OCR-VQA: Visual Question Answering by Reading Text in Images

Anand Mishra[1] Shashank Shekhar[1] Ajeet Kumar Singh[2] Anirban Chakraborty[1]

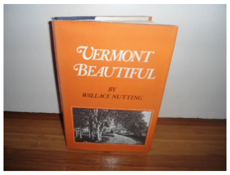
[1]Indian Institute of Science, Bangalore, India
[2]TCS Research, Pune, India

***Abstract*—The problem of answering questions about an image is popularly known as visual question answering (or VQA in short). It is a well-established problem in computer vision. However, none of the VQA methods currently utilize the text often present in the image. These "texts in images" provide additional useful cues and facilitate better understanding of the visual content. In this paper, we introduce a novel task of visual question answering by reading text in images, i.e., by optical character recognition or OCR. We refer to this problem as *OCR-VQA*. To facilitate a systematic way of studying this new problem, we introduce a large-scale dataset, namely *OCR-VQA*–200K. This dataset comprises of 207,572 images of book covers and contains more than 1 million question-answer pairs about these images. We judiciously combine well-established techniques from OCR and VQA domains to present a novel baseline for *OCR-VQA*–200K. The experimental results and rigorous analysis demonstrate various challenges present in this dataset leaving ample scope for the future research. We are optimistic that this new task along with compiled dataset will open-up many exciting research avenues both for the document image analysis and the VQA communities.**

***Keywords*-Optical Character Recognition (OCR), Visual Question Answering (VQA), Document image analysis, textVQA.**

I. Introduction

Given an input image and a related natural language question, the visual question answering (VQA) [4] task seeks to provide a natural language answer. In recent times, VQA has emerged as an important problem spanning computer vision, natural language understanding and artificial intelligence. However, VQA tasks and datasets are often limited to scene images. In this work, we introduce a novel task of visual question answering by reading text. Consider a scenario where a visually-impaired person picks up a book in the library (Figure 1), and asks the following questions to an intelligent conversational agent "*Who is the author of this book?*" or "*What type of book is this?*." Even, answering these apparently simple questions requires reading text in the image, interpreting the question and arriving at an accurate answer. Despite significant progress in VQA literature [4, 8] in the past few years, this important problem has not been studied. We fill this gap by introducing a novel dataset namely *OCR-VQA*–200K which contains 207,572 images of book covers and 1 million question-answer pairs about these images. This dataset can be explored and downloaded from our project website: https://ocr-vqa.github.io/.



Q. What is the title of this book?
A. Vermont Beautiful
Q. Who is the author of this book?
A. Wallace Nutting
Q. What type of book is this?
A. Travel

Figure 1: *We introduce a novel task of visual question answering by reading text in images, an accompanying large-scale dataset and baseline for this task.* ***[Best viewed in color].***

The optical character recognition (OCR) has a long history in computer vision and pattern recognition communities [17]. In the early years, OCR research has been restricted to handwritten digits [3] and clean printed document images [19]. Recently, OCR has manifested itself into various forms, e.g., photoOCR, popularly known as scene text recognition [5] and unconstrained handwritten text recognition [15]. There has been significant progress in all these forms of OCR problem. Nevertheless, many problems still remain open, e.g., recognizing text with arbitrary fonts and layout. In this paper, we further multiply the aforementioned challenges in OCR with that in the VQA, and introduce a novel task of answering visual questions by reading and interpreting text appearing in the images.

Further, we provide a novel deep model for VQA by reading texts in images. To this end, we rely on state-of-the-art text block identification and OCR modules, and present a trainable visual question answering system. Our baseline VQA system constitutes of the following representations: (i) pretrained CNN features for visual representation, (ii) text block coordinates and named entity tags on OCRed text for textual representation, and (iii) bi-directional LSTM for question representation. All these representations are fed to a trainable feed foreword neural network to arrive at an accurate answer.

Contributions of this paper

1) We draw attention to a novel and important problem of visual question answering by reading text in images. We refer to this new problem as *OCR-VQA*.
2) We introduce *OCR-VQA*–200K, the first large-scale dataset for the proposed VQA task by reading text in the image. This dataset can be downloaded from

2379-2140/19/$31.00 ©2019 IEEE
DOI 10.1109/ICDAR.2019.00156

our project website: https://ocr-vqa.github.io/.

3) We judiciously combine well-established techniques in OCR and VQA literature, and demonstrate baseline performance on *OCR-VQA*–200K. We are optimistic that *OCR-VQA*–200K will open-up various new research avenues for the document image analysis as well as the VQA communities.

