NISTIR 8199

The Text Recognition Algorithm Independent Evaluation (TRAIT)

Afzal Godil
Patrick Grother
Mei Ngan

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8199

National Institute of
Standards and Technology
U.S. Department of Commerce

NISTIR 8199

The Text Recognition Algorithm Independent Evaluation (TRAIT)

Afzal Godil
Patrick Grother
Mei Ngan
Information Access Division
Information Technology Laboratory

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8199

December 2017

U.S. Department of Commerce
Wilbur L. Ross, Jr., Secretary

National Institute of Standards and Technology
Walter Copan, NIST Director and Under Secretary of Commerce for Standards and Technology

Disclaimer

Specific hardware and software products identified in this report were used in order to perform the evaluations described in this document. In no case does identification of any commercial product, trade name, or vendor, imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the products and equipment identified are necessarily the best available for the purpose.

Key Words

Text detection; text recognition; evaluation; competition; benchmarking.

Table of Contents

List of Tables

List of Figures

Summary

The Text Recognition Algorithm Independent Evaluation (TRAIT) was conducted to assess the capability of text detection and recognition algorithms to correctly detect and recognize text appearing in unconstrained imagery. NIST invited all organizations, particularly universities and corporations, to submit their technologies to TRAIT-2016. The evaluation was a sequestered evaluation of text detection and recognition algorithms and open worldwide.

The primary driver of the evaluation was to support forensic investigations of digital media. These images are of interest to NIST's partner law enforcement agencies that seek to employ text recognition in investigating the serious crime of child exploitation. The primary applications are the identification of previously known victims and suspects, as well as detection of new victims and suspects. The presence of text may allow a location to be identified or to generate leads.

The primary dataset is an operational child exploitation collection containing images and videos seized in criminal investigations. Many of the images contain geometrically unconstrained text. This text is human-legible and sometimes has investigational value. Such text is visible on certificates, posters, logos, uniforms, sports apparel, computer screens, business cards, newspapers, books lying on tables, cigarette packets and a long list of rarer objects.

The TRAIT best results summary is presented in Table 1. The evaluation results show that the performance of unconstrained text recognition was low. However, from Phase 1 to Phase 3, the performance of the text recognition algorithms have shown significant improvement. From first test (in either Phase 1 or 2) to Phase 3, we saw an average improvement across the classes of 61%, with Megvii improving by 70% and the Czech algorithm by 52%. Based on these results and other text detection and recognition competitions/evaluations, we conclude that there is still much room for improvement in unconstrained text recognition. Although three groups participated in the TRAIT challenge, we hope this evaluation will spearhead more research in this exciting field.

Table 1. TRAIT Best Results Summary

Test	Metric		Value	Participant/Method
Class A - Text detection	F-measure		0.34	Megvii C30A
Class B - Text Recognition (given location)	Edit Distance based Accuracy	Character	39.9%	Megvii C32B
		Word (ordered)	21.4%	
		Word (unordered)	22.8%	
Class C - Text detection and recognition	Edit Distance based Accuracy	Character	44.8%	Megvii C32C
		Word (ordered)	22.8%	
		Word (unordered)	25.6%	Megvii C30C
Class D - URLs	Accuracy	Detection	30.5%	Megvii C32C
		Recognition	13.1%	

Abstract

The report describes and presents the results for Text Recognition Algorithm Independent Evaluation (TRAIT) in support of forensic investigations of digital media. These images are of interest to NIST's partner law enforcement agencies that seek to employ text recognition in investigating serious crimes. This evaluation used images seized in child exploitation investigations. The primary application is the identification of previously known victims and suspects, as well as detection of new victims and suspects. The presence of text, for example, on a wall poster or on an item of clothing, may allow a location to be identified and linked to prior cases. In total, 3 groups took part in this evaluation over three Phases. The evaluation results show that the initial performance of text recognition is low. However, from Phase 1 to Phase 3, the performance of text recognition algorithms has shown significant improvement. We hope this evaluation will stir more research in this field.

1. Introduction

The field of text detection and recognition in unconstrained imagery (images, videos, and media) is receiving renewed interest because of an increasing number of important applications. Text is present everywhere; it is often embedded in documents and imagery, or present incidentally in scene imagery, and sometimes left unintentionally in forensic media evidence. The papers described in [1, 2] have more details about the field, applications, algorithms, datasets, and evaluations.

Applications of text detection and recognition are numerous, including media indexing and retrieval, data mining, media forensics and security, scene understanding, autonomous navigation, law enforcement investigations, industrial automation (by reading text on packages), help with accessibility assistance (via providing text-to-speech and machine translation), mobile phone applications, and many others.

While in-plane text recognition in documents is a well-researched and understood problem, text recognition in unconstrained, complex imagery has low performance due to the following challenges: scene complexity with a variety of backgrounds, level of occlusion and clutter, low or uneven lighting, low and variable resolution, blur and focus issues, motion, out-of-plane text, font size or type variation, text along a curve and multi-language environments. These factors necessitate advancement of dedicated computer vision and pattern recognition algorithms.

The terms text detection, localization, and recognition are often used reciprocally in the literature. However, since all the images in the TRAIT dataset have text in them, we define: text detection is the process of determining the location of text in the image and generating bounding boxes around the text; and text recognition is the process of recognizing the text contained within the images.

The TRAIT dataset used for the evaluation is an operational child exploitation collection containing illicit images. (The images are present in digital media seized in criminal investigations.) This text is human-legible and sometimes has investigational value. These images are of interest to NIST's partner law enforcement agencies that seek to employ text recognition in investigating this area of serious crime.

The images in the dataset are mainly of text found in indoor scenes. Also, we did not provide any training dataset or separate training and test vocabularies compared to other datasets.

NIST, and in particular the Information Access Division (IAD), has been organizing and conducting different evaluations in biometrics [3–5], text retrieval [6], machine translation [7], speaker recognition [8], video retrieval [9], etc., in order to support development of error measurement and standards to advance the state of the art in the respective technologies. These efforts have helped to improve the robustness and performance accuracy of these technologies. Hence, we are hopeful that our effort will motivate research in the field of unconstrained text recognition as well..

This report is organized as follows. Section 2 provides a short survey of previous work in this field. The evaluation tasks and conditions are discussed in Section 3. The TRAIT dataset and ground truth (GT) issues are discussed in Section 4. The participating organizations are listed in Section 5. Section 6 introduces the evaluation and comparison metrics and protocols for text detection and recognition. Section 7 presents the results for all three Phases. Finally, concluding remarks and perspectives are given in Section 8.

2. Previous Work

In this report, we describe and present the results of a TRAIT text detection and recognition evaluation. Therefore, we will mainly focus on datasets and evaluation of text detection and recognition systems. Since the last decade, the International Conference on Document Analysis and Recognition (ICDAR) has been developing datasets and running competitions under the title Robust Reading Competitions or Challenges. Since 2003, ICDAR 2003 [10] was the first publicly available dataset for text detection and recognition, two different competitions were run in ICDAR 2005 [11], one in ICDAR 2011 [12] and [13], one in ICDAR2013 [14] and finally in ICDAR 2015 [15].

Six other well-known datasets for text detection and recognition include Coco-text: Dataset for text detection and recognition in natural images [16] with 63 686 images and 145 859 text instances. MSRA Text Detection 500 Database (MSRA-TD500) [17] with 500 natural images, with Chinese, English or mixture of both. The KAIST Scene_Text Database 2010 [18] has 3000 images of indoor and outdoor scenes containing text in Korean, English (Number), and Mixed (Korean + English + Number) and is used for text location, segmentation, and recognition. The Street View Text (SVT) dataset [19] was harvested from "Google Street View" images. The Downtown Osaka Scene Text dataset consists of sequential images captured in shopping streets with an omnidirectional camera [20]. Finally, the Synthetic Word Dataset [21] [22] contains 9 million images covering English words and supports tasks in text recognition and segmentation.

For a complete overview, the following review papers [1, 2] on text detection and recognition have more details about the field, applications, algorithms, and issues.

3. Evaluation Tasks and Conditions

The TRAIT test included detection and recognition tasks for still images. As described in Table 2, the test is intended to support operations in which an automated text recognition engine yields text that can be indexed and retrieved using mainline text retrieval engines. For Phase 3 an optional task was introduced that detected the number of lines of Uniform Resource Locator (URL) along with text recognition of the URLs. We also requested the particpants provide timing information for all the algorithms.

Table 2. Subtests supported under the TRAIT 2016 activity

	Class A	**Class B**	**Class C**	**Class D**
Aspect	Text detection	Text recognition	Text detection and recognition	URLs detection and recognition (optional)
Languages	Mostly English. Some French, Spanish and German. While some Cyrillic and Chinese appear also, evaluation was confined to English roman alphabets only.			
Input	Image(s)	Image(s) and location(s) of text	Image(s)	Images(s)
Output	Given an input image, output detected locations of text. This does not require the algorithm(s) to produce strings of text.	Given an input image and location(s) of text in the image, output strings of text.	Given an input image, output strings of text along with their corresponding locations in the image.	Given an input image, detect zero or more URLs and recognize the text in the image.

3.1 Offline Testing

TRAIT is intended to mimic operational reality. As an offline test intended to assess the core algorithmic capability of text detection and recognition algorithms, it does not extend to do real-time transcription of live image sources. Offline testing is attractive because it allows uniform, fair, repeatable, and efficient evaluation of the underlying technologies. Testing of implementations under a fixed Application Programming Interface (API) allows for a detailed set of performance related parameters to be measured. The algorithms are run only on NIST machines by NIST employees.

3.2 Phased Testing

To support development, TRAIT was conducted in three Phases. In each Phase, NIST has evaluated implementations on a first-come-first-served basis and returned results to participants. In Phase 3, for Class B we provided bounding box coordinates to algorithms where available; otherwise, we provided lines. After Phase 1 and Phase 2, a combined report card was sent to the participants, which included the Phase 1 and 2 results, cropped images of some of the text in the TRAIT images and spatial sampling rate in the TRAIT images. The provided performance feedback may have helped in algorithm improvement in the subsequent Phase.

