. Benchmark Datasets

As mentioned above, we here introduce the **Interior Design Dataset** (IDD). It consists of 7,708 images of scanned product catalogues for interior design and decoration. We selected 4,708 of them as training images, 1,500 as validation images, and the remaining 1,500 for testing. The document images used for training contain more than 600,000 image regions with text and the number of image regions with text used for testing are 251,074. The ground-truth labels are the polygon coordinates of each image region that contains a word and a textual representation (i.e., ASCII representation) of the word itself.

We also created a subset of the UCLA Protest Image Dataset [18], which is a collection of social media images that can be used to analyze protest activities in street scenes. The original dataset consists of 40,764 images, among which 11,659 images show a protest. Among the protest images, we identified 816 images that contain

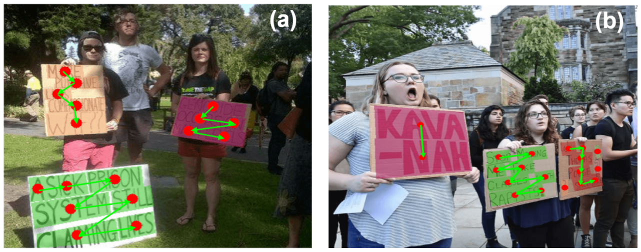

Fig. 3: Visualization of WOGA on TPID images. The center of each recognized word is shown as a red dot. A green arrow indicates the semantic connection of the words within a text region.

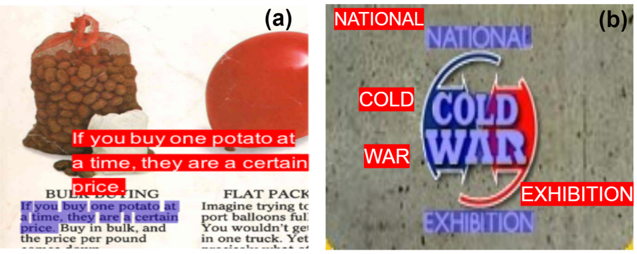

Fig. 4: Sample results of SSR. In the IDD image in (a), the baseline method omits two occurrences of the word ’a,’ but SSR correctly predicts these two words based on the context of the sentence. In the BDI images in (b), the baseline method incorrectly predicts ’WAR’ as ’WAS’ (probably due to the special font), but SSR makes a correct prediction based on context of the phrase ’cold war’.

mostly hand-made signs and text that is hand written and select 656 of them as training images and 160 as testing images. We created ground-truth polygons and textual representations of all the words on every sign. We refer to the protest-sign image collection as the **Text-containing Protest Image Dataset** (TPID). The total number of words for testing is 2,293. We make IDD and TPID publicly available for download [http://anonymous], and thus enable other researchers to engage in semantic-based scene text recognition work.

The **Born-Digital Images dataset** (BDI) [3] is one standard benchmark. BDI consists of 410 images for training and 141 images for testing. Since BDI was created as a text reading challenge, the test set was published without ground truth. We therefore selected the 410 BDI training images as a test set to evaluate our SSR method.

4. EXPERIMENTAL RESULTS

The sequence-to-sequence model is widely used by start-of-the-art SWR models. We used **CTC-based**[19] and **Attention-based** [20] sequence-to-sequence-based SWR models as a component of SSR and as baseline methods for performance comparison. Since the state-of-the-art changes quickly in SWR [21, 22], the above two models may not be the currently strongest text recognition models. However, our emphasis here is to show that semantics-based processing can improve text recognition accuracy.

The proposed CTC-based SSR system recognizes 226,067 out of 251,074 (90%) words in IDD, 1,630 out of 2,293 (71%) words in TPID, and 2,738 out of 3,558 (77%) words in BDI correctly (Table 2(a)). Our results reveal that it outperforms the CTC-based SWR for all datasets. In particular, on IDD, the CTC-based SSR beats the CTC-based SWR by 4.7 percent points and, on TPID, the CTC-based SSR beats the CTC-based SWR by 5.5 percent points. On BDI, the CTC-based SSR beats the CTC-based SWR by 2.2 percent points.