II. Related Work

A. Traditional OCRs to recent advancements

Optical character recognition (OCR) is a historically well-studied problem in the literature [17]. Many research challenges such as binarization [11], skew correction [23], digit recognition [23], document layout analysis [19] and word spotting [7] have emerged as subproblems under the bigger umbrella of OCR literature. Despite several success cases, the problem of OCR is far from being solved in an unconstrained setting. In other words, many of the classical methods fall short when tested on the images captured in the wild. Besides continuous research efforts in OCR domain, the focus, of late, has shifted towards camera-captured images, scene text and handwritten text images in the wild. Especially, for the task of scene text recognition, we have witnessed significant boost in performance over the past years [5, 9, 12, 14, 25, 20]. These approaches utilize the availability of large annotated datasets and advancements in deep learning, and are currently the best performing methods for various OCR tasks. Despite these successes, OCR tasks are often limited to detection and recognition. However, down-stream important tasks like visual question answering has not been studied in the OCR literature.

B. Dataset efforts

The area of OCR has greatly benefited by the availability of public datasets such as MNIST [3], GW [6], IIIT-5K [16], SVT [24], etc. However, there does not exist a large-scale dataset for studying visual question answering by reading text, i.e., *OCR-VQA* task.[1]

C. Visual question answering literature

VQA has gained huge interest in recent years [4, 8]. Improvement in image classification and introduction of large-scale visual question answering benchmarks, such as VQA [4] and VQA v2.0 [8] have played important role in triggering a push in VQA research. However, most of these conventional VQA datasets and methods are mainly focused towards scene and object recognition, and they tend to ignore text in images which can be useful for VQA tasks, especially depending on the types of questions asked. We are optimistic that our new dataset will fill this gap in the traditional VQA literature.

Summary of *OCR-VQA*–200K statistics:	
Number of Images	207,572
Number of QA pairs	1,002,146
Number of unique authors	117,378
Number of unique titles	203,471
Number of unique answers	320,794
Number of unique genres	32
Average question length (in words)	6.46
Average answer length (in words)	3.31
Average number of questions per image	4.83

Table I: Composition of *OCR-VQA*–200K dataset in brief.

III. Dataset

The available datasets for visual question answering are not meant for the task of VQA by reading text as they mainly focus on non-textual objects. We, therefore, introduce a large-scale dataset specifically for the task of OCR-driven VQA and to facilitate future research efforts on this novel problem across document image analysis and VQA communities. The contents as well as the data-collection procedure for the proposed dataset, named as *OCR-VQA*–200K, are explained in detail in the subsequent sections. The summary of our dataset and a few selected sample images and QA pairs are given in Table I and Figure 2 respectively.

A. Data collection and annotation

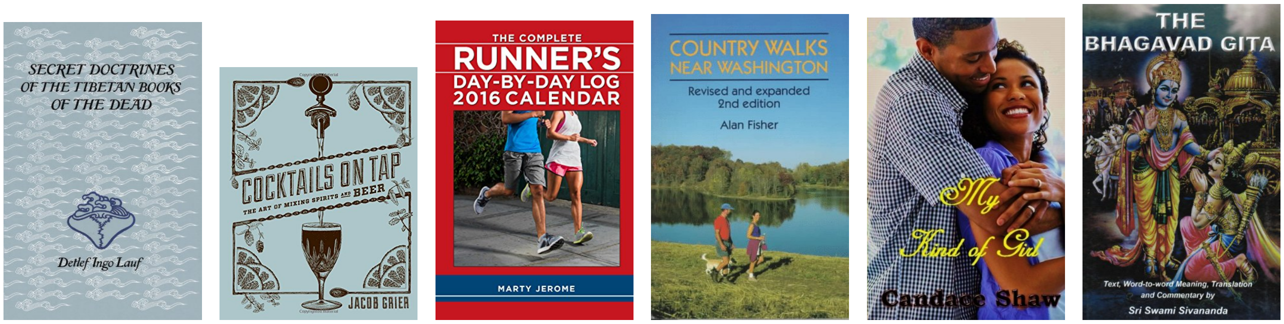
OCR-VQA–200K is constructed in the following stages.

