ICDAR 2021

International Conference on Document Analysis and Recognition

September 5-10, 2021, Lausanne, Switzerland

EVALUATING THE INFLUENCE OF OCCLUSION ON THE QUALITY OF DEEP LEARNING-BASED SYSTEMS FOR NATURAL SCENE TEXT DETECTION AND RECOGNITION

Workshop on Camera-Based Document Analysis and Recognition – CBDAR 2021

Authors: Aline G. Soares, Byron Leite and Estanislau Lima.

SCHEDULE

01. INTRODUCTION

Contextualization, motivation, and problem statement with objectives.

02. LITERATURE REVIEW

State-of-the-art overview.

03. ISTD-OC DATASET

Description of the proposed occlusion generation method and the ISTD-OC dataset.

04. EXPERIMENTATION

Planning and execution of deep models' evaluation under the ISTD-OC dataset.

05. RESULTS

Presentation and discussion of the obtained results.

06. CONCLUSIONS

Final considerations, contributions, limitations, and future works.

INTRODUCTION

01

TEXT HAS PLAYED AN ESSENTIAL ROLE IN HUMAN LIFE.

Key tool for communication and understanding of the environment.

02

WIDE RANGE OF COMPUTER VISION-BASED APPLICATIONS.

Rich and precise semantic information embodied.

03

STILL CHALLENGING

Significant variation of text diversity, scene complexity, and distortion factors.

04

DEEP LEARNING

Demand for data to achieve satisfactory results in natural scenes.

PROBLEM STATEMENT

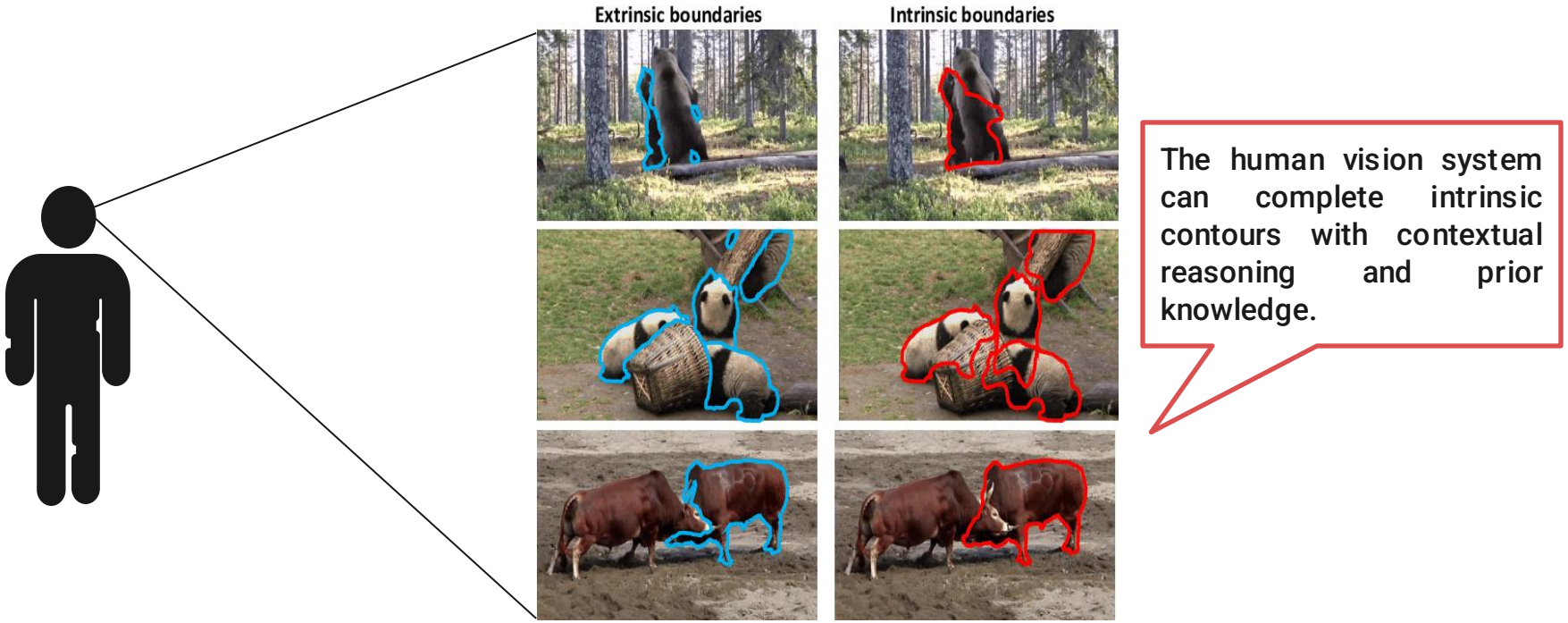
- Identify texts in natural scenes provides environment understanding and benefits the human-machine interaction.
- However, in real-world scenes, texts and objects rarely occur in isolation. Occlusion is one of the recurring problems and represents a severe threat to the system's performance.
- Even with Deep Neural Networks, little has yet been done in natural settings due to the demand for data to achieve satisfactory results.

STUDY OBJECTIVES

- **General Objective:** present a systematic methodology to implement occlusion and generate large datasets of occluded scene text in natural images.
- To achieve the general objective, we define the following **specific objectives**:

  - ✓ investigate the state-of-the-art in text detection and recognition;
  - ✓ evaluate the current effectiveness of baseline algorithms without any occlusion;
  - ✓ propose a systematic methodology to generate a large dataset of occluded scene text in natural images;
  - ✓ evaluate the current effectiveness of baseline algorithms under the generated dataset;
  - ✓ validation and comparison of the benchmark models' ability to handle different occlusion levels.