Table 2: Accuracy (%) and p-value (%) of the CTC and attention-based versions of our SSR method and the corresponding SWR baseline methods when tested on three datasets IDD, TPID, and BDI.

(a) accuracy	IDD	TPID	BDI
CTC-based SWR	85.36	65.63	74.73
CTC-based SSR	**90.04**	**71.09**	**76.95**
Attention-based SWR	93.80	78.65	86.42
Attention-based SSR	**96.34**	**89.39**	**88.11**
(b) p-value	IDD	TPID	BDI
CTC-based	$\ll 0.01$	$\ll 0.01$	2.9
Attention-based	$\ll 0.01$	$\ll 0.01$	3.3

The attention-based SSR beats the attention-based SWR by 2.5 percent points on IDD, 10.7 percent points on TPID, and 1.7 percent points on BDI.

To determine whether the accuracy improvement between SWR and SSR may be due chance, we performed a statistical significance analysis. The p-value measures the confidence of obtaining such an increase based on the sample size of the dataset. We use the N-1 Chi-Squared test [23, 24] to calculate the p-values in Table 2(b). The first row shows the p-values for evaluating the improvement between CTC-based SSR and SWR, and the second row the p-values for evaluating the improvement between Attention-based SSR and SWR, on three benchmark datasets. A typical threshold for declaring statistical significance is a p-value of less than 0.05. Our result shows that the impact of SSR is statistically significant on all datasets. For example, we are 99.99 % (1 minus p-value) confident to declare that the accuracy increase of the CTC-based SSR on IDD is statistical significant. For BDI, we are 97.13% confident to declare a gain of 2.2 percent points.

The high accuracy rate of SSR (e.g., 96.3% for the attention-based SSR) on the document images (IDD) is due to the fact that the task is relatively easy. The words are all machine generated, horizontally aligned, and the characters within a word have fixed fonts. This reduces the difficulty for SWR and WOGA to be effective. In contrast, the task of interpreting a word in a protest image is significantly more difficult. Due to the diversity of handwritten word layouts in TPID, the tasks of word ordering and grouping are not easy. At the same time, the variability of the aspect ratio and font type of the handwritten word also makes it more challenging to order and group the words on the protest signs in TPID.

5. CONCLUSIONS

In this work, we proposed a new deep learning model called SSR. This model can efficiently understand the context between regions of text or between words in images. SSR can extract sentences or paragraphs from images instead of only isolated text regions or words like state-of-the-art frameworks do. By ordering words and grouping them into phrases, sentences, or paragraphs, and interpreting on semantic information, our model is able to effectively improve prediction results compared to single-word recognition approaches. Our experimental analysis shows that this improvement is statistically significant. The combination of bi-modal information to obtain improved prediction results is of great significance to the research work in the field of computer vision.

6. REFERENCES

[1] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman, "Reading text in the wild with convolutional neural networks," *International Journal of Computer Vision*, vol. 116, no. 1, pp. 1–20, Jan 2016.

[2] Baoguang Shi, Xiang Bai, and Cong Yao, "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition," arXiv:1507.05717, 2015.

[3] Dimosthenis Karatzas, Sergi Robles Mestre, Joan Mas, Farshad Nourbakhsh, and Partha Pratim Roy, "ICDAR 2011 robust reading competition - challenge 1: Reading text in born-digital images (web and email)," in *2011 International Conference on Document Analysis and Recognition, ICDAR 2011, Beijing, China, September 18-21, 2011*. 2011, pp. 1485–1490, IEEE Computer Society.

[4] Andrew V. Goldberg, "An efficient implementation of a scaling minimum-cost flow algorithm," *Journal of Algorithms*, vol. 22, pp. 1–29, 1992.

[5] Shangxuan Tian, Yifeng Pan, Chang Huang, Shijian Lu, Kai Yu, and Chew Lim Tan, "Text flow: A unified text detection system in natural scene images," *CoRR*, vol. abs/1604.06877, 2016.