4. The TRAIT Dataset

The TRAIT dataset is an operational child exploitation collection containing illicit images. The images are present in digital media seized in criminal investigations. The text in these images is human-legible and sometimes has investigational value. These images are of interest to NIST's partner law enforcement agencies that seek to employ text recognition in investigating this area of serious crime. The images are of children ranging from infant through adolescent. Their faces are the subject of a separate face recognition evaluation and development effort (CHEXIA-FACE 2016). Many of the images contain geometrically unconstrained text. Such text is visible on certificates, posters, logos, uniforms, sports apparel, computer screens, business cards, newspapers, books lying on tables, cigarette packets and a long list of rarer objects. Images also contain text on clothing, low-resolution text, exotic fonts, symbols, and numbers. There also are instances where watermarks or logos were post-processed into the image. The text is most commonly in English with French, Spanish, German and Cyrillic present in significant quantity. We did not intend to test non-Roman alphabets. We are unable to show a sample of these images due to the sensitivity, instead we took a few pictures to illustrate the complexity (Figure 1).

Compared to other text recognition datasets or competitions, the images in the TRAIT dataset are mainly of incidental text in indoor scenes. Also, the TRAIT dataset are based on an operational data collection compared to data capture and curation for other datasets. Finally, we did not provide any training dataset or training/test vocabularies as included in other datasets.

The annotators used the Netclean Analyze tool which was modified to allow export of ground truth information into Comma Separated Values (CSV) files.

Out of the approximately 130K images that were ground-truthed, 5348 images (4%) contain text. However, text is more common than this suggests, because this calculation does not account for visual groups (i.e., duplicate photographic events). The TRAIT dataset statistics are shown in Table 3. There is some variation in image sizes as shown in the scatterplot between the image width and the image height in Figure 2. The histograms of the image width and the image height are also shown on the top and on the right side of Figure 2. Table 4 shows the six most common image sizes and their frequency.

4.1 Ground Truth and Issues

There are a number of ways text can be annotated in an image. The most common type of ground truth annotations are shown in Figure 3: (a) character based; (b) word based;

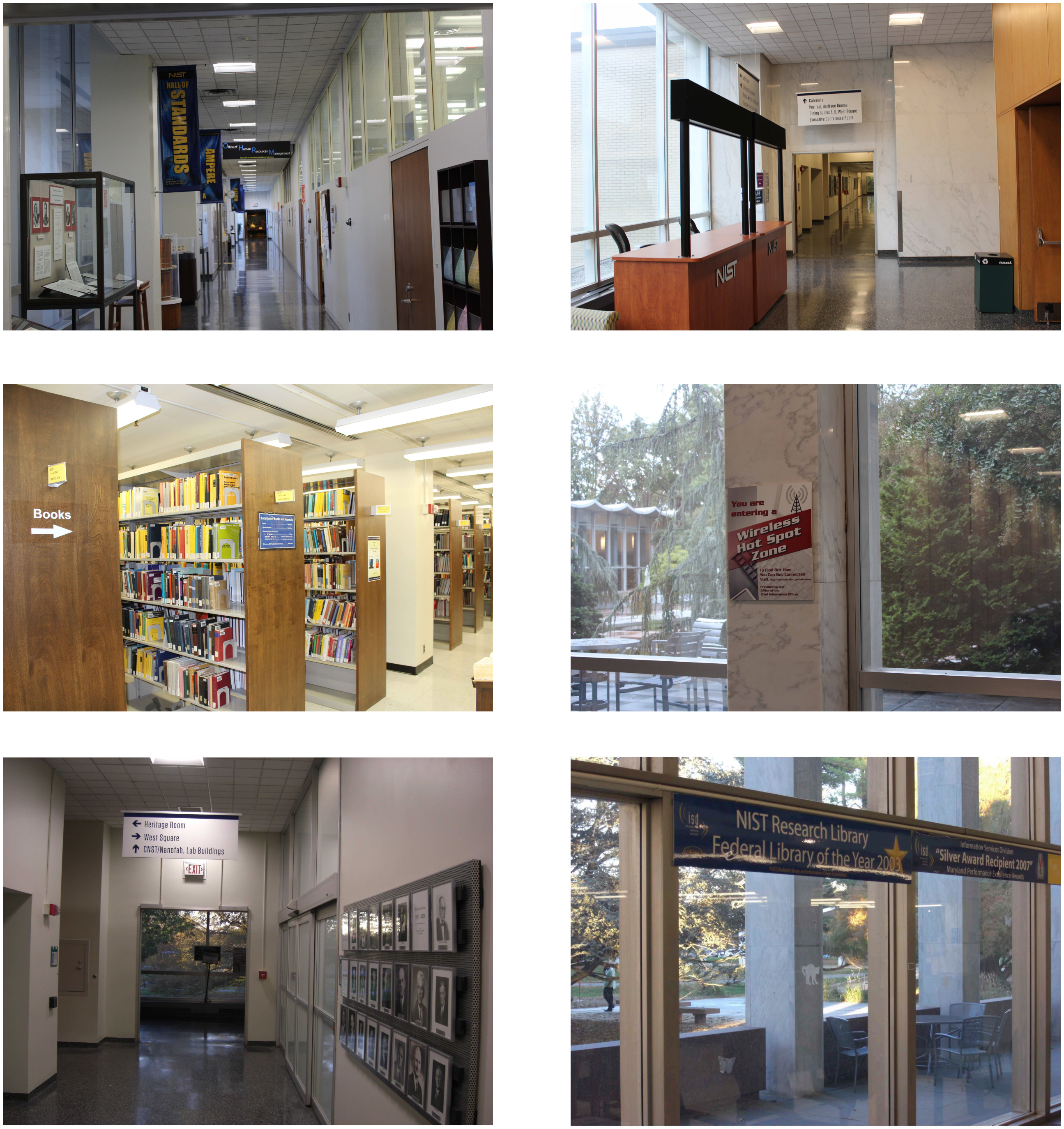

NIST
HALL OF
STANDARDS
AMPERE
Office of Human Resources Management
NIST
Books
You are entering a
Wireless
Hot Spot
Zone
Heritage Room
West Square
CNST/Nanofab, Lab Buildings
EXIT
NIST Research Library
Federal Library of the Year 2003
FLICC/Federal Library and Information Center Committee
isd
Information Services Division
"Silver Award Recipient 2007"
Maryland Performance Excellence Awards

Fig. 1. Sample images captured at NIST to illustrate the complexity of the data for text recognition

Table 3. TRAIT Dataset Statistics

- Number of unique images = 5348
- Number of annotations = 8741
  - Total number of ground truth bounding boxes = 4942
    - 186 have no associated ground truth text; Reason is unknown
  - Total number of ground truth lines = 3799
    - 52 have no associated ground truth text; Reason is unknown
- Number of annotations with Uniform Resource Locator (URL) = 2164
- Number of annotations with numbers = 835
- Mean number of annotations per image = 1.6
- Mean number of letters per annotation = 17.6
- Mean number of words per annotation = 2.4

Table 4. Common Size of Images (pixels). 81% of the images come in six sizes in both portrait and landscape mode.

Size of Images	# Images	Proportion of Total (5348)
768x1024	1382	26%
960x1280	1127	21%
640x480	722	14%
800x600	493	9%
1600x1200	324	6%
576x768	310	6%

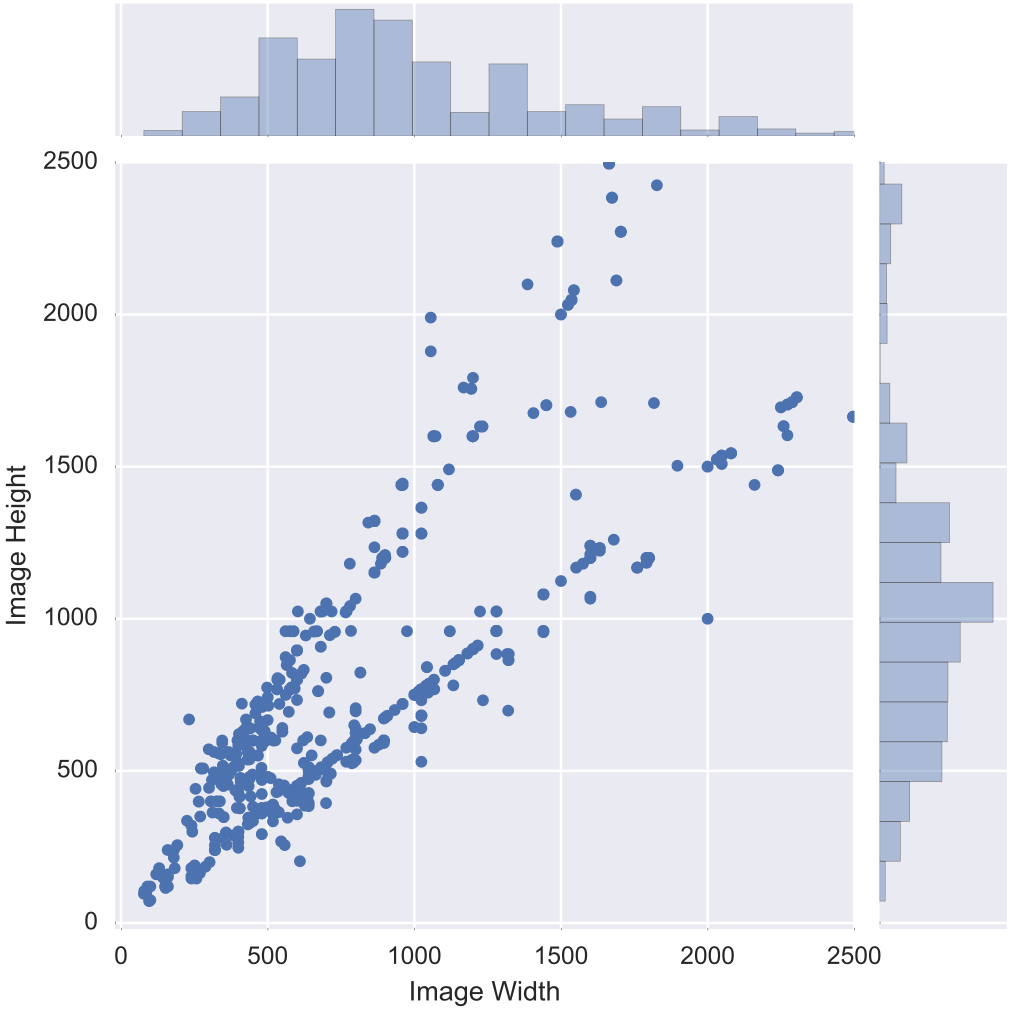


Fig. 2. Scatter plot of image width vs. image height for the TRAIT dataset and the histogram of the image width on the top and the histogram of the image height on the right. The image has two symmetrical lines along the diagonal due to the fact that the dataset has the same size images in both landscape and portrait mode

(c) line going through the center of text; (d) bounding box around the text; (e) pixel based ground truth; and finally (f) annotation using polygons around the text. The text recognition community, in general, has moved toward annotation using polygons. This supports annotation of out of plane text. Polygons were not used in TRAIT, because ground-truthing is time intensive. Because of the sensitivity of the images, crowdsourcing annotation was not considered.

