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)
Le Kremlin-Bicêtre, France
Email: firstname.lastname@lrde.epita.fr

***Abstract*—Text detection is an important topic in pattern recognition, but evaluating the reliability of such detection algorithms is challenging. While many evaluation protocols have been developed for that purpose, they often show dissimilar behaviors when applied in the same context. As a consequence, their usage may lead to misinterpretations, potentially yielding erroneous comparisons between detection algorithms or their incorrect parameters tuning. This paper is a first attempt to derive a methodology to perform the comparison of evaluation protocols. We then apply it on five state-of-the-art protocols, and exhibit that there indeed exist inconsistencies among their evaluation criteria. Our aim here is not to rank the investigated evaluation protocols, but rather raising awareness in the community that we should carefully reconsider them in order to converge to their optimal usage.**

***Keywords*-Text detection; Evaluation protocol; Comparison**

I. Introduction

Text detection is an important pattern recognition task. Numerous challenges has been organized around this task such as imagEVAL06 and ICDAR Robust Reading (RR) challenges [1], [2], [3]. Several text detection algorithms have been developed [4] and have different performances depending on the context in which they are used. Many applications require an accurate text detection algorithm (automatic image and video indexing, blind people assistance system, etc.). It is necessary to accurately evaluate text detection algorithms in order to be able to choose and fine-tune the most suited one with respect to the concerned application. An error in the evaluation of the text detection can potentially lead to selecting an inappropriate method or a wrong parameterization. Therefore, the choice of the evaluation protocol is determinant to ensure the best choice of detection algorithm.

As many applications do not require reading, end-to-end recognition evaluation is not always relevant. Moreover, every step of the text extraction process must be evaluated independently. We thus limit ourselves to text detection evaluation (notice that the scheme of this article could be extended to many other contexts such as object detection evaluation). There exist many evaluation protocols (EPs) for text detection algorithms. The evaluation should depend on the application but, in practice, the community requires a standard EP to be able to compare detection methods. For text detection, EPs can be unfair in some situations, leading to wrong evaluations. Researchers are aware that any novel text detection algorithm must be evaluated rigorously. Most of the time, a comparison of this new method against state-of-the-art ones is unavoidable, as it is essential to assess the quality of the new research. However, there is no rigorous protocol to validate EPs, which are used as such by the community. Nevertheless, a weakness in the EP can lead to erroneous comparisons and conclusions.

In practice, we can notice that usual EPs fail in many common situations. In the evaluation of the detection presented in Fig. 1 for example, some EPs conclude that 66% of the text is well detected and 33% of the detection is a false positive. The provided score is counter intuitive as there is clearly more than 66% of text detected, without any false positive. The reason is that the size of the box around the word "parcel" is too small and the EP does not validate the detection, which is then considered as a false positive. This EP failure in a simple and very usual situation points out the necessity to finely analyze common EPs behaviors. It also questions the level of trustworthiness that should be granted to EPs.

It is thus important to evaluate these protocols and to determine their optimal conditions of use, in what situations they are relevant, what are their limits, etc. These characteristics can then be used by any user as guidelines on which EP should be used depending on the context, and how to correctly interpret the results of this EP. But how can EPs be characterized? This paper is a first attempt to derive a methodology to perform the comparison of EPs. We then apply it on various state-of-the-art EPs, and show that a single text detection scenario can lead to widely divergent interpretations, when looked through the prism of different EPs. Our aim here is not to rank the investigated EPs, but rather to raise awareness in the community that some extra attention should be paid to these tools, for which their "correct behavior" is often taken for granted by researchers.

II. Evaluation protocols for text detection

There are multiple ways to evaluate text detection algorithms and these evaluations rely on many performance measurements. The most common performance measurements are the precision P and the recall R. Denoting respectively by TP, FP and FN the number of true positives, false positives and false negatives, the recall R and the precision P are defined as:

$$R = TP/(TP+FN), \quad \text{and} \quad P = TP/(TP+FP). \quad (1)$$

* These authors contributed equally to this work.

DAS 2018: 13th IAPR International Workshop on Document Analysis Systems
DOI: 10.1109/DAS.2018.55
URL: https://doi.org/10.1109/DAS.2018.55

Figure 1: Example of text detection (green boxes) where a common flaw may occur for EPs: some EPs conclude that 66% of the text has been detected, 33% of the detection being a false positive.

The recall *R* evaluates the ability to detect the text (and not to miss any) and the precision *P* measures the ability not to *invent* text. They are usually combined by an harmonic mean to provide a final global score (called F-score).

