A first step toward a fair comparison of evaluation protocols for text detection algorithms

Aliona Dangla, Elodie Puybareau, Guillaume Tochon, Jonathan Fabrizio
EPITA Research and Development Laboratory (LRDE), France
jonathan.fabrizio@lrde.epita.fr

At a Glance

Problem. Can we trust evaluation protocols?

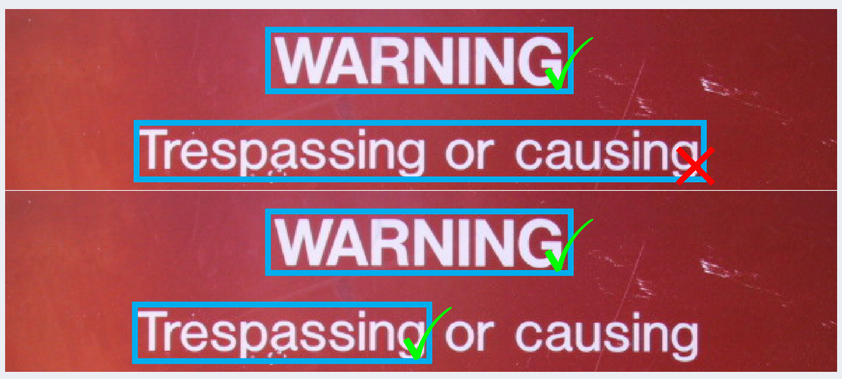
Objective. How can we validate/compare evaluation protocols?

Contribution. We propose a strategy to validate/compare evaluation protocols. We apply it on text detection evaluation protocols.

Can we trust evaluation protocols?

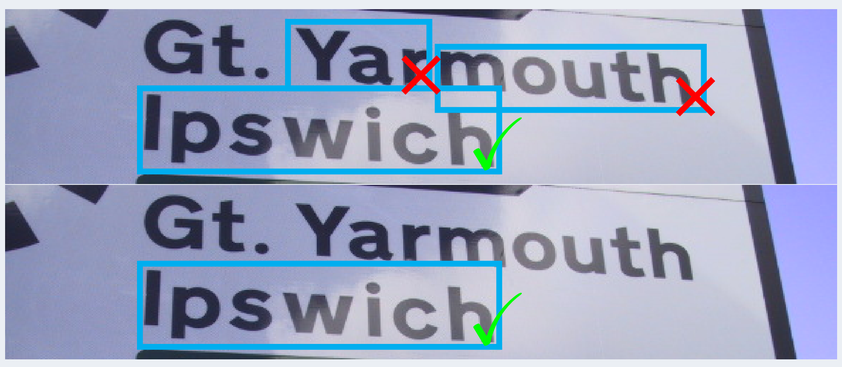
Detection results (✓ ⇒ validated by the evaluation protocol, × ⇒ rejected by the evaluation protocol)

Scores provided by the evaluation protocol:



Detection 1:
- Precision: 50%
- Recall: 25%

Detection 2:
- Precision: 100%
- Recall: 50%



Detection 1:
- Precision: 33%
- Recall: 33%

Detection 2:
- Precision: 100%
- Recall: 33%

Detection 1:
- Precision: 0%
- Recall: 0%

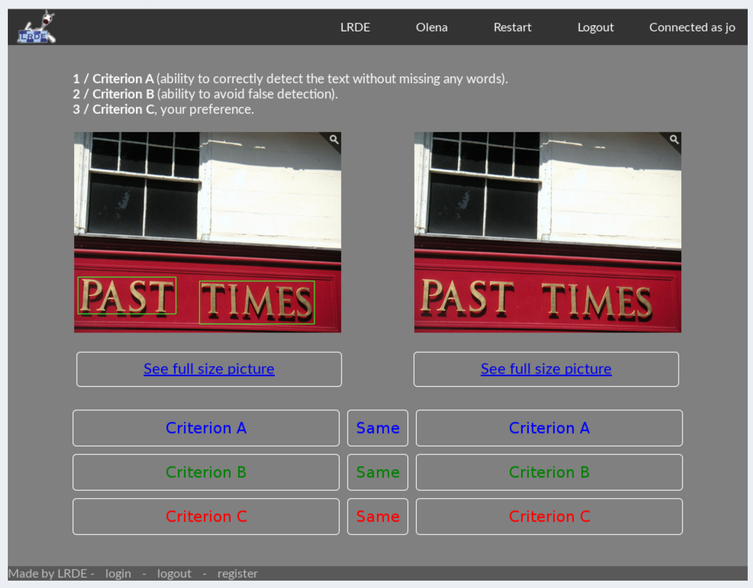
Detection 2:
- Precision: 10%
- Recall: 100%

Common evaluation protocols fail on simple and common situations.

Evaluation of evaluation protocols

How to evaluate evaluation protocols. To validate evaluation protocols, we have to compare evaluation protocol results with a reference. The reference must be made by human,
- but we can not ask people to score detection results,
- however we can ask people to compare and rank detection results.

Annotators can easily rank results thanks to a clever interface.



- Creation of a website to collect annotations.
- For each image of a dataset, the annotators sort the results. of many text detectors.

Ranks can be compared. Inspired by the Levenstein distance, we can score evaluation protocols on one image:

distance between
a = d > b > c > f > e
and
a > b > c > d > e > f

total cost: 3.5

a = d > b > c > f > e — swap: 1
a = d > b > c > e > f — split: 0.5
a > d > b > c > e > f — swap: 1
a > b > d > c > e > f — swap: 1
a > b > c > d > e > f

Cumulative results on the whole dataset provide an evaluation of the evaluation protocols.

Criteria. Text detection results are sorted three times according to:
1. the capacity of a method to correctly detect the text and not to miss it,
2. the capacity to precisely detect the text without false positives,
3. the overall preference of a detection.

Results

Results on ICDAR Robust Reading Challenge database that contains 233 images processed by 10 text detection methods.

Method	Best	Worst	Score
EvalTex EMD (no split)	**123**	**23**	**3.22**
EvalTex (no split)	96	41	3.58
DetEval	75	48	4.43
EvalTex EMD	81	52	4.57
ICDAR13	62	44	4.72
EvalTex	63	80	4.86
IOU	53	117	6.22

(a) Ranking w.r.t. recall

Method	Best	Worst	Score
EvalTex EMD (no split)	**105**	59	**8.36**
EvalTex EMD	**105**	59	**8.36**
IOU	100	**40**	8.70
ICDAR13	69	55	9.55
DetEval	70	57	9.56
EvalTex (no split)	20	143	12.09
EvalTex	20	143	12.09

(b) Ranking w.r.t. precision

Method	Best	Worst	Score
DetEval	**103**	**47**	**6.75**
ICDAR13	84	51	6.95
EvalTex EMD (no split)	81	69	7.77
EvalTex EMD	68	60	7.90
EvalTex (no split)	57	88	8.01
EvalTex	46	77	8.14
IOU	65	92	8.26

(c) Ranking w.r.t. preference

Best/Worst : number of times an EP gets the best/worst score (*i.e.*, has the smallest/largest distance with the ground truth) - *Score* : the mean of our Levenshtein-like distance over the whole database.

Conclusions

- Current evaluation protocols must be improved.
- Definition of evaluation criteria must be improved (such as precision).
- Too many common situations are not well handled.
- We propose a new way to evaluate evaluation protocols.
- We apply it on text detection evaluation.
- We have ranked evaluation protocols according to evaluation criteria.