Before the start of the Phase 1 evaluation, we were expecting only line annotation through the text. After delivery of the data to NIST, we realized that almost 50% of the data was annotated using bounding boxes. However, since the TRAIT API had already been published and implemented, TRAIT Phase 1 and Phase 2 proceeded with only line annotations. This was suboptimal, so for Phase 3 evaluation, both bounding boxes (where available) and lines were provided.

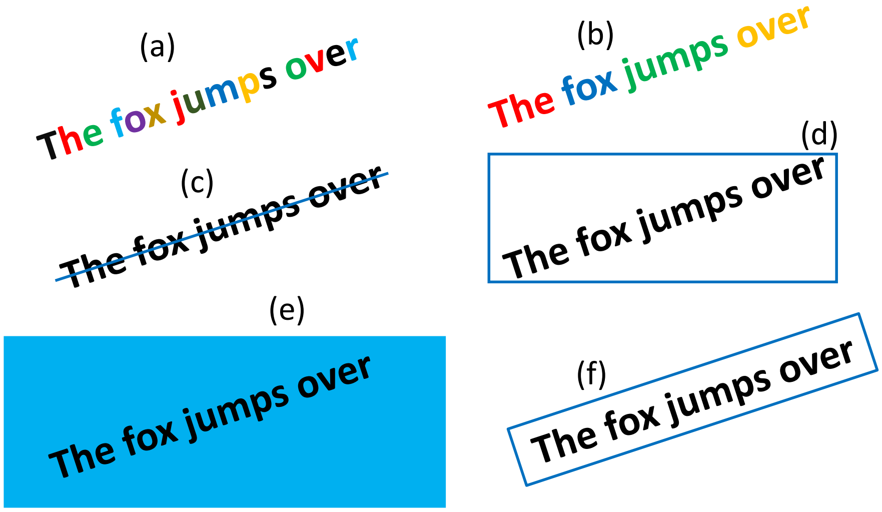


Fig. 3. The different types of GT annotations: (a) character based; (b) word based; (c) line based with GT line going through the center of text; (d) bounding box around the text; (e) pixel based ground truth; and finally (f) annotation using polygons around the text

For any given piece of text, the text was annotated with either a line or a bounding box as shown in Figure 4. The lines go through the center of the text and bounding boxes surround the text in the images. For Phase 1 and 2, the bounding box annotations are converted into lines and this is what was provided to the algorithm as shown in Figure 5. For Phase 3, both lines and bounding boxes are provided to the algorithm.

In some cases, the conversion from bounding boxes to lines created the location lines at incorrect locations, for example, when the bounding box was drawn with an inappropriately large spatial extent. This caused the derived ground truth line to be displaced, and in the case of text that was at an angle, the resulting line was not centered through the text.

There were several rare markup inconsistencies observed in the ground truth of the dataset:

- The annotation bounding boxes were drawn around multiple lines of text.
- Words describing a texture of a curtain when there was no text on the curtain.
- Bounding boxes and lines that indicated the presence of text but did not contain corresponding ground-truth text strings.
- Ground truth lines that didn't go through the middle of the text but underlined the text.
- There was a 90-degree rotation bug introduced by the annotation tool, rotating the image, if Exif (in image metadata) said portrait, but not transposing the text coordinates back accordingly.

By analyzing cropped images of text extracted from the TRAIT images and comparing them to text reported by algorithms, we can state that the ground truth accuracy is at least an order of magnitude higher than those obtained by the text recognition algorithms being evaluated. The recognition errors that are caused by annotation issues are rare compared to errors due to the natural difficulty of text in TRAIT images. So, the results presented are valid. In the future, we will try to correct some of the issues with the ground truth annotation data.

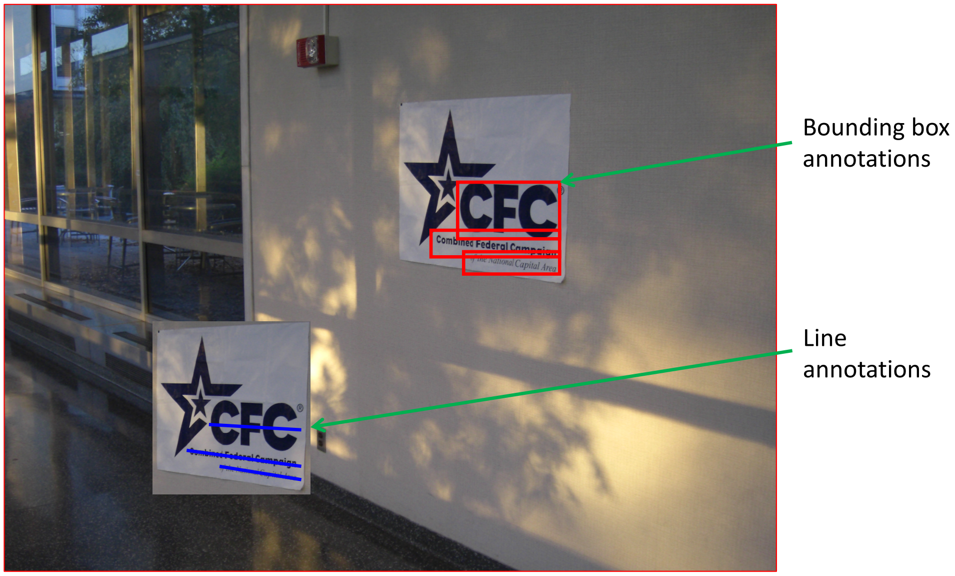
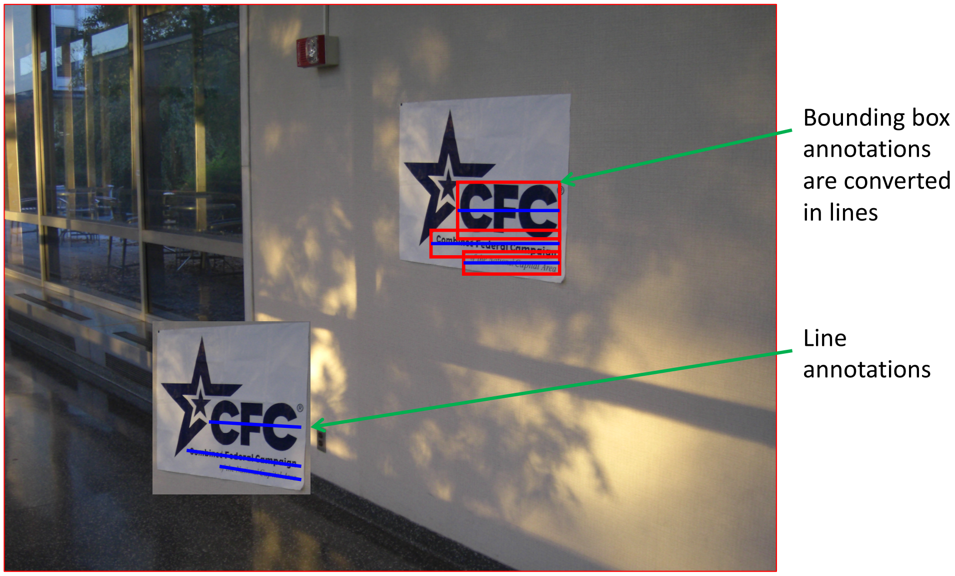


Fig. 4. Types of ground truth image annotations with bounding boxes and lines used in TRAIT



Bounding box annotations are converted in lines

Line annotations

Fig. 5. Bounding box annotations are converted into lines for Phase 1 and 2, and this is what was provided to Class B algorithms. For Phase 3, we used boxes when available and lines otherwise

4.2 Text Properties in TRAIT Imagery

This subsection contains data related to the spatial sampling rate of text in the TRAIT images.

Figure 6 shows the distribution of bounding box width over the number of characters. This distribution has a second peak. This is due to the fact that 12% of the bounding boxes have two lines. Also, the text has a mean resolution of 15 pixels per character.

Figure 7 shows the distribution of line length over the number of characters. The main difference is the second peak does not show up in the line case compared to the bounding box case, since the line annotations, only have one line. The text in the line case has a mean resolution of 17 pixels per character.

Next, Figure 8(a, b), shows bounding box height vs. number of characters. The large scatter in the figure is because of the large variety of fonts being used and because of annotation issues.

As illustrated in Figure 9, the second peak shows up in the bounding box case, since there are two lines in the bounding box annotations.

5. Participating Organizations

In total three groups took part in TRAIT evaluation over three Phases, which are listed in Table 5 and Table 6. TRAIT was open to a worldwide developer audience. Participation

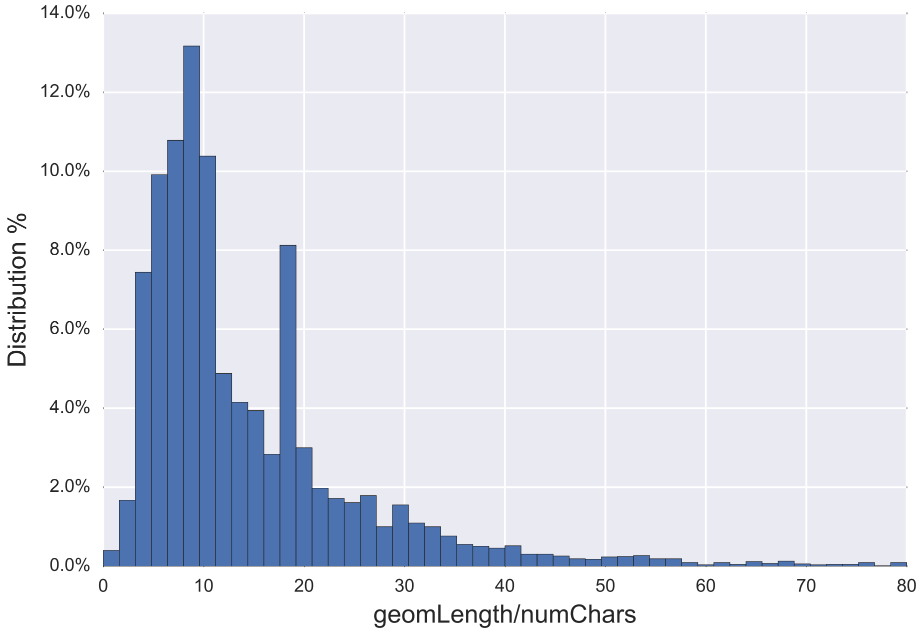


Fig. 6. Histogram of bounding box width over the number of characters contained. The second peak only shows up in the bounding box case, not for the line case. This could be due to the fact that 12% of the bounding boxes have two lines. The text has a mean resolution of 15 pixels per character.