While these performance measurements are computed in many different EPs, it is not an easy task to determine TP, FP and FN. As a matter of fact, the way these scores are evaluated varies from one protocol to the other, leading to wide discrepancies in the final score values. These performance measurements are computed by comparing the actual detection results against some ground truth (GT) data. As the detection and the GT generally do not exactly match, it is difficult to robustly estimate TP, FP or FN. The simplest way to evaluate text detection results is simply to compute the *Intersection over Union* (IOU) between the detection and the ground truth. The GT is a collection of N text areas GT_i. The result D is also a collection of M text areas D_j. To decide whether a GT_i has been detected, a detection D_j has to be found such as the ratio of the area they shared over the union of their surfaces is above a threshold s:

$$\exists j \in [\![1, M]\!],\ |D_j \cap GT_i| \,/\, |D_j \cup GT_i| \geqslant s, \quad (2)$$

and D_j is then considered as a TP. A D_j does not contribute to a detection if we have:

$$\forall i \in [\![1, N]\!],\ |D_j \cap GT_i| \,/\, |D_j \cup GT_i| < s, \quad (3)$$

and it is then considered as a FP. A FN is a GT_i that has not been detected; it satisfies:

$$\forall j \in [\![1, M]\!],\ |D_j \cap GT_i| \,/\, |D_j \cup GT_i| < s. \quad (4)$$

The most commonly used EP is certainly DETEVAL, which was proposed by Wolf and Jolion in [5]. DETEVAL is also a threshold-based method: it involves the computation of two values related to the precision (τ_{ij}) and the recall (σ_{ij}) for any GT_i and D_j. Those are then thresholded to achieve a decision. DETEVAL also comes with a smart visualization that allows the user to see the influence of each threshold value on the detection quality and to optimize their numerical values.

Another common EP (ICDAR13) has been derived from the latter and was used during *ICDAR RR 2013 challenge* [6]. All these EPs take a binary decision to consider a text as detected or not.

More recently EVALTEX [7] has been proposed. This tool is able to evaluate text detection at line, word or character levels. It is not restricted to horizontal and vertical rectangle bounding boxes [8]. Besides, it comes with a convenient visualization based on histograms to compare at a glance different results [9].

A new version has been proposed, EVALTEX EMD, based on a novel strategy. It consists in computing histogram of detected texts and computing the earth mover distance between this histogram and an ideal expected one. Compared to previous EPs, EVALTEX and its variant EVALTEX EMD do not take a binary decision: they take into account the proportion of correctly detected text. EVALTEX and EVALTEX EMD also penalize *R* when a word is fragmented into multiple detections. This penalty can be disabled, in which case the resulting score will be denoted by "EVALTEX no split".

While plenty other EPs exist (a more complete review can be found in [10]), we focus in this work on IOU, DETEVAL, ICDAR13, EVALTEX and EVALTEX EMD.

III. HOW TO ASSESS EVALUATION PROTOCOLS?

The assessment of EPs is an arduous task. Evaluation in text detection usually measures three characteristics: the recall R, the precision P and a final score based on them. One then has to check if these characteristics are well evaluated by EPs and represent the efficiency of the evaluated method.

We may now wonder how to evaluate the representativeness of the results of an EP? Usually, evaluations are conducted by comparing the output of the process to evaluate (P and R in our case) against some ground truth data. The creation of this latter is in practice a challenging exercise, often achieved by human annotators. In our case, it is unrealistic to expect humans to give reliable P and R scores for a specific detection. However, sorting the results of different text detection methods according to different criteria (the ability to detect text and the precision of the detection) is easily feasible by humans. Thus, instead of directly comparing scores, it becomes possible for each image to rank text detection methods according to the scores provided by EPs and compare them against the human annotators rankings. We finally compute for each EP an overall score from this comparison.

A. Collecting rankings from annotators

We created a website able to collect ranking from different annotators (Fig. 2). The website selects an image, and the user sorts 10 methods of text detection applied on this image. As the task is tedious, the process has been simplified: the website always asks for the comparison of only two detection results at the same time for a particular image. To get the complete ranking among all detection results for a given image, the process tries to ask as few comparisons as possible to reach the final ranking. Particularly, the process

- uses transitivity to deduce ranking between all results instead of asking comparison of each result against all others,
- uses a binary search (dichotomy) to find the position where to insert the new result in the final ranking,

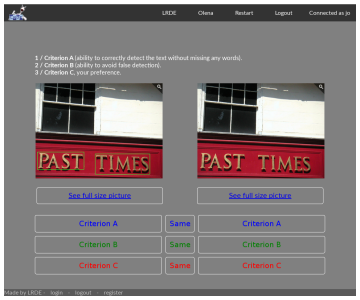

Figure 2: Ranking interface for annotators. The annotator sees two different results for the same image at the same time, and has to give his/her preferences with respect to three criteria.