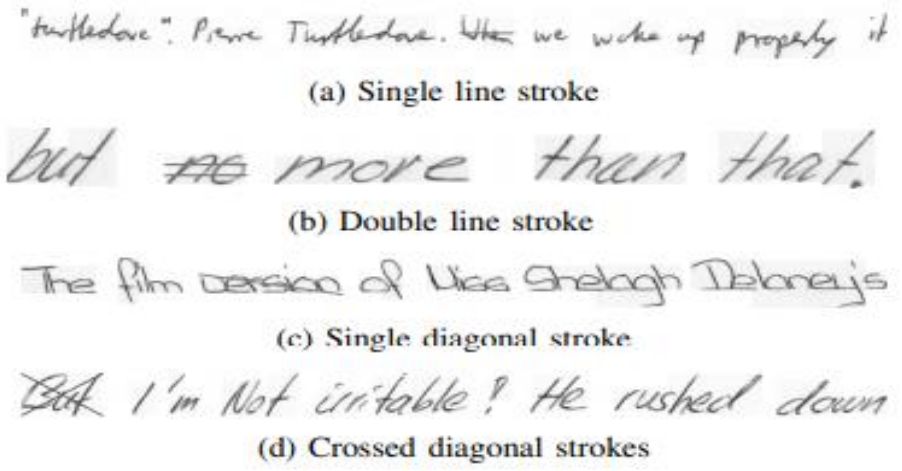
OCCLUSION



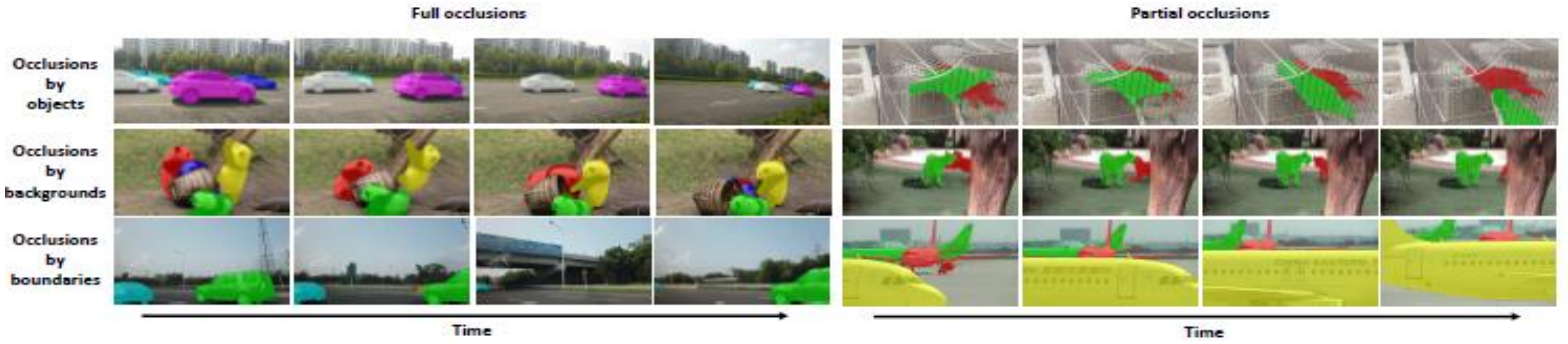
Source: Occluded Video Instance Segmentation (2021).

OCCLUSION

- In real-world scenes, texts and objects rarely occur in isolation.
- Occluded text instances embody a critical threat to a scene text recognition system's performance.
- In the past years, significant progress has been made in handle with occlusion in handwritten text recognition, video instance segmentation, object detection and recognition, and image-to-image translation.



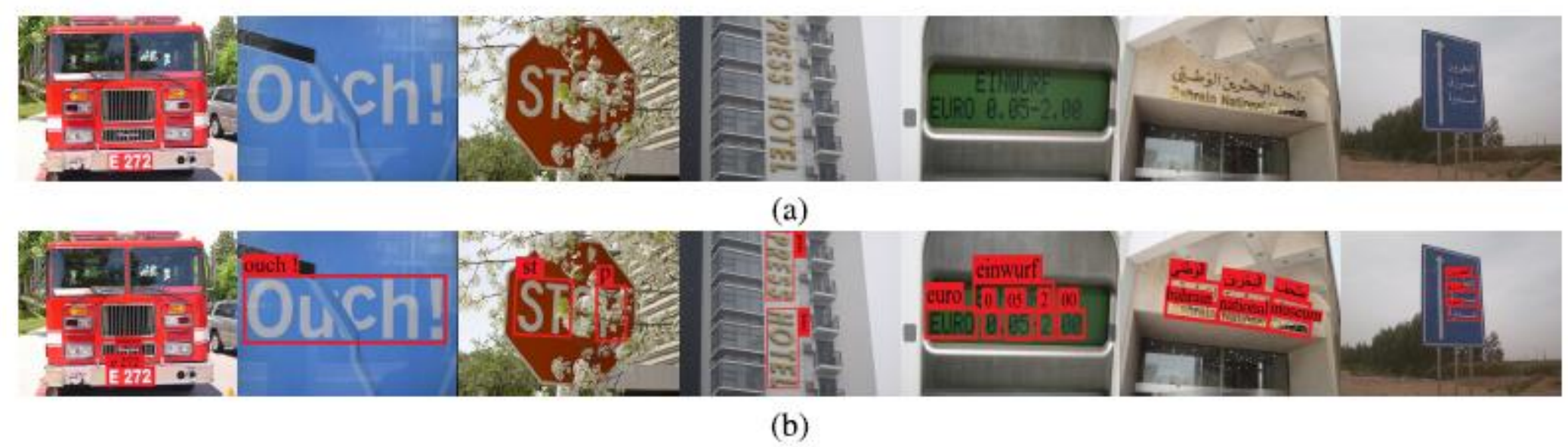
Source: A deep learning approach to handwritten text recognition in the presence of struck-out text (2019).