1) Stage 1–obtaining images: In [13], Iwana et al. provided a dataset containing cover images of books including meta-data containing author names, titles and genres. This dataset was proposed towards classifying book covers into one of the 32 genres, e.g., science, religion, art, children book, comics, history, etc. using visual features. We adopt these images for visual question answering task by obtaining question and ground-truth answer pairs as discussed in Stage 2 and Stage 3 below.

2) Stage 2 – obtaining questions and ground-truth answers: We begin with a set of template questions inquiring about title, author name, genre (type) of book, year and edition. It should be noted that while title and author name are mostly available in all book covers, year and edition are only available in a few. Further, we expect that genre can be inferred from visual (e.g., graphics) and textual cues (e.g., title and surrounding text). The example template questions include: *What is the title of this book? Who is the author of this book?, What type of book is this?, Is this book related to religion?*, etc. The ground truth answer is obtained using meta-data made available with every book title.

3) Stage 3 – paraphrasing questions: As in any VQA task, the main challenge comes from the large variability and complexity in the set of possible natural language questions that can be framed with the same target answers. In order to add this natural complexity into our dataset, we ask human annotators to paraphrase the template-based

[1] Note that a couple of independent works in this area, viz. ICDAR 2019 Robust Reading Challenge on Scene Text Visual Question Answering [2] and TextVQA [22], have appeared very recently and were not available at the time of original submission of this manuscript.



(a)
Q. Who is the author of this book?
A. Detlef Ingo Lauf
Q. What type of book is this?
A. Religion and Spirituality

(b)
Q. Who wrote this book?
A. Jacob Grier
Q. Is this book related to cooking?
A. Yes

(c)
Q. Which year's calendar is this?
A. 2016
Q. Is this an exam preparation book?
A. No

(d)
Q. Who wrote this book?
A. Alan Fisher
Q. What is the edition of this book?
A. 2

(e)
Q. What is the title of this book?
A. My Kind of Girl
Q. Is this book related to travel?
A. No

(f)
Q. Who wrote this book?
A. Swami Sivananda
Q. Is this a religious book?
A. Yes

Figure 2: A selection of images and question-ground truth answer pairs from *OCR-VQA*–200K dataset. The dataset contains a large variations in type of questions and book covers.

question obtained in Stage 2. Some examples of paraphrased questions are: (i) *Template:* Who is the author of this book?, *Paraphrased:* Who wrote this book? (ii) *Template:* Is this book related to Bank and Money?, *Paraphrased:* Is this a finance book? (iii) *Template:* Which year's calendar is this? *Paraphrased:* What is the year printed on this calendar? We randomly choose between paraphrased and template questions for every image.

4) Stage 4 – preparing data split: We split images of the dataset into train, validation and test. To this end, we randomly divide 80%, 10% and 10% of images respectively for train, test, and validation, respectively. This leads to approximately 800K, 100K and 100K QA pairs in train, validation and test splits.

B. Challenges in OCR-VQA*–200K*

Our dataset presents several practical challenges and research avenues for the community as listed below:

- **Challenges for document image analysis community**: Our dataset presents two challenges for document image analysis community - (i) robust layout analysis scheme for handling wide variety of layouts present in book covers, and (ii) good optical character recognition engine which can read fancy fonts with different scales and orientations.
- **Challenges for VQA community**: Large answer-space, need for reading and interpreting text in image, identifying book category based on book cover, paraphrased questions, unseen answers during test time[2], are few of the practical challenges that do not appear in the conventional VQA datasets.

IV. Approach

In this section, we present our baseline approach to evaluate visual question answering performance on *OCR-VQA*–200K. Leveraging well-established techniques in VQA and OCR domains, our approach is composed of the following four modules.

A. Text block extraction

Our first step is to perform text block extraction in the images so that subsequently each text block and corresponding OCRed text can be represented as features for VQA task. To this end, we use the following three approaches, and choose the best performing approach for the subsequent stage: (i) Tesseract[3] text block segmentation (with Otsu binarization), (ii) EAST [26] and (iii) VGG text detector [9]. It should be noted that among the above three approaches (i) is conventional OCR-style method, whereas, (ii) and (iii) are modern scene text based methods. Further, (ii) and (iii) only provide word bounding boxes. In order to obtain text blocks from detected words, we first make detected word regions as black and no detection region as white, and then run a projection profile-based text block extraction, namely X-Y cut [18]. We choose maximum five text blocks in this

[2] In contrast to conventional VQA, few answers in the test set are not observed at training time, e.g., a particular book title or author name.