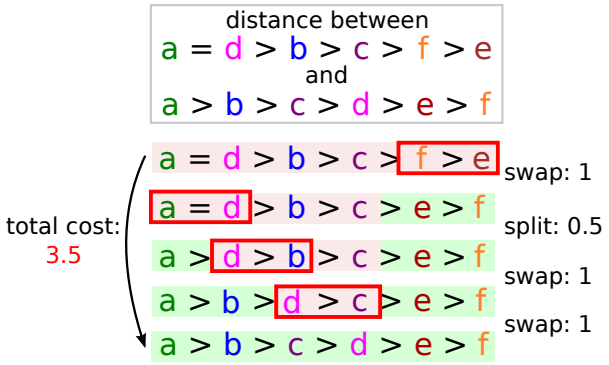

Figure 3: Illustration of the proposed measure for comparing rankings.

- treats only once (as a group) all results tagged as equal by the annotator.

The annotator is asked to rank the text detection algorithms according to three criteria: i) the capacity of a method to correctly detect the text and not to miss it, ii) the capacity to precisely detect the text without generating false positives and iii) a more subjective criterion which is the overall preference of a detection by the annotator. If the first two criteria are well described, the last one is voluntarily less precise to leave the user choose his favorite detection.

Each image of the database has been ranked at least by three different annotators. For one image, the ranking of the text detection algorithms can hence be different from one annotator to the other. A global rank of the algorithms for each image is deduced by computing a mean from every individual ranking. This global rank is considered as our ground truth for this image.

B. Comparing rankings provided by humans and ranking from evaluation protocols

To compare ranking of the ground truth and ranking from EPs, we get inspired by the Levenshtein distance. The Levenshtein distance between two strings counts the minimum number of simple character edits to change the first string into the other. Usual character edits are insertion, suppression, or even substitution. In our case, the idea is to form a string with the list of methods (represented by one symbol) sorted according to their rank. These usual edits are then not well adapted. We hence use, as edit, the swap of two consecutive elements (which costs 1 point). As some methods can have the same ranks, to be able to compare $a = b$ and $a > b$, we added the split of a set (remove rank equality) and the merge of a set (merge ranks). These operations cost 0.5 as this mistake is less serious than a swap. An example is given in Fig. 3.

IV. Evaluation of protocols

To evaluate EPs, we used the ICDAR 2013 RR challenge dataset [6]. This challenge comes with 233 images processed by 10 detection methods, so 10 results for each image of the dataset. Thanks to the great effort of the challenge organizers, scores (image by image) given by DetEval, ICDAR13 and IOU are also available, and their interface eased our work. The advantage of relying on these results is that they are well known by the community, and we do not favor any EP against the others as we directly consider provided scores. The size of the dataset is high enough to be representative, but still allows an annotation task. In addition, we use the tools provided by Calarasanu [10] and collect EvalTex and EvalTex EMD scores.

For our tests, the annotations (the ranking) has been done by around 10 people for a total of 680 ranks. All annotators are in the scope of image processing (but not necessarily in the document). In our results *Best* counts the number of times an EP gets the best score, i.e. has the smallest distance with the ground truth, *Worst* counts the number of times an EP get the worst score, i.e. has the biggest distance with the ground truth and *Score* is the mean of the Levenshtein distance over the whole database.

Recall R and Precision P

We compare, for each image, the ranking of the methods based on R and P computed by EPs and the ranking done by annotators. Based on our proposed measure, we rank EPs with respect to their scores. Those results are presented in Table Ia and Table Ib, respectively. Their analysis reveals that the performance behavior of EPs varies greatly depending on the investigated criterion (R or P). For instance, IOU performs poorly with respect to R as it ranks last. However, it is much more consistent in regards of P as it scores third overall. EvalTex (no split) has the opposite behavior, being second best for R but penultimate for P. On the other hand, some EPs have a consistent behavior with respect to both criteria: EvalTex yields the worst score for P and ranks second to last for R. Contrarily, EvalTex EMD (no split) ranks first for both criteria. Still, the lowest score in Table Ib being 8.36 highlights the fact that the P criterion has an ill-posed definition.

Preference

To determine the preference, we compare rankings based on the preference of the annotators and the ranking computed with the classically used F-score: $F_1 = 2\,PR\,/\,(P + R)$ [11]. The result of this comparison is given Table Ic. Although EvalTex EMD (no split) is ranked first with respect to R and P, it is ranked at the third place regarding the preference. On the other

Table I: Rankings.

(a) Ranking w.r.t. recall

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

(b) Ranking w.r.t. precision

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

(c) Ranking w.r.t. preference

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

(a) (b)

Figure 4: According to usual definition of *precision*: (b) is better than (a). According to *annotators*: the *precision* of (a) is better than (b).

hand, DetEval gets the first place in spite of the fact that it averagely performed in respect of *R* and *P*. These observations raise the question on the relevance of these indicators.