Source: Occluded Video Instance Segmentation (2021).

OCCLUSION

- **For scene** text detection and recognition, Bagi et al. (2020) proposed an end-to-end trainable light-weight scene text spotter for cluttered environment.



Source: Cluttered TextSpotter: An End-to-End Trainable Light-Weight Scene Text Spotter for Cluttered Environment (2020).

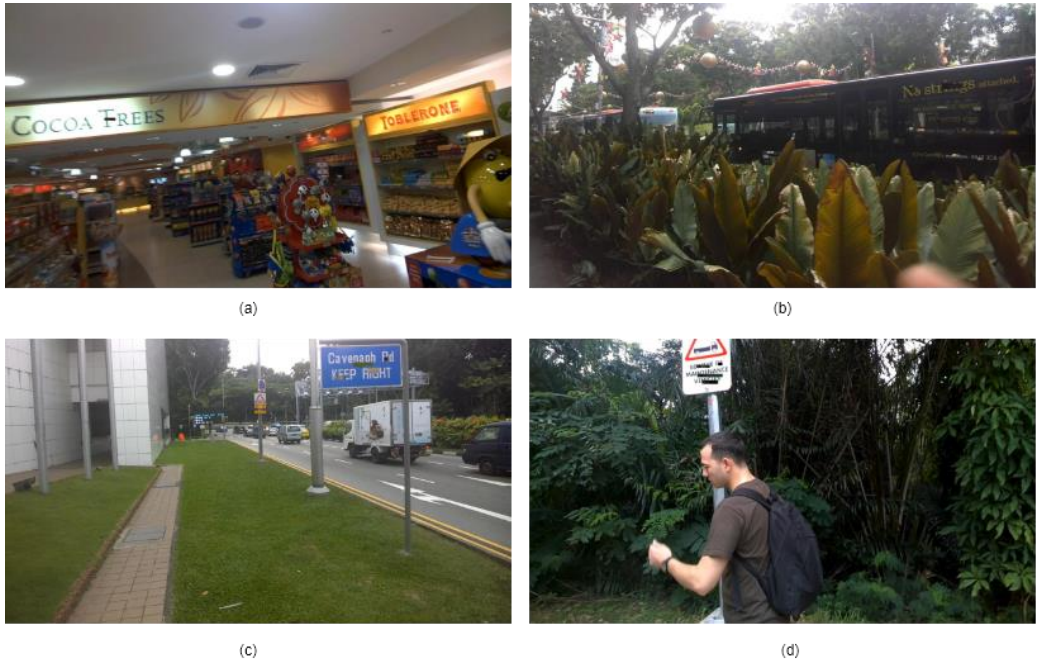
DATASETS OF TEXT IN NATURAL IMAGES

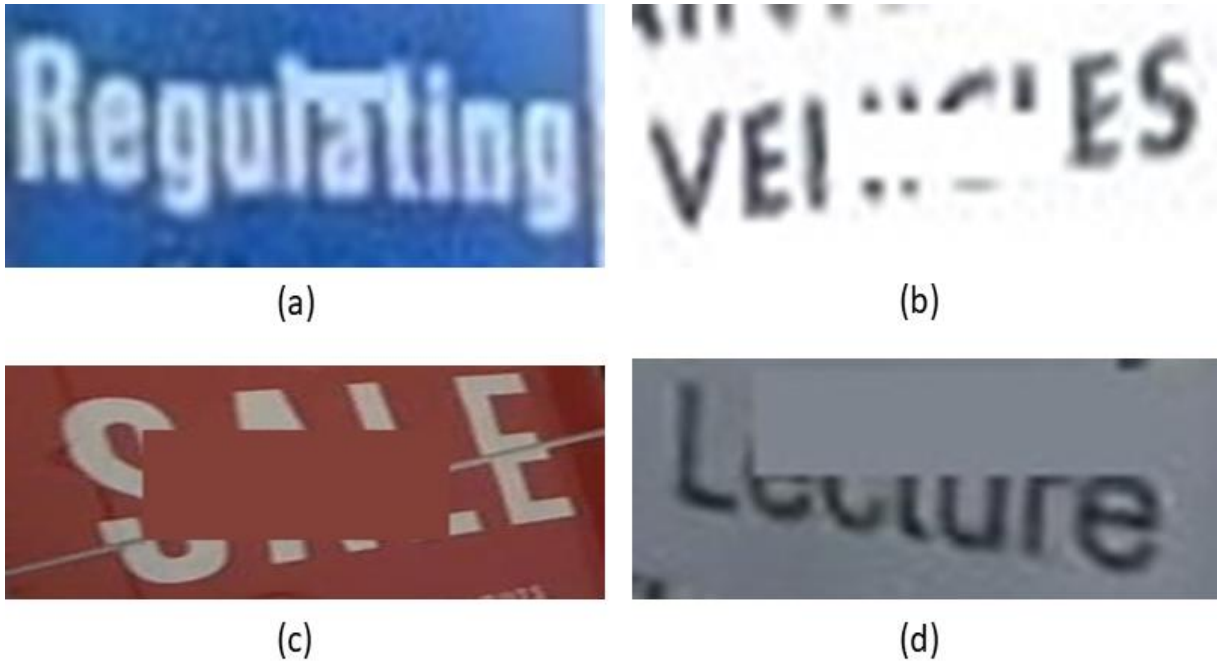
- The primary purpose is to achieve better performances in challenges such as arbitrary shape, adjacent instances, and languages but none of the methods proposed in state of the art focuses on images where a part of the text is missing due to **occlusion**.