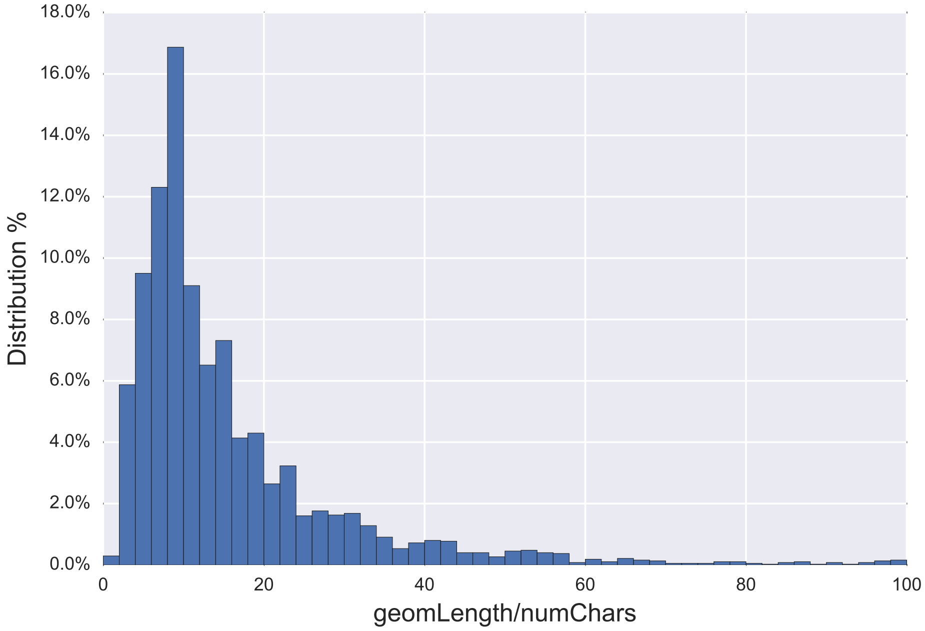
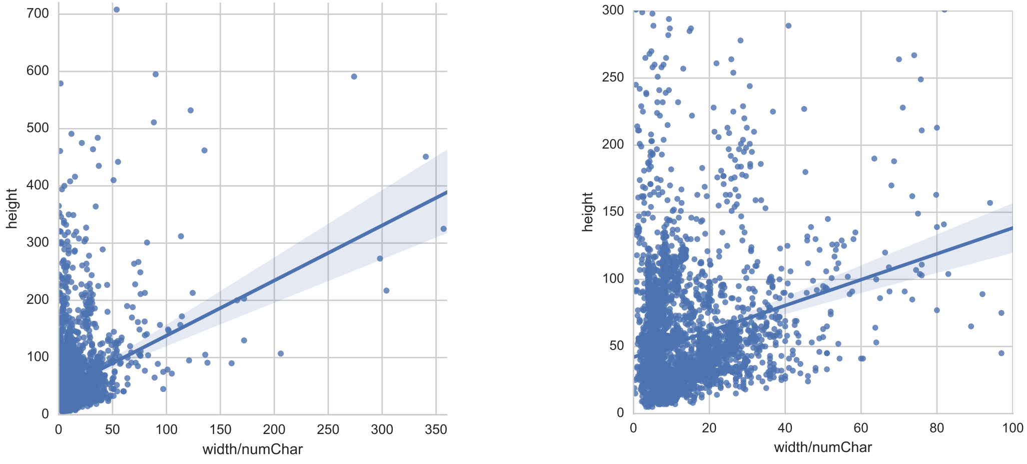


Fig. 7. The histogram of length of line over the number of characters. The text has a mean resolution of 17 pixels per character.



(a) Scatterplot of height vs. width/numChars

(b) Zoomed in version of Figure a

Fig. 8. Scatterplot of bounding box height vs. width over number of characters. Figure b is a zoomed in version of Figure a, to show more details of the scatterplot.

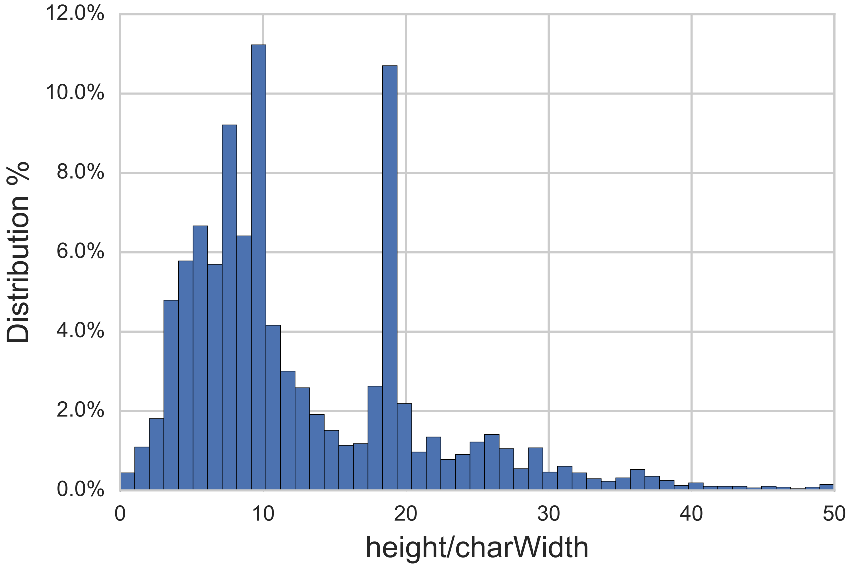


Fig. 9. The histogram of bounding box height over character width

was free, the only cost being that associated with implementing the NIST API.

Table 5. Organizations that participated by different Phases

Participants	Submissions		
	Phase 1	Phase 2	Phase3
Czech Technical University, Prague	✓		✓
Glyphin, Belgium		✓	
Megvii, China		✓	✓

Table 6. Organizations that participated by Class

Participants	Submissions			
	Class A	Class B	Class C	Class D
Czech Technical University, Prague		✓	✓	
Glyphin, Belgium		✓	✓	
Megvii, China	✓	✓	✓	✓

6. Performance Metrics

In the following subsections, we describe the metrics for text detection, text recognition, URL detection and, recognition and timing information.

6.1 Text Detection Metrics

Previously there have been a number of papers on the evaluation of text detection and localization [23–25]. Usually, the detection results are evaluated by comparing the bounding box of the ground truth with the bounding box detected by the algorithm. In some evaluations, the pixel-level ground truth is used to measure localization accuracy. The most common metrics are Precision and Recall measures, which are computed from the intersection area of these two bounding boxes as shown in Figure 10. Precision is the fraction of detected items that are correct. Recall is the fraction of items that were correctly detected among all the items that should have been detected. Details of how to calculate Precision and Recall are in the equations 1, 2 and 3.

Usually, the match between the bounding boxes can have issues: a split error, a one-to-many match with one ground truth bounding box; a merge error, a one-to-many match with one detected bounding box, as shown in Figure 11 and Figure 10. For our study, we didn't calculate the splitting and merging errors. Also, in the case of the ground truth being defined as a line, we used the line intersection with the bounding box as the criteria, as shown in equation 3.

$$
\begin{array}{ll}
TP: & \text{True Positive (based on comparing the bounding boxes/lines)} \\
N_{GT}: & \text{Number of bounding boxes/lines in the GT} \\
N_{Alg}: & \text{Number of detected bounding boxes/lines returned by the Algorithm} \\
TH: & \text{Threshold value} \\
Precision: & P = TP/N_{ALG} \\
Recall: & R = TP/N_{GT} \\
\text{F1-Score}: & F1 = 2\,P\,R/(P+R)
\end{array}
\tag{1}
$$

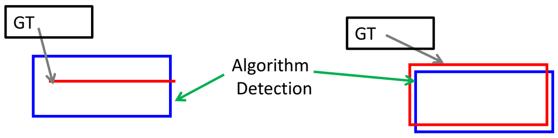


Fig. 10. Text detection match criteria

In case of bounding box annotation:

$$
\begin{array}{l}
\text{If}(A_{GT} \cap A_{ALG})/(A_{GT} \cup A_{ALG}) > TH \\
\text{Then } TP = 1, \text{ else } TP = 0 \\
A_{GT} \text{ and } A_{ALG} \text{ are areas of ground truth and algorithm}
\end{array}
\tag{2}
$$

In case of line annotation:

$$\begin{aligned}&\text{If the \% of line which intersects with the Bbox } > TH\\&\text{Then } TP = 1,\ \text{else } TP = 0\end{aligned} \tag{3}$$

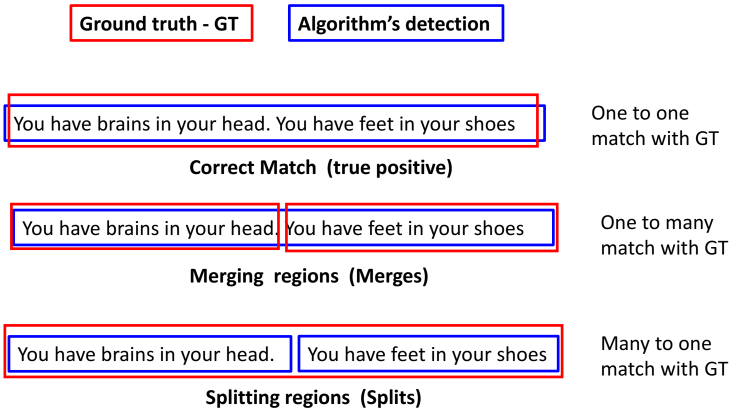

Ground truth - GT
Algorithm's detection
You have brains in your head. You have feet in your shoes
One to one match with GT
Correct Match (true positive)
You have brains in your head.
You have feet in your shoes
One to many match with GT
Merging regions (Merges)
You have brains in your head.
You have feet in your shoes
Many to one match with GT
Splitting regions (Splits)

Fig. 11. Different match types for text detection: 1) one to one match 2) one to many match, and 3) many to one match

Usually, the TH value is set to 0.5, but in our case, the value is set at 0.35. We also enforce a one-to-one correspondence between the GT and detection objects. We have used a lower TH of 0.35 is because the algorithms have split the detection results in several cases.