V. Discussion and Questioning

We discuss below several remarks arising from the analysis of Table I.

A. About Recall and Precision

Two classes emerge in the set of the considered EPs: the first one, comprising DetEval, ICDAR13, IOU, relies on a threshold to decide whether a detection is valid or not. The second one (all variants of EvalTex) provides a score proportional to the detection. Threshold-based methods can easily deal with margins, with respect to a certain limit. An example of this capacity is given in Fig. 5b: the margin around the detection may lead to wrong interpretation of the detection but these three EPs handles correctly this situation. On the contrary, EvalTex and EvalTex EMD under-estimate *R*. These EPs compute a penalty when the detection is split into multiple parts; the margin in the detection overlays surrounding text and the EP consider it as a split, leading to erroneous scores. The "no split" version of these EPs corrects this weakness. More issues arise when the boundary of the detection grows. Threshold-based EPs do not consider the margins as false detection, which is debatable. Over-detection on Fig. 5a is sufficiently large, however all tested threshold-based EPs provide a different *R* and *P* from 0% to 100%. DetEval correctly evaluates that the whole text has been detected thanks to good parametrization but fails to evaluate *P*, giving a 100% for *P*. ICDAR13 only validates the word "Roland", and fails in *P* too. IOU considers no valid detection at all. In that case, the binary decision shows significant limitations. EvalTex and EvalTex EMD do not suffer from these limitations and well handle it, penalizing only *P*.

Table II: Rankings with rounding.

Method	Recall 10^{-2}	Recall 10^{-3}	Recall 10^{-4}	Precision 10^{-2}	Precision 10^{-3}	Precision 10^{-4}
EvalTex EMD (no split)	2.92	3.22	3.22	8.01	8.36	8.36
EvalTex (no split)	3.06	3.53	3.58	11.52	12.01	12.09
EvalTex EMD	4.25	4.57	4.57	8.01	8.36	8.36
EvalTex	4.45	4.81	4.86	11.52	12.01	12.09
IOU	6.22	6.22	6.22	8.69	8.70	8.70
ICDAR13	4.72	4.72	4.72	9.54	9.55	9.55
DetEval	4.43	4.43	4.43	9.55	9.56	9.56

Undersized detections may also lead to erroneous evaluations: in Fig. 5c the word "St.Helena" is partially detected. DetEval and ICDAR13 do not validate the detection of this word, *R* is then under-estimated (50%). As a consequence, this detection is considered as a false positive, also leading to an under-estimated *P* (50%). Moreover, a detection that have missed the word "St.Helena" would be granted with a better score: *R* equal to 50% and *P* equal to 100%. This leads to a wrong ranking on a common situation. On the contrary IOU considers the detection as complete and then grant the detection with a 100% for the recall even if a part of the text is missing. In this case, IOU do not differentiate a method that successfully detect the word and a partial detection. EvalTex EMD and EvalTex correctly evaluate the detection but EvalTex has a noisy *P*.

The last point about *P* is the behaviour of EPs when an algorithm does not detect anything in an image. All methods grant *P* with 0% except EvalTex EMD (with or without penalty) which put 100% for *P*. A human annotator naturally favors an image without any detection rather than an image with a lot of false positives. Fig. 4 illustrates this point: how to rank these two methods according to *P*? EvalTex EMD matches more human expectation on this point. The definition of *P* must be revisited and formula 1 should be completed.

B. Influence of rounding

If threshold-based EPs are too coarse to allow precise comparisons, all variants of EvalTex allow much finer comparisons. However these EPs have a pixelwise accuracy and rank as "not equal" two detections having only one pixel different. This accuracy yields to unfair distinctions in ranking: in Fig. 5c, *P* given by EvalTex is 99.54% instead of 100%. In Fig. 5f, the text is correctly detected but the recall for EvalTex and EvalTex EMD are respectively 99.21% and 99.33%. This lack of "equalities" can penalize these methods in our evaluation. To measure this penalty, we wandered what was the influence of the rounding of the scores on the rankings. Results are shown in table II, in which we rounded them at 10^{-2}, 10^{-3} and 10^{-4}. The noticeable point here is that the scores are improved for

all variants of EVALTEX when the rounding is higher. IOU, ICDAR13 and DETEVAL remain stable. Indeed, as their decisions are binary, they allow less diversity in their scores. Thus, methods quite similar have the same rank. For all variant of EVALTEX, the rounding erases the small differences between two similar methods. This observation confirms that the pixelwise accuracy is too discriminant to be compared with human annotators, and raise a question about the level of precision the EPs are supposed to reach.

C. Preference

We can not find the ultimate way to evaluate all the detection methods as the evaluation is different according to the application. Some applications require a good P even if R is not so good while other necessitate the detection of all the text regardless false positives. However on the website, we asked the annotators to chose the detection they prefer