[6] Shaona Ghosh and Per Ola Kristensson, "Neural networks for text correction and completion in keyboard decoding," *CoRR*, vol. abs/1709.06429, 2017.

[7] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, "Neural machine translation by jointly learning to align and translate," *CoRR*, vol. abs/1409.0473, 2014.

[8] Sina Ahmadi, "Attention-based encoder-decoder networks for spelling and grammatical error correction," *CoRR*, vol. abs/1810.00660, 2018.

[9] Minh-Thang Luong, Hieu Pham, and Christopher D. Manning, "Effective approaches to attention-based neural machine translation," *CoRR*, vol. abs/1508.04025, 2015.

[10] Anand Mishra, Karteek Alahari, and CV Jawahar, "Scene text recognition using higher order language priors," in *BMVC-British Machine Vision Conference*. BMVA, 2012.

[11] S. M. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong, and R. Young, "Icdar 2003 robust reading competitions," in *Proceedings of the Seventh International Conference on Document Analysis and Recognition - Volume 2*, USA, 2003, ICDAR '03, p. 682, IEEE Computer Society.

[12] Dimosthenis Karatzas, Faisal Shafait, Seiichi Uchida, Masakazu Iwamura, Lluis Gomez i Bigorda, Sergi Robles Mestre, Joan Mas, David Fernandez Mota, Jon Almazan Almazan, and Lluis Pere De Las Heras, "ICDAR 2013 robust reading competition," in *2013 12th International Conference on Document Analysis and Recognition*. IEEE, 2013, pp. 1484–1493.

[13] Xinyu Zhou, Shuchang Zhou, Cong Yao, Zhimin Cao, and Qi Yin, "ICDAR 2015 text reading in the wild competition," *CoRR*, vol. abs/1506.03184, 2015.

[14] Kai Wang, Boris Babenko, and Serge Belongie, "End-to-end scene text recognition," in *2011 International Conference on Computer Vision*. IEEE, 2011, pp. 1457–1464.

[15] Trung Quy Phan, Palaiahnakote Shivakumara, Shangxuan Tian, and Chew Lim Tan, "Recognizing text with perspective distortion in natural scenes," in *ICCV*. 2013, pp. 569–576, IEEE Computer Society.

[16] Anhar Risnumawan, Palaiahnakote Shivakumara, Chee Seng Chan, and Chew Lim Tan, "A robust arbitrary text detection system for natural scene images," *Expert Syst. Appl.*, vol. 41, pp. 8027–8048, 2014.

[17] Andreas Veit, Tomas Matera, Lukas Neumann, Jiri Matas, and Serge J. Belongie, "Coco-text: Dataset and benchmark for text detection and recognition in natural images," *CoRR*, vol. abs/1601.07140, 2016.

[18] Donghyeon Won, Zachary C Steinert-Threlkeld, and Jungseock Joo, "Protest activity detection and perceived violence estimation from social media images," in *Proceedings of the 25th ACM international conference on Multimedia*. ACM, 2017, pp. 786–794.

[19] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd International Conference on Machine Learning*, New York, NY, USA, 2006, ICML '06, pp. 369–376, ACM.

[20] B. Shi, M. Yang, X. Wang, P. Lyu, C. Yao, and X. Bai, "Aster: An attentional scene text recognizer with flexible rectification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 9, pp. 2035–2048, Sep. 2019.

[21] Deli Yu, Xuan Li, Chengquan Zhang, Tao Liu, Junyu Han, Jingtuo Liu, and Errui Ding, "Towards accurate scene text recognition with semantic reasoning networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[22] Zhi Qiao, Yu Zhou, Dongbao Yang, Yucan Zhou, and Weiping Wang, "Seed: Semantics enhanced encoder-decoder framework for scene text recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[23] Karl Pearson F.R.S., "X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 50, no. 302, pp. 157–175, 1900.

[24] Ian Campbell, "Chi-squared and fisher-irwin tests of two-by-two tables with small sample recommendations.," *Statistics in medicine*, vol. 26 19, pp. 3661–75, 2007.