- **Baek et al. (2019)** highlight how each of the works differs in constructing and using their datasets and investigate the bias caused by the inconsistency when comparing performance between different works.

ISTD-OC DATASET

- The Incidental Scene Text Dataset - Occlusion, also named ISTD-OC, is derivated from the irregular real-world dataset ICDAR 2015.





OCCLUSION GENERATION METHOD

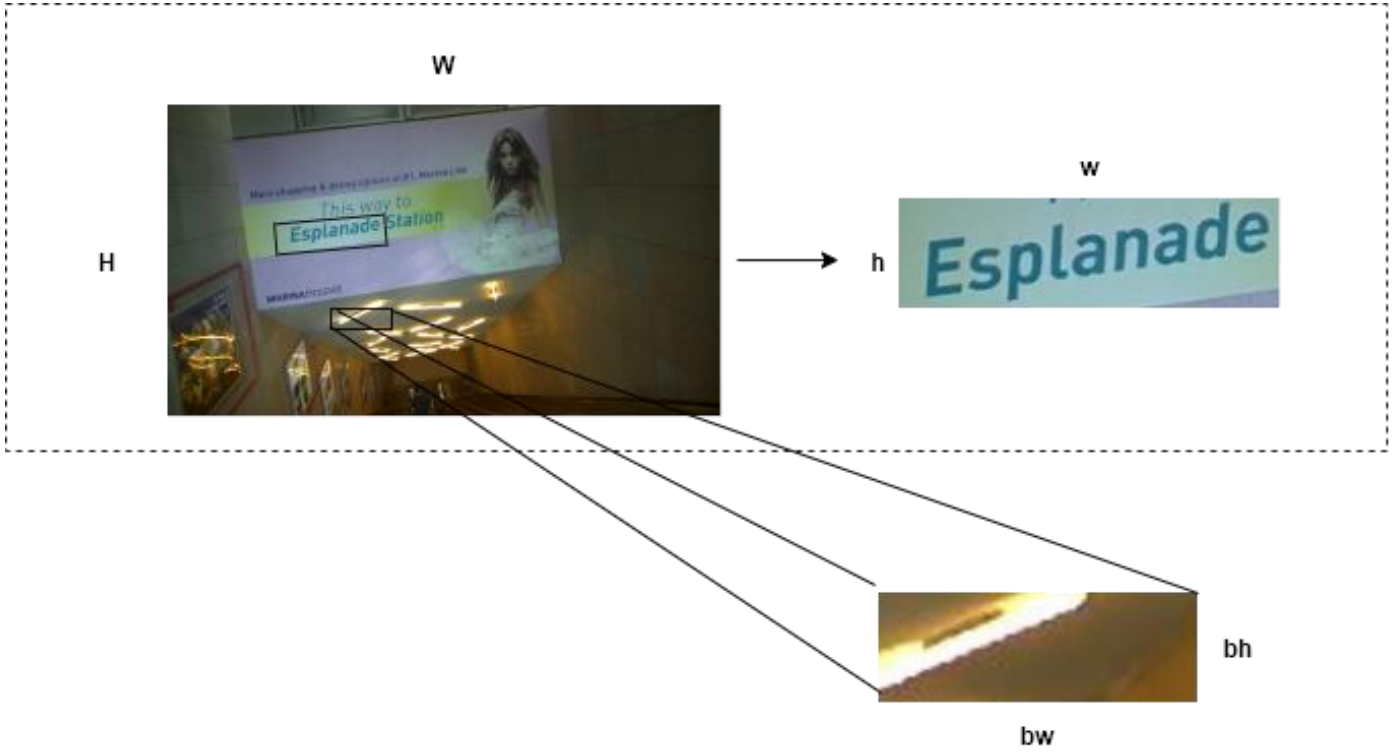
#1

For each image of the dataset, all instances of text are occluded differently.

OCCLUSION GENERATION METHOD

#2

The prediction was made considering the proportion of the text size in relation to the original image.