6.2 Text Recognition Metrics

The text recognition accuracy metrics are based on both character and word-based normalized Edit Distance [26] (Also, called Levenshtein Distance). The Edit Distance is a common similarity measure between two strings. It is defined as the minimum number of insertions, deletions or substitutions of single characters needed to transform one of the strings into the other one. We also have a recognition accuracy metric based on the unordered word list, because the algorithms have split the text recognition results, and ground truth issues.

Since text retrieval is almost always performed on a case-insensitive basis, we have converted all the text into lowercase. We have removed all the punctuations from the text, except hyphens between words.

Stopwords: In text retrieval, commonly used words such as “the,” “of,” “are”, “is“, “and,” “in,” etc., are normally not indexed because they provide essentially no retrieval value; The accuracy based on non-stopwords is even more relevant than the total word accuracy to a text retrieval type application. We have not removed the stop words for this study.

Character based Edit Distance (# errors) :

$$Ed_C = (\text{'cat', 'car'}) \Rightarrow Ed_C = 1$$

Word based Edit Distance (# errors) :

$$Ed_W = (\text{'This is a cat', 'This is a dog'}) \Rightarrow Ed_W = 1$$

You can normalize the Edit Distance with the number of characters or words in the reference (ground truth).

$$\begin{aligned} normEd_C &= ed_C / numChars \quad \text{range } \{\ 0\ ,\ 1\ \} \\ normEd_W &= ed_W / numWords \quad \text{(Also, called Word Error Rate)} \end{aligned} \tag{4}$$

Some evaluations have used Total normalized Edit Distance ($totalNormEd_C$ and $totalNormEd_W$) as the ranking metric, which is the sum of normalized Edit Distance over all annotations over all images.

The Character Edit Distance is used in the evaluation of optical character recognition (OCR) and text recognition, natural language processing, etc. The Word Edit Distance is used for evaluation in machine translation, speech recognition, text recognition, natural language processing, etc.

Character Accuracy:

$$AC_C = 100(1 - normEd_C)/\#Annotations \tag{5}$$

where AC_C is Character based Accuracy.

Word Accuracy (ordered):
Here the word order is taken into account for calculating the accuracy.

$$AC_W = 100(1 - normEd_W)/\#Annotations \tag{6}$$

where AC_W is Word based Accuracy.

Word Accuracy (unordered):
In a text retrieval type application, the correct recognition of words is much more important than the order of words and correct recognition of numbers or punctuation. We calculate the unordered word accuracy %, with a set of threshold values (TH_w) of the normalizedEditDistance = { 0.0, 0.1, 0.2, 0.3 } for the word matching criteria.

$$AC_{Wuo} = \#wordMatched / numWords \tag{7}$$

The AC_{Wuo} is unordered Word based Accuracy. Where $\#wordMatched$ is words matched correctly based on using the normalizedEditDistance ($Walg, Wgt$) $\leq TH_w$ criteria at different threshold values { 0.0, 0.1, 0.2, 0.3 } in an image and $numWords$ is the total number words in the ground truth.

Finally, for Class C evaluation, when the ground truth (GT) lines or bounding boxes intersect with multiple algorithm polygons by more than 1% we select the GT and algorithm polygon with the lowest normalized character Edit Distance. We also enforce the one-to-one correspondence between GT and algorithm results.

6.3 URL Metrics

The Uniform Resource Locator (URL) metrics are based on reporting the, % of URLs detected correctly, % of the number of URLs detected correctly and the % accuracy of recognition of URLs. The accuracy of recognition of URLs was based only on the domain name. We ignored punctuation, (e.g., {., :, /, ?, @, etc.}), protocol text, (e.g., http, https, etc.), and top level domain words (e.g., com, org, info, net, etc.)

6.4 Timing Metrics

The timing metrics are based on reporting the median timing/image and the boxplot of the time duration for all the detection and recognition algorithms. All timing tests were executed on unloaded machines running a single process. We executed the software on Dual Intel Xeon E5-2695 3.3 GHz CPUs (14 cores each; 56 logical CPUs total) with 227 Dual NVIDIA Tesla K40 Graphics Processing Units (GPUs) with 64-bit version of CentOS 7 operating system running Linux kernel 3.10.0. The time limits for text detection and recognition algorithms were less than 10 seconds for Phase 1 and 2. This limit was removed from Phase 3.

7. Results of the Evaluations

In this section, we compare the results for the participants' runs by task, for the three Phases. Only Megvii provided results for Class A. In total, three groups participated, with a total of 10 results for Class B and Class C. For Phases 1 and 2 of Class B, line location alone was provided to the algorithms. Whereas in Phase 3, the location(s) of the text was being provided to the algorithm either in the form of a line through the centroids of the text or a closed bounding box polygon around the text. Most of the text locations provided via bounding boxes contain a single line of text, but some contain multiple lines of text. The bounding boxes can be both tight around the text or contain extra space. For Phase 3, a optional task for URL detection and recognition was introduced (Class D); also the algorithm time constraints for text detection and recognition were removed.

7.1 Text Detection Results (Class A)

For Class A, the algorithm detects the location of text in the images as bounding boxes for a given input image. Table 7 shows detection performance based only on Class A results. Only Megvii provided the submissions for Class A. For all the other algorithms, we use Class C submissions to generate the Precision, Recall and F1-Score results as shown in Table 8. The Precision, Recall and F1-Score results for Megvii for both Class A and Class C are the same. Megvii C30A/C30C had the best performance based on F1-score, followed by Megvii C32A/C32C and both Megvii C31A/C31C and C21A/C21C came in third. However for performace based on Recall, Czech A30C has the best result.

Table 7. Class A: The detection performance based on Precision, Recall and F1-Score for the three Phases of evaluation. Based on F1-score, Red is number one, Blue is number two and Green is number three in the evaluation

Phases	Participants Methods	Prec	Recall	F1-score
Phase3	Megvii C32A	0.24	0.42	0.31
	Megvii C31A	0.25	0.34	0.29
	Megvii C30A	0.29	0.40	0.34
Phase2	Megvii C22A	0.20	0.42	0.27
	Megvii C21A	0.24	0.37	0.29
	Megvii C20A	0.22	0.37	0.27

Table 8. The detection performance based on Precision, Recall and F1-Score for the three Phases of evaluation based on Class C results. Based on F1-score, Red is number one, Blue is number two and Green is number three in the evaluation

Phases	Participants Methods	Prec	Recall	F1-score
Phase3	Megvii C32C	0.24	0.42	0.31
	Megvii C31C	0.25	0.34	0.29
	Megvii C30C	0.29	0.40	0.34
	Czech A30C	0.11	0.47	0.18
Phase2	Megvii C22C	0.20	0.42	0.27
	Megvii C21C	0.24	0.37	0.29
	Megvii C20C	0.22	0.37	0.27
	Glyphin B21C	0.08	0.10	0.09
	Glyphin B20C	0.06	0.07	0.07
Phase1	Czech A10C	0.10	0.23	0.14

7.2 Text Recognition Results (Class B)

For Class B, the algorithm recognizes text based on an input image and the ground truth locations of the text provided either as a line or bounding box. Table 9 and Figure 12 shows the results, calculating text recognition accuracy by character and word Edit Distance. Accuracy required the words to be reported in the right sequence. The results show that Megvii C32B is most accurate with a word-based accuracy of 21.4% and character based accuracy of 39.9%. The next most accurate algorithms were Megvii C31B and Megvii C30B. The performance of participants improved from Phase 1 to Phase 3.

The results in Table 10 consider the accuracy of algorithms without considering the reported order of words. We calculate the unordered word accuracy with a set of threshold values (TH_w) of the normalizedEditDistance= { 0.0, 0.1, 0.2, 0.3 } for the word matching criteria. If R out of N words are matched correctly, the unordered word accuracy is R/N. When matched correctly, normalizedEditDistance (W_{alg}, W_{gt}) $<= TH_c$ (threshold value). Tesseract results are presented as baseline performance and basis for comparison. In the case of Tesseract, cropped Bounding boxes were provided, instead of the whole image and location information. Tesseract is a open source OCR engine that has unicode (UTF-8) support, and can recognize more than 100 languages.

Table 9. Class B: The character and word based % accuracy of recognition for all the algorithms for the three Phases. The accuracy is based on the edit distance of the ordered word list. For character and word based accuracy, Red is number one, Blue is number two and Green is number three in the evaluation.