[3] https://github.com/tesseract-ocr/tesseract

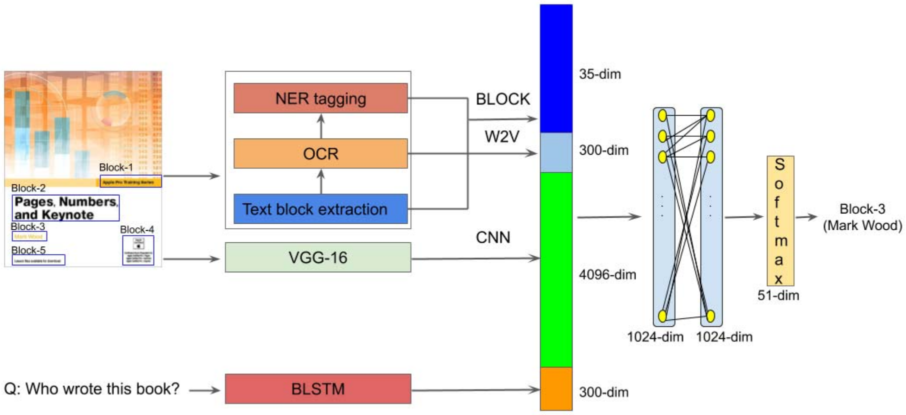


Figure 3: We designed a baseline approach for the novel task of answering visual questions by reading text in images. Please refer to Section IV for more details.

module. If there are more than five detected blocks, then we only take largest (area-wise) five blocks and represent each text block by a 5-dimensional vector containing block index, top-left and bottom-right $x-y$ coordinates. Note that blocks are indexed using numbers 1 to 5 based on their top-left y-coordinates. If there are less than five detected blocks, then each of the remaining blocks is represented by a 5-dimensional zero vector.

B. Optical Character Recognition (OCR)

Once text blocks are identified, we fed them to one of the following OCR engines: (i) Tesseract (ii) CRNN [20] and (iii) VGG deep text spotter [9]. In order to measure the effectiveness of these OCR engines, we compute recall of recognizing book titles and author names in our validation set. Note that recall is computed as fraction of correctly recognized words in author name and book title. This result is summarized in Table II. We show results without any correction, as well as with minimum edit-distance based correction (with ED) using a lexicon containing all the book titles and author names in our set. We observe that textspotter [9] achieves significantly better performance in recognizing book titles and author names. At the end of this module, each block index and corresponding OCR of the best performing method is stored as a look-up table.

C. NER tagging

Named entity recognition (NER) is a well-studied problem in natural language processing. Here, the goal is to locate and classify named entity mentions in unstructured text into predefined categories such as the person names, organizations, geo-political entities, year, etc. We use spacy[4] on OCRed text of every text block to identify person and year. Further, if OCRed text of a block contains word ‘edition’, we convert numeric information in that block to appropriate edition number, e.g., first to 1, 2nd to 2, etc. We then append one of the following numbers to block features computed using Section IV-A: 0 (if neither person name nor year is found), 1 (if person name is found), recognized year (if year is found). Further, in above feature, we append 0 (if edition is not found) or edition number (if edition is found). In other words, NER tags are 2-dimensional representation for each block.