for each text box:
pbwW = w/W
pbhH = h/H
alpha = 0.55

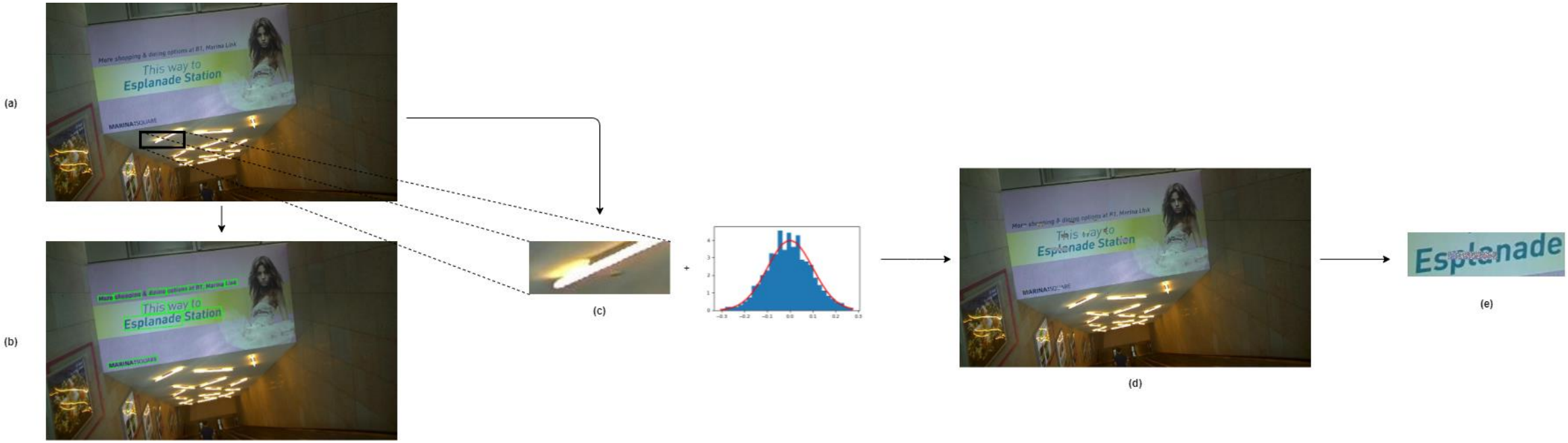
$bw = \alpha \times w \times rand + (1- \alpha) \times w \times pbwW$

$bh = \alpha \times w \times rand + (1- \alpha) \times w \times pbhH$

OCCLUSION GENERATION METHOD

The occlusion corresponds to a random part of the original image between 0 and 100%. In order to prevent the models from learning patterns of equal regions of occlusion, we added noise to the RGB channels of the occlusion.

#3

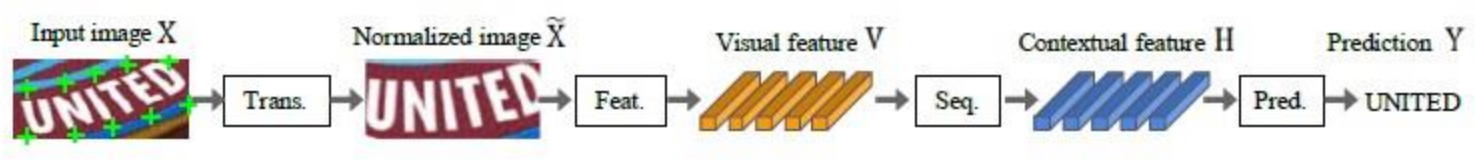


STD PLANNING

- CRAFT[1], EAST[2], PAN[3], and PSENet[4] were used for text detection evaluation.
- Except for EAST and PSENet we used the corresponding pre-trained model directly from the authors' GitHub page trained on the ICDAR15 dataset.
- For testing, the ISTD-OC has been used.

STR PLANNING

- For scene text recognition schemes, the deep learning-based techniques chosen are based on the Baek et al. (2019)[5] study.



Source: What Is Wrong With Scene Text Recognition Model Comparisons? Dataset and Model Analysis (2019).

- Models such as CRNN, RARE, STAR-Net and ROSETTA were re-implemented with the proposed framework and under consistent settings.

- All recognition models have been trained on combination of SynthText and MJSynth datasets. For evaluating, 2077 occluded cropped word instances images of each occlusion level from ISTD-OC were used.

EVALUATION METRICS FOR STD

- For scene text detection schemes, we chose to adopt the ICDAR protocol[6] and their standard evaluation metrics: Precision(P), Recall (R) and F1-Score metrics.
- Precision and Recall are based on using the ICDAR15 intersection over union (IoU) metric.

$$IoU = \frac{Area\,(G_j \cap D_i)}{Area\,(G_j \cup D_i)}$$

- F1-Score were used as follow:

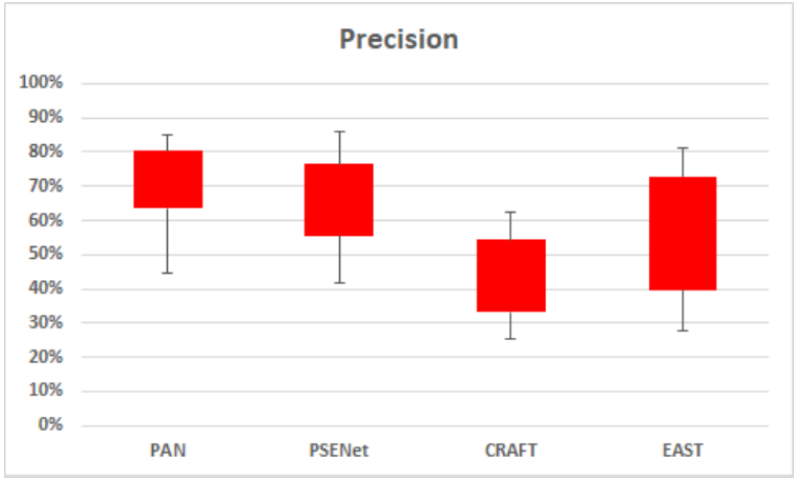
$$F1 - Score = 2\ x\ \frac{P\ x\ R}{P + R}$$