Phases	Participants Methods	Accuracy of Recognition (%) (Char)	Accuracy of Recognition (%) (Word)
Phase 3	Megvii C32B	39.9	21.4
	Megvii C31B	37.3	20.5
	Megvii C30B	36.7	18.1
	Czech A30B	27.1	5.9
Phase 2	Megvii C22B	29.2	11.6
	Megvii C21B	27.7	12.8
	Megvii C20B	28.0	12.9
	Glyphin B20B	5.8	4.2
	Glyphin B21B	6.6	4.5
Phase 1	Czech A10B	15.1	4.1
	Tesseract T10B	7.4	3.4

7.3 Text Detection and Recognition Results (Class C)

For Class C, for a given input image the algorithm reports the location of text and the content of text in the image. Table 11 and Figure 13 show the character and word accuracy % results. Notice that the participants' performance improved from Phase 1 to Phase 3. When

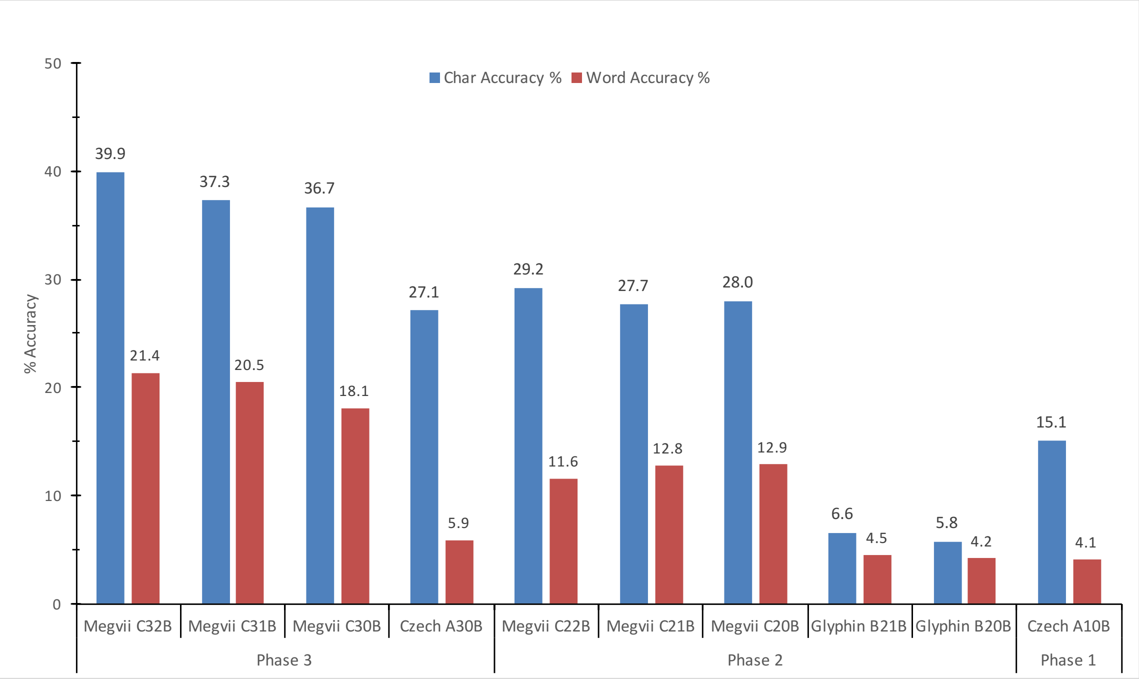


Fig. 12. Class B: The bar chart of character and word based accuracy of recognition for all the algorithms

Table 10. Class B: The unordered word based % accuracy of recognition for the three phases. The accuracy is based on the edit distance of the unordered word list and calculated with four levels of threshold $TH_c = \{0.0, 0.1, 0.2, 0.3\}$. For TH_c=0.0, Red is number one, Blue is number two and Green is number three in the evaluation.

Phases	Participants Methods	Accuracy of Recognition (%) TH_c=0.0	Accuracy of Recognition (%) TH_c=0.1	Accuracy of Recognition (%) TH_c=0.2	Accuracy of Recognition (%) TH_c=0.3
Phase 3	Megvii C32B	22.8	44.8	50.9	56.7
	Megvii C31B	17.3	32.6	37.8	41.5
	Megvii C30B	22.1	54.8	65.0	73.3
	Czech A30B	7.4	9.0	12.8	15.5
Phase 2	Megvii C22B	10.9	21.5	26.5	28.8
	Megvii C21B	11.7	25.4	30.3	33.1
	Megvii C20B	11.0	24.9	29.7	32.3
	Glyphin B21B	1.7	1.8	2.0	2.3
	Glyphin B20B	1.8	1.9	2.3	2.4
Phase1	Czech A10B	6.1	8.0	11.1	14.0
	Tesseract T10B	7.1	9.9	10.8	12.0

the GT lines or bounding boxes intersect a few of the algorithm polygons, by more than 1%, then, we calculate the Edit Distance between all of them and select the one with the lowest normalized Edit Distance (Char). We then use Edit Distance to determine the character and (ordered) word based accuracy. We also enforce the one-to-one correspondence between GT and algorithm results. The best results for ordered words, in order, were Megvii C32C, Megvii C31C, and Megvii C30C. Table 12 shows the results if we consider the unordered word accuracy instead. In that case, the best results are Megvii C30C, Megvii C32C and Megvii C31C. We calculate the unordered word accuracy with a set of threshold values (TH_c) of the normalizedEditDistance = $\{0.0, 0.1, 0.2, 0.3\}$ for the word matching criteria. If R out of N words are matched correctly, the unordered word accuracy is R/N, where match correctly is normalizedEditDistance $(W_{alg}, W_{gt}) <= TH_c$ (threshold value).

Table 11. Class C: The character and word based % accuracy of recognition for the three phases. The accuracy is based on the edit distance of the ordered word list. For character and word based accuracy, Red is number one, Blue is number two and Green is number three in the evaluation.

Phases	Participants Methods	Accuracy of Recognition (%) (Char)	Accuracy of Recognition (%) (Word)
Phase 3	Megvii C32C	44.8	22.8
	Megvii C31C	39.6	20.2
	Megvii C30C	41.1	19.9
	Czech A30C	35.5	12.0
Phase 2	Megvii C22C	29.2	9.6
	Megvii C21C	27.1	10.2
	Megvii C20C	27.5	10.4
	Glyphin B21C	9.5	4.9
	Glyphin B20C	6.7	3.5
Phase 1	Czech A10C	24.1	7.2

7.4 URL Detection and Recognition Results (Class D)

We added three optional URL related tasks to Phase 3, as shown in Table 13. Algorithms were meant to both detect images with URLs and number of URLs in images and then recognize the text in all the URLs detected. Only Megvii took part in these tasks, and of their multiple algorithms, C32C had the overall best performance. It detected 30.5% of URLs correctly, 13.1% of which were accurately recognized.

7.5 Timing Results

The timing information is presented in Table 14 and Figure 14 for all the runs for the different algorithms for Class A. The most time consuming algorithms for Class A was Megvii's C32A, followed by Megvii's C30A. The timing information is presented in Table 15 and Figure 15 for all the runs for the different algorithms for Class B. Table 16 and

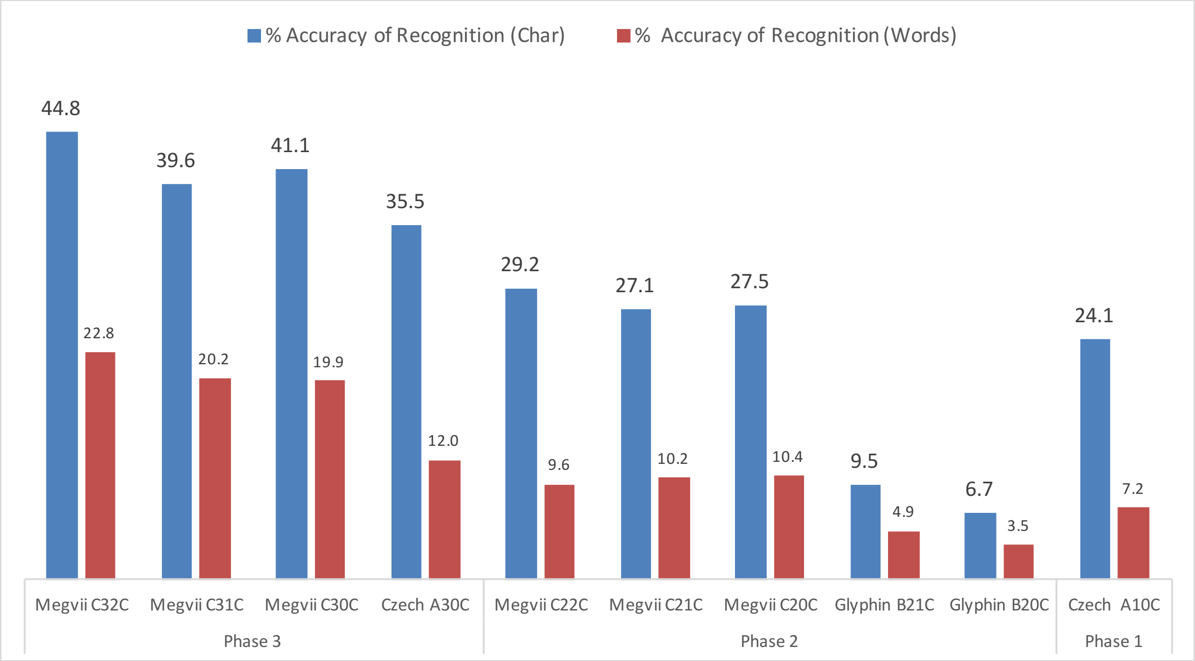


Fig. 13. Class C: The character and word based accuracy for the algorithms

Table 12. Class C: The unordered word based % accuracy of recognition for the three Phases. The accuracy is based on the edit distance of the unordered word list and calculated with four levels of threshold TH_c = 0.0, 0.1, 0.2, 0.3. For TH_c = 0.0, Red is number one, Blue is number two and Green is number three in the evaluation.

Phases	Participants Methods	Accuracy of Recognition (%) TH_c=0.0	Accuracy of Recognition (%) TH_c=0.1	Accuracy of Recognition (%) TH_c=0.2	Accuracy of Recognition (%) TH_c=0.3
Phase 3	Megvii C32C	25.4	47.3	53.4	59.5
	Megvii C31C	20.0	36.8	42.3	46.7
	Megvii C30C	25.6	48.3	56.1	62.4
	Czech A30C	17.52	25.4	32.9	38.7
Phase 2	Megvii C22C	10.1	21.6	24.1	25.4
	Megvii C21C	12.7	28.1	30.9	32.8
	Megvii C20C	11.9	27.6	30.5	32.7
	Glyphin B21C	2.2	2.3	2.5	2.6
	Glyphin B20C	2.1	2.3	2.5	2.7
Phase 1	Czech A10C	10.0	12.4	14.7	17.3

Table 13. Detect images with URLs correctly, number of URLs detected in images and accuracy for text recognition of URLs

Participants Methods	% URLs Detected Correctly	% #URLs Detected Correctly	% Accuracy of Recognition of URLs
Megvii C32C	**30.5**	**22.6**	**13.1**
Megvii C31C	23.2	17.6	9.5
Megvii C30C	23.8	17.8	9.3

Figure 16 show the same for the Class C. The most time consuming algorithm for Class B was Megvii's C32B, followed by Megvii's C31B. The most time consuming algorithms for Class C was Megvii's C32C, followed by Megvii's C30C. Glyphin's algorithms took the least time, with B20B as the fastest algorithm in Class B and B20C as the fastest algorithm in Class C. It should be noted that Glyphin's algorithms were timing out on a large number of images. The time limit for Phase 1 and Phase 2 was 10 seconds. From the results, we noticed that for Megvii, from Phase 2 to 3, their accuracy improved by 70%, but their time more than doubled in order to accomplish this (more time, more accuracy); however, for the Czech algo from Phase 1 to 3, their accuracy also improved (by 52%), but their time was reduced by much more (more than an order of magnitude for Class C). Also, from Table 15 and Table 16, the time duration for Class C (both text detection and recognition) is less than the time duration for Class B (recognition only).