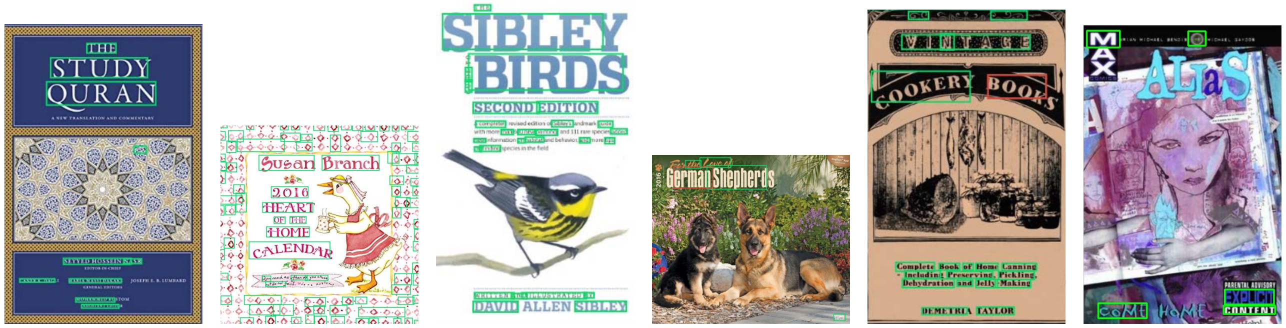
D. *Trainable* OCR-VQA *model*

We now describe our trainable *OCR-VQA* model. We represent questions using bidirectional long short term memory (BLSTM) [10]. Further, images are represented using VGG-16 [21] pretrained CNN features, and text blocks using their indices, coordinate positions and NER tags. We also use average of word2vec representations for all the words in the OCRed text as a feature. All these features, i.e., question (300-dim), image (4096-dim), average word2vec (300-dim) and block features (35-dim) (BLOCK) are concatenated to form a 4731-dimensional composite vector. This vector is then fed to a fully connected feed forward network (2 layers of size 1024) followed by a softmax layer. The softmax layers predicts, one of the *five block indices*, *32 book genres*, *yes or no*, *book edition* (1 to 5), or a *year between 2000 to 2016* as an answer. In other words, output dimension of our model is 5 (block index) + 32 (genre) + 2 (yer or no) + 5 (edition number) + 7 (year) = 51. The fully connected feed forward network is trained by back propagating categorical cross entropy loss using RMSprop optimizer. We used learning rate as 0.01, batch size of 128 and maximum epoch of training as 30. Please refer to Figure 3 for schematic representation of our proposed *OCR-VQA* framework.

V. Results and Analysis

In this section, we provide results of our approach and its variants on *OCR-VQA*–200K. We use four sets of features:

[4] https://spacy.io/



(a) Q. Is this a religious book? A. Yes

(b) Q. What is the year in the Calendar? A. 2016

(c) Q. What is the edition of this book? A. 2

(d) Q. What is the year in the Calendar? A. 'NULL'

(e) Q. What is the title of this book? A. Looker boon

(f) Q. What is the title of this book? A. Explicit Content

Figure 4: Some sample results of our best performing method (BLOCK+CNN+W2V) on *OCR-VQA*–200K. (a), (b) and (c) are some of the successful examples, whereas (d), (e) and (f) shows a few failure cases. Current best text recognition engine and VQA techniques fall short to deal with challenges present in our large-scale dataset. **[Best viewed in color].**

Method	Recall (book title)		Recall (author name)	
	Without	With ED	Without	With ED
Tesseract	0.28	0.43	0.25	0.40
CRNN	0.30	0.50	0.35	0.46
Textspotter	**0.53**	**0.83**	**0.52**	**0.77**

Table II: OCR performance on reading author names and titles correctly. We observed that the TextSpotter [1] performs best on our dataset. We use the the same for our trainable VQA model.

Method	accuracy
BLOCK	42.0
CNN	14.3
BLOCK+CNN	41.5
BLOCK+CNN+W2V	**48.3**

Table III: Ablation Study:*OCR-VQA* results by our proposed baseline and its variants.

(i) only text-block features (BLOCK), (ii) only VGG-16 features (CNN), (iii) block and CNN features, (iv) block, CNN and word2vec (BLOCK+CNN+W2V) features. The results of these ablations are reported in Table III. We observe that BLOCK and CNN features alone are not sufficient in answering questions. This is primarily because BLOCK features are designed to answer only author name, book title or year related questions, and similarly CNN features only deal genre related question. The best performing variant of our method is BLOCK+CNN+W2V. Here, text embeddings of all the words on the cover, i.e. the average word2vec specially help in improving genre related questions.

A. Error analysis and challenges

The lower performance in this dataset is primarily due to wide variations in scale, layout and font-styles of text. Secondly, variations in questions asked (e.g., paraphrasing) and questions related to genre of book also limit the performance. We have shown some examples of successful and failure cases in Figure 4. In Figure 4(a), (b) and (c), we observe that despite large variations in layout, the proposed baseline is able to answer question inquiring book type, year and edition of book. Some failure cases are also shown in Figure 4(d), (e) and (f). The major failures are due to fancy fonts and cluttered background on the book cover, where text detection and recognition perform poorly.

We also show results for various question types in Table IV. We observe that binary questions yield the maximum success rate, but the method is observed to be less successful on other factual questions, e.g., inquiring about book genre, author name and year. To verify the effectiveness of block features, we also experimented with OCR+NER baseline without any block features. This baseline yields 39.5% as compared to our method which achieves 42.9% for authorship questions. This result indicates the utility of our block features which encode layout and spatial positioning information along with the NER-tags.

While addressing all the challenges is beyond the scope of a single paper, we believe our dataset will give a strong test bed for future research in these areas.