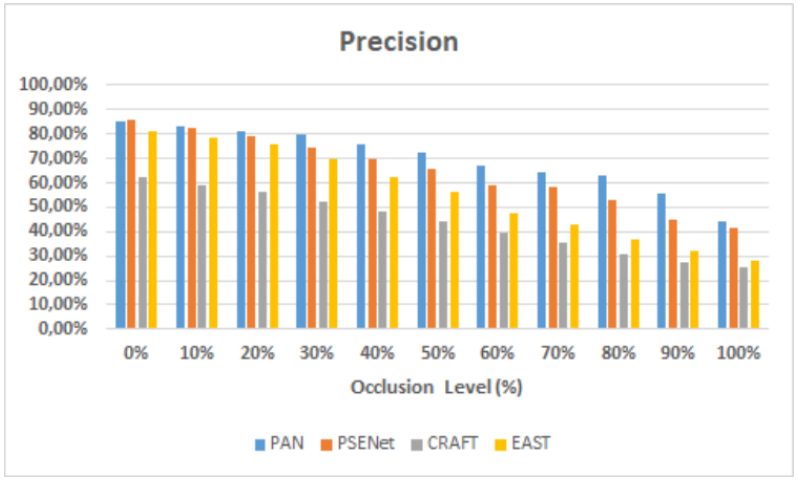
EVALUATION METRICS FOR STR

- For scene text recognition schemes, we chose to adopt the most common evaluation metrics in STR systems: the word error rate (WER) and the character error rate (CER).
- CER is defined as the minimum number of editing operations at the character level, considering the respective ground truth. WER is specified in the same way, but when it comes to words.
- C is the total number of characters, and C_r represents the number of correctly recognized characters. In the similar way, WER is defined.

$$CER = \frac{C_r}{C} \qquad WER = \frac{W_r}{W}$$

EVALUATION OF STD APPROACHES





PAN

70%

of precision for ranges of occlusion between 10% and 50%.

PSENet

~60%

of precision for ranges of occlusion between 70% and 80%.

CRAFT

37%

precision decline when compared with images without any generated occlusion.

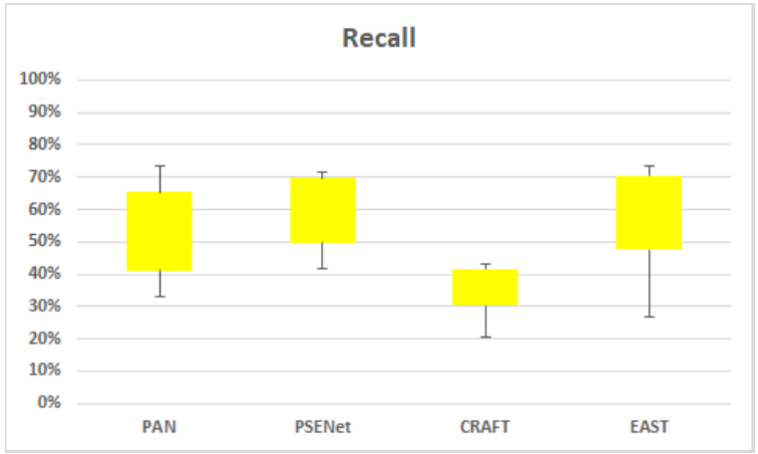
EAST

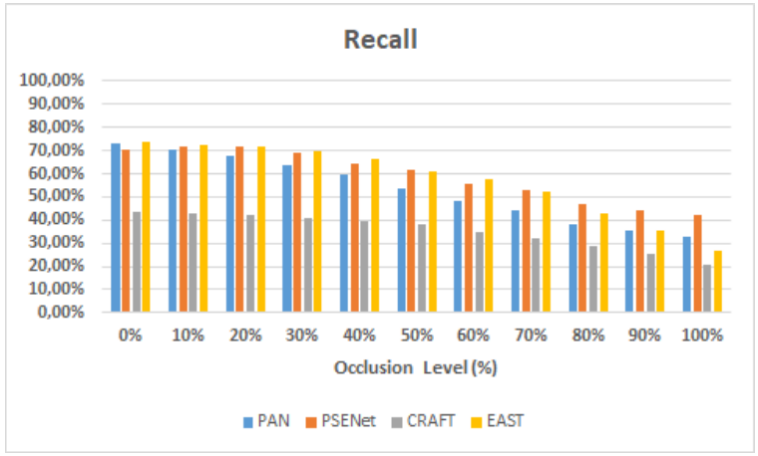
~22%

more precise for detecting occluded texts then CRAFT for levels of occlusion between 20% and 40%.

RESULTS

EVALUATION OF STD APPROACHES





PAN

73%

of recall decline when comparing images without any occlusions and images with heavy occlusions.

PSENet

52%

of recall for images with text instances ~70% occluded.

CRAFT

~30%

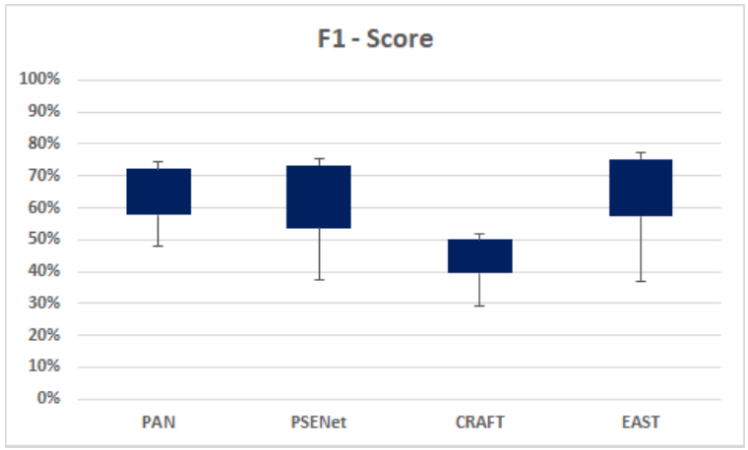
of recall for heavy occlusions.

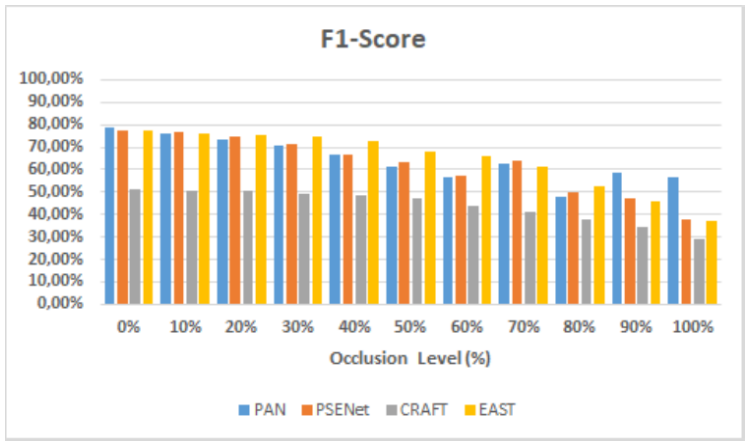
EAST

~70%

of recall for levels of occlusion between 10% and 30%.

EVALUATION OF STD APPROACHES





PAN

57%

of F1-Score for images with text instances ~100% occluded.

PSENet

40%

of F1-Score decline when comparing images without any occlusions and images with heavy occlusions.

CRAFT

~20%

of F1-Score decline for non occluded images to heavy occluded images.

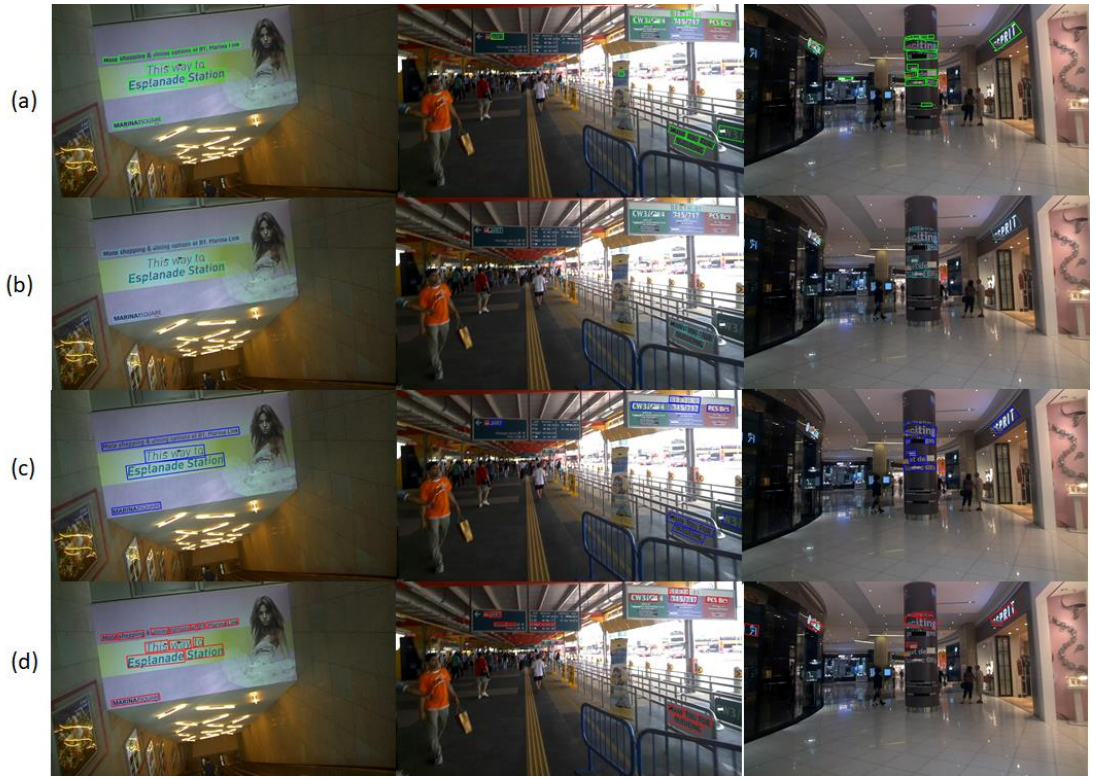
EAST

~10%

of F1-Score decline for partial occluded images to heavy occluded images.