Table 14. The text detection durations for Class A

Phases	Participants Methods	Median timing/image (millisecs)
Phase3	Megvii C32A	57545
	Megvii C31A	20202
	Megvii C30A	26721
Phase2	Megvii C22A	8673
	Megvii C21A	10650
	Megvii C20A	10717

7.6 Text Resolution Results

This subsection contains data related to the resolution of text in the TRAIT images. Figure 17 shows the overlay of two histograms, one for all text entries, the other where the algorithm produced no results for six different submissions. The main reason B20B has many values where the algorithm produces no results is that the algorithm timed out for many of the images. From the figure, we can conclude that text resolution is not the main cause of the detection failure.

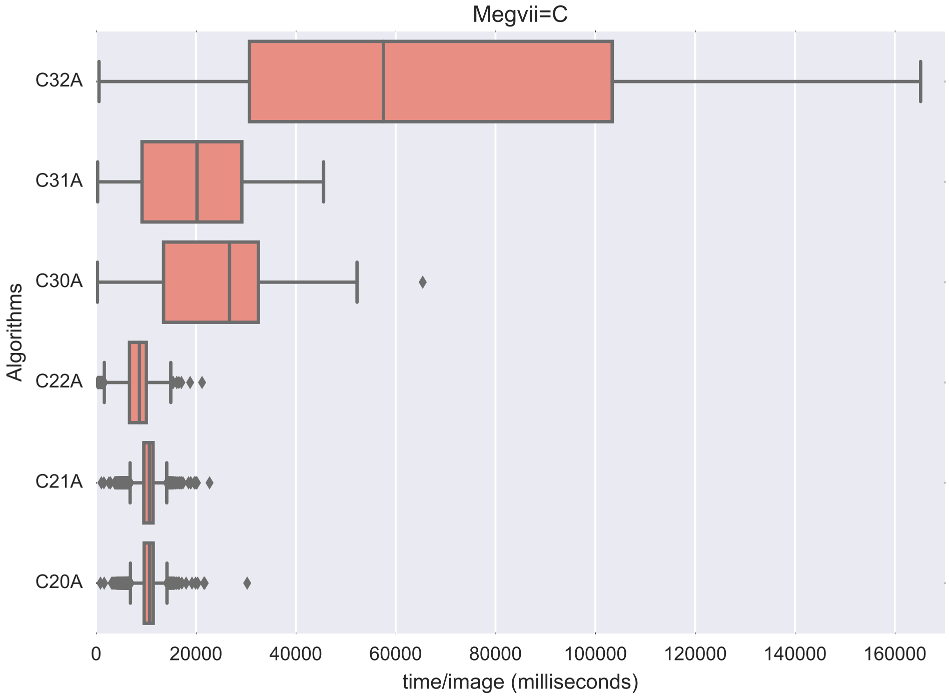


Fig. 14. The boxplot of the text detection durations for all the algorithms in Class A evaluation

Table 15. The text recognition durations for Class B

Phases	Participants Methods	Median timing/image (millisecs)
Phase3	Megvii C32B	59916
	Megvii C31B	20032
	Megvii C30B	24783
	Czech A30B	490
Phase2	Megvii C22B	8607
	Megvii C21B	10664
	Megvii C20B	10731
	Glyphin B20B	6006
	Glyphin B21B	65
Phase1	Czech A10B	2538

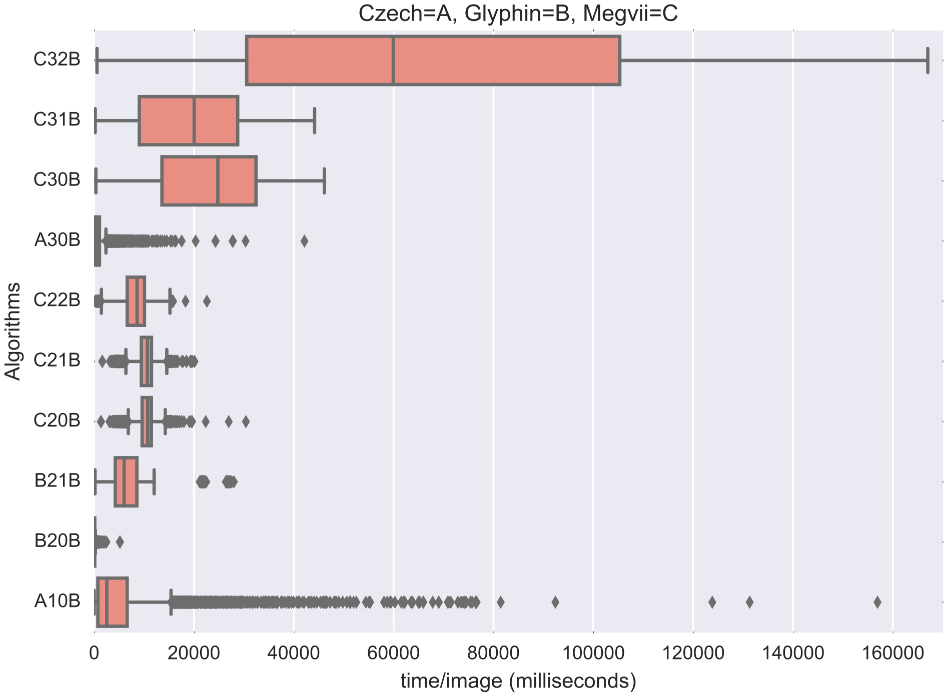


Fig. 15. The boxplot of the text recognition durations for all the algorithms in Class B evaluation

Table 16. The text detection and recognition durations for Class C

Phases	Participants Methods	Median timing/image (millisecs)
Phase3	Megvii C32C	56380
	Megvii C31C	20064
	Megvii C30C	24847
	Czech A30C	509
Phase2	Megvii C22C	8504
	Megvii C21C	10640
	Megvii C20C	10629
	Glyphin B21C	5980
	Glyphin B20C	125
Phase1	Czech A10C	19540

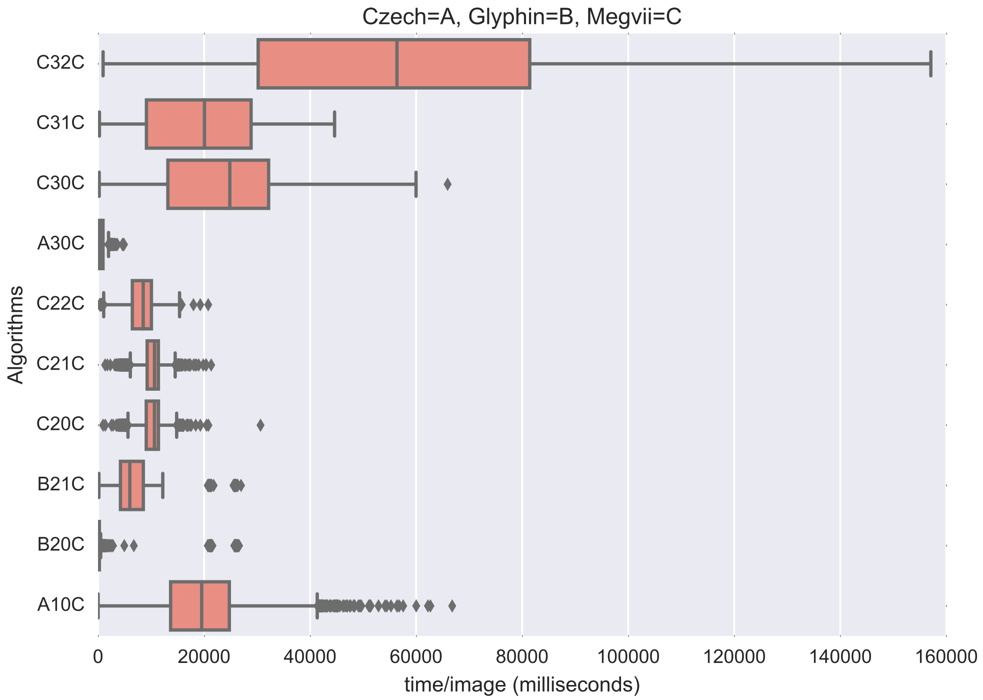
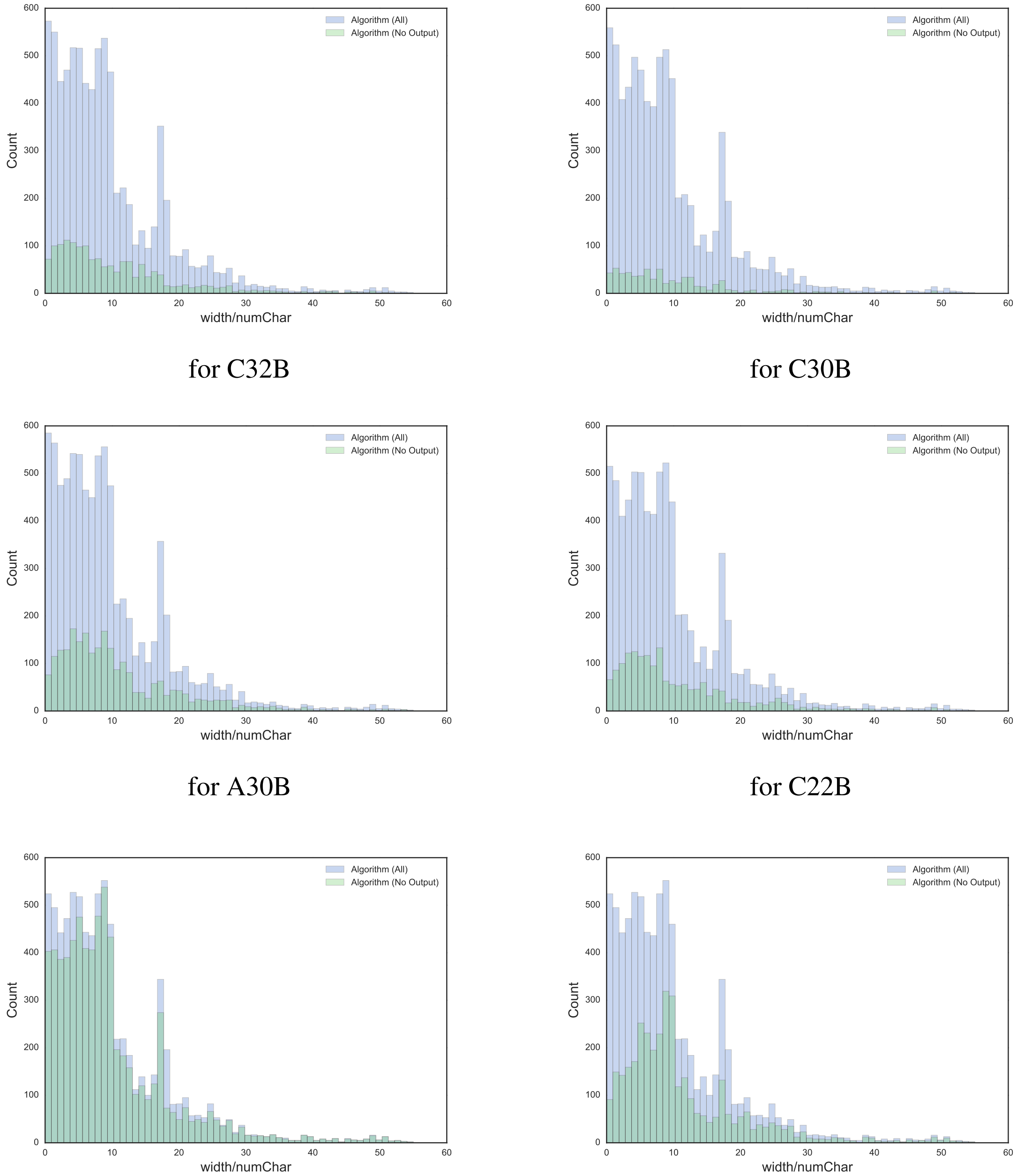