VI. Summary and Future Work

In this paper, we introduced the novel task of visual question answering by reading text in images, and an ac-

Question type	Accuracy (in %)
Binary	58.2
book title	48.5
author name	42.9
book genre	22.0
Year	46.2
Edition	42.5

Table IV: *OCR-VQA* results on different question types.

companying large-scale dataset (*OCR-VQA*–200K). To the best of our knowledge, this is the first large-scale dataset identifying the need for reading text in the images to answer visual questions. Towards benchmarking this novel dataset we composed a baseline deep model that builds atop state-of-the-art VQA and OCR modules. Our results and error analyses suggest that the proposed task and the dataset present several challenges and research avenues for both document image analysis as well as the VQA community. We are optimistic that the introduction of this task and dataset will encourage research community to improve text recognition in the wild as well as the downstream practical task, i.e., visual question answering by reading text. We plan to extend our dataset to scene text and born-digital images, and focus on combining image and text understanding for more challenging VQA tasks in the future.

Acknowledgements

This work is partially supported by the Pratiksha Trust Young Investigator Award to Dr. Anirban Chakraborty.

References

[1] http://textspotter.org.

[2] ICDAR 2019 Robust Reading Challenge on Scene Text Visual Question Answering. https://rrc.cvc.uab.es/?ch=11. Accessed: 2019-06-01.

[3] MNIST database. http://yann.lecun.com/exdb/mnist/. Accessed: 2019-02-28.

[4] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. VQA: Visual Question Answering. In *ICCV*, 2015.

[5] A. Bissacco, M. Cummins, Y. Netzer, and H. Neven. PhotoOCR: Reading Text in Uncontrolled Conditions. In *ICCV*, 2013.

[6] A. Fischer, A. Keller, V. Frinken, and H. Bunke. Lexicon-free handwritten word spotting using character hmms. *Pattern Recognition Letters*, 33(7):934–942, 2012.

[7] V. Frinken, A. Fischer, R. Manmatha, and H. Bunke. A novel word spotting method based on recurrent neural networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(2):211–224, 2012.

[8] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *CVPR*, 2017.

[9] A. Gupta, A. Vedaldi, and A. Zisserman. Synthetic data for text localisation in natural images. In *CVPR*, 2016.

[10] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 1997.

[11] N. R. Howe. Document binarization with automatic parameter tuning. *IJDAR*, 16(3):247–258, 2013.

[12] W. Huang, Y. Qiao, and X. Tang. Robust scene text detection with convolution neural network induced mser trees. In *ECCV*, 2014.

[13] B. K. Iwana, S. T. Raza Rizvi, S. Ahmed, A. Dengel, and S. Uchida. Judging a book by its cover. *arXiv preprint arXiv:1610.09204*, 2016.

[14] M. Jaderberg, A. Vedaldi, and A. Zisserman. Deep Features for Text Spotting. In *ECCV*, 2014.

[15] M. Liu, Z. Xie, Y. Huage, L. Jin, and W. Zhou. Distilling gru with data augmentation for unconstrained handwritten text recognition. In *ICFHR*, 2018.

[16] A. Mishra, K. Alahari, and C. V. Jawahar. Scene Text Recognition using Higher Order Language Priors. In *BMVC*, 2012.

[17] G. Nagy. Twenty years of document image analysis in PAMI. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(1):38–62, 2000.

[18] G. Nagy and S. C. Seth. Hierarchical representation of optically scanned documents. In *ICPR*, 1984.

[19] F. Shafait, D. Keysers, and T. M. Breuel. Performance evaluation and benchmarking of six-page segmentation algorithms. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(6):941–954, 2008.

[20] B. Shi, X. Bai, and C. Yao. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(11):2298–2304, 2017.

[21] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[22] A. Singh, V. Natarajan, M. Shah, Y. Jiang, X. Chen, D. Batra, D. Parikh, and M. Rohrbach. Towards VQA models that can read. In *CVPR*, 2019.

[23] R. Smith. A simple and efficient skew detection algorithm via text row accumulation. In *ICDAR*, 1995.

[24] K. Wang, B. Babenko, and S. Belongie. End-to-End Scene Text Recognition. In *ICCV*, 2011.

[25] T. Wang, D. Wu, A. Coates, and A. Ng. End-to-End Text Recognition with Convolutional Neural Networks. In *ICPR*, 2012.

[26] X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou, W. He, and J. Liang. EAST: an efficient and accurate scene text detector. In *CVPR*, 2017.