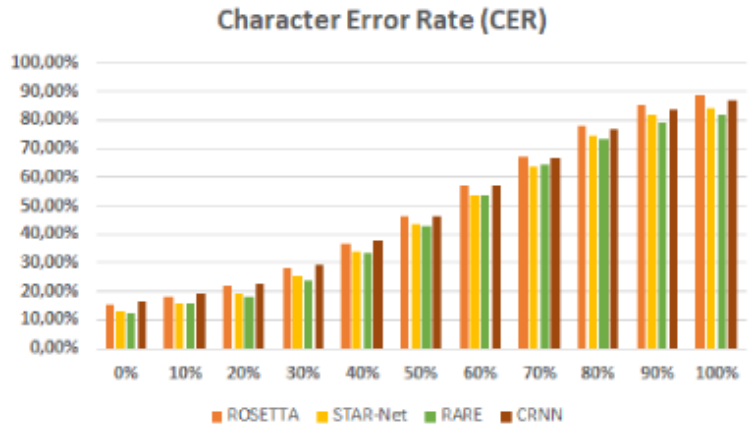
EVALUATION OF STD APPROACHES

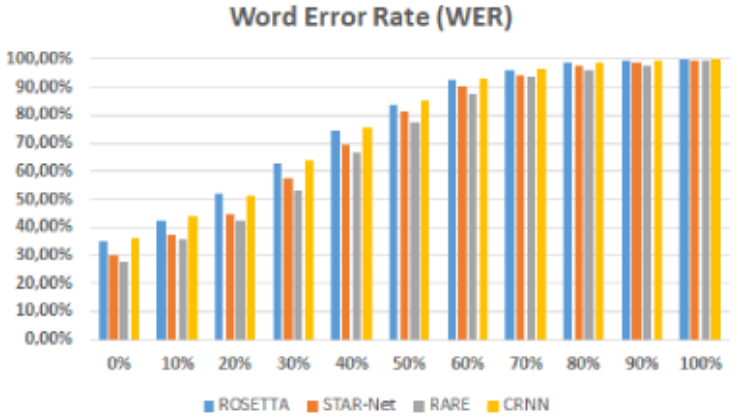
- Qualitative evaluation of text detection state-of-the-arte models on ISTD-OC. Each row presents a sample of results for PSENet (a), EAST (b), CRAFT (c) and PAN (d) in levels of 20%, 40% and 80% of occlusion.



RESULTS

EVALUATION OF STR APPROACHES





ROSETTA

+90%

WER for levels of occlusion over 60%.

STAR-Net

2nd

best STR model on both CER and WER evaluation.

RARE

- 20%

of CER for occlusion levels between 10% and 30%.

CRNN

+75%

of WER for occlusion levels over 40%.

EVALUATION OF STR APPROACHES

- Samples of cropped word instances from ISTD-OC dataset under 10%(a), 40% (b), 60% (c) and (d) 70% occlusion levels.

FINAL CONSIDERATIONS

- Automatically reading text in natural scenes has a tremendous practical value due to its potential applications in numerous fields.
- Recent datasets and competitions show that the community moves toward more challenging text recognition tasks. However, it is still hard to assess the problem of incidental and diversified text detection and recognition in natural scenes.
- In this work, we investigated a selected number of state-of-the-art deep architectures for scene text detection and recognition in case of occlusion.
- As a proposal, we proposed a systematic methodology to generate occlusion, which resulted in the ISTD-OC dataset.
- The experimental results suggest that these existing deep architectures for STD and STR are far from the human visual system's ability to read occluded texts in natural daily-life situations with little supervision learning.

FUTURE WORKS

- Several areas may be helpful for further investigation as future works to this research. The causes of these existing models' failure could be examined for further improvement of a single and lightweight network for end-to-end scene text detection and recognition to handle different levels of occlusion.

- Also, we should improve the way occlusion is generated, representing more real-world scenarios.

BIBLIOGRAPHICAL REFERENCES

- YUAN, T.-L.; ZHU, Z.; XU, K.; LI, C.-J.; HU, S.-M. Chinese text in the wild. arXiv preprint arXiv:1803.00085, 2018.
- ZHOU, X.; YAO, C.; WEN, H.; WANG, Y.; ZHOU, S.; HE, W.; LIANG, J. EAST: An Efficient and Accurate Scene Text Detector. 2015.
- BABU, S.C (2021). **Automating Receipt Digitization with OCR and Deep Learning.**
- LONG, S.; HE, X.; YAO, C. Scene text detection and recognition: The deep learning era. International Journal of Computer Vision, Springer, p. 1–24, 2020.
- RAISI, Z.; NAIEL, M. A.; FIEGUTH, P.; JUN, C. V. Text Detection and Recognition in the Wild : A Review. p. 13–15, 2020.
- QI, J.; GAO, Y.; LIU, X.; HU, Y.; WANG, X.; BAI, X.; TORR, P. H.; BELONGIE, S.;YUILLE, A.; BAI, S. Occluded video instance segmentation. arXiv preprint arXiv:2102.01558,2021.
- NISA, Hiqmat et al. A deep learning approach to handwritten text recognition in the presence of struck-out text. In: **2019 International Conference on Image and Vision Computing New Zealand (IVCNZ)**. IEEE, 2019. p. 1-6.
- BAGI, Randheer; DUTTA, Tanima; GUPTA, Hari Prabhat. Cluttered TextSpotter: An End-to-End Trainable Light-Weight Scene Text Spotter for Cluttered Environment. **IEEE Access**, v. 8, p. 111433-111447, 2020.
- CHEN, Xiaoxue et al. Text recognition in the wild: A survey. **arXiv preprint arXiv:2005.03492**, 2020.
- BAEK; HAN, Y.; KIM, J. O.; LEE, J.; PARK, S. What Is Wrong With Scene Text Recognition Model Comparisons? Dataset and Model Analysis. 2019.
- CAO, D.; ZHONG, Y.; WANG, L.; HE, Y.; DANG, J. Scene text detection in natural images: A review.Symmetry, Multidisciplinary Digital Publishing Institute, v. 12, n. 12, p. 1956,2020

THANKS!

Do you have any questions?

{ags4,ebl2}@ecomp.poli.br; byron.leite@upe.br

You can access our Github repository here:

SCAN ME