Fig. 16. The boxplot of the text detection and recognition durations for all the algorithms in Class C evaluation



for C32B

for C30B

for A30B

for C22B

for B20B

for A10B

Fig. 17. Shows the overlay of two histograms, one for all algorithm text entries, the other where the algorithm produced no results for six different submissions. The width/numChar is based on the ground truth data of both lines and bounding boxes. For B20B, the algorithm timed out for most of the images. From the figure, we can conclude that text resolution is not the main cause of the detection failure. The issues could be the annotations, or, image quality, font type, compression, blur, view, background, etc.

8. Conclusions

An overview of the TRAIT evaluation and the results are presented in this report. The evaluation results show that the performance of unconstrained text recognition was low. However, from Phase 1 to Phase 3, the performance of the text recognition algorithms have shown significant improvement. From first test (in either Phase 1 or 2) to Phase 3, we saw an average improvement across the classes of 61%, with Megvii improving by 70% and the Czech algorithm by 52%. Based on these results and other text detection and recognition competitions/evaluations, we conclude that there is still much room for improvement in unconstrained text recognition. We hypothesize that the likely factors affecting text recognition performance are: low resolution of the text in the images, compression of images, issues with the annotations (multiple lines, describing an image, etc.), text on clothing, partial occlusion, view angle, blur, focus issues, lighting issues, and shadows. In addition, the text text has many URLs and punctuation (some of the algorithms performed poorly on punctuation and numbers).

Although three groups participated in the TRAIT challenge, we hope this evaluation will motivate more research in this exciting field. We plan to present the results of the competition in a journal paper.

NIST has been organizing and conducting different evaluations in biometrics, text retrieval, machine translation, speech recognition, video retrieval, etc., which has helped to improve the robustness and performance accuracy of these technologies. Similarly, we are hopeful that in the future we can help improve the measurement infrastructure for unconstrained text recognition through 1) development of large and challenging benchmark data sets with highly accurate ground truth for localization and recognition, and 2) development of performance measurement methodologies, tools and evaluation infrastructure.

Acknowledgments

The authors would like to thank the Department of Homeland Security (DHS) Science and Technology (S&T) Directorate, First Responders Group, Office for Public Safety Research for support of this effort. The TRAIT activity falls under their Child Exploitation Image Analytics Program (CHEXIA). In addition, we are indebted to DHS Homeland Security Investigations for access to the image data and to Pacific Northwest National Laboratory for coordinating the extensive ground truthing effort.

References

[1] Jung K, Kim KI, Jain AK (2004) Text Information Extraction in Images and Video: a Survey. *Pattern Recognition* 37(5):977–997.

[2] Ye Q, Doermann D (2015) Text Detection and Recognition in Imagery: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37(7):1480–1500.

[3] Grother PJ, Ngan ML, Quinn GW (2017) Face In Video Evaluation (FIVE) Face Recognition of Non-Cooperative Subjects. *NIST Interagency/Internal Report (NISTIR)-8173* .

[4] Ngan ML, Quinn GW, Grother PJ (2016) NISTIR 8078-Tattoo Recognition Technology-Challenge (Tatt-C) -Outcomes and Recommendations. *NIST Interagency/Internal Report (NISTIR)-8078-rev1* .

[5] Grother PJ, et al. (2012) IREX III-Performance of Iris Identification Algorithms. *NIST Interagency/Internal Report (NISTIR)-7836* .

[6] Robertson SE, Soboroff I (2002) The TREC 2002 Filtering Track Report. *TREC*, Vol. 2002 Vol. 2002, p 5.

[7] Tong A, Przybocki M, Margner V, El Abed H (2014) NIST 2013 Open Handwriting Recognition and Translation (Open HaRT'13) Evaluation. *Document Analysis Systems (DAS), 2014 11th IAPR International Workshop on* (IEEE), pp 81–85.

[8] Greenberg CS, et al. (2014) The NIST 2014 Speaker Recognition i-Vector Machine Learning Challenge. *Odyssey: The Speaker and Language Recognition Workshop*, pp 224–230.

[9] Over P, et al. (2014) TrecVid 2014–An Overview of the Goals, Tasks, Data, Evaluation Mechanisms and Metrics. *Proceedings of TRECVID*, p 52.

[10] Lucas SM, et al. (2003) ICDAR 2003 Robust Reading Competitions. *Document Analysis and Recognition, 2003. Proceedings. Eighth International Conference on* (IEEE), pp 80–84.

[11] Lucas SM (2005) ICDAR 2005 Text Locating Competition Results. *Document Analysis and Recognition, 2005. Proceedings. Eighth International Conference on* (IEEE), pp 80–84.

[12] Karatzas D, Mestre SR, Mas J, Nourbakhsh F, Roy PP (2011) ICDAR 2011 Robust Reading Competition-Challenge 1: Reading Text in Born-Digital Images (Web and Email). *Document Analysis and Recognition (ICDAR), 2011 International Conference on* (IEEE), pp 1485–1490.

[13] Shahab A, Shafait F, Dengel A (2011) ICDAR 2011 Robust Reading Competition Challenge 2: Reading Text in Scene Images. *Document Analysis and Recognition (ICDAR), 2011 International Conference on* (IEEE), pp 1491–1496.

[14] Karatzas D, et al. (2013) ICDAR 2013 Robust Reading Competition. *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on* (IEEE), pp 1484–1493.

[15] Karatzas D, et al. (2015) ICDAR 2015 Competition on Robust Reading. *Document Analysis and Recognition (ICDAR), 2015 13th International Conference on* (IEEE), pp 1156–1160.

[16] Veit A, Matera T, Neumann L, Matas J, Belongie S (2016) Coco-text: Dataset and Benchmark for Text Detection and Recognition in Natural Images. *CoRR* 1601.07140.

[17] Yao C, Bai X, Liu W, Ma Y, Tu Z (2012) Detecting Texts of Arbitrary Orientations in Natural Images. *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on* (IEEE), pp 1083–1090.

[18] Lee S, Cho MS, Jung K, Kim JH (2010) Scene Text Extraction with Edge Constraint and Text Collinearity. *Pattern Recognition (ICPR), 2010 20th International Conference on* (IEEE), pp 3983–3986.

[19] Wang K, Babenko B, Belongie S (2011) End-to-End Scene Text Recognition. *Computer Vision (ICCV), 2011 IEEE International Conference on* (IEEE), pp 1457–1464.

[20] Iwamura M, et al. (2016) Downtown Osaka Scene Text Dataset. *European Conference on Computer Vision* (Springer), pp 440–455.

[21] Jaderberg M, Simonyan K, Vedaldi A, Zisserman A (2014) Reading Text in the Wild with Convolutional Neural Networks. *CoRR* 11412.1842.

[22] Jaderberg M, Simonyan K, Vedaldi A, Zisserman A (2014) Synthetic Data and Artificial Neural Networks for Natural Scene Text Recognition. *CoRR* 1406.2227.

[23] Nascimento JC, Marques JS (2006) Performance Evaluation of Object Detection Algorithms for Video Surveillance. *IEEE Transactions on Multimedia* 8(4):761–774.

[24] Wolf C, Jolion JM (2006) Object Count/Area Graphs for the Evaluation of Object Detection and Segmentation Algorithms. *International Journal of Document Analysis and Recognition (IJDAR)* 8(4):280–296.

[25] Baumann A, et al. (2008) A Review and Comparison of Measures for Automatic Video Surveillance Systems. *EURASIP Journal on Image and Video Processing* 2008(1):1–30.

[26] Fiscus JG, Ajot J, Radde N, Laprun C (2006) Multiple Dimension Levenshtein Edit Distance Calculations for Evaluating Automatic Speech Recognition Systems During Simultaneous Speech. *The International Conference on Language Resources and Evaluation* (LERC), pp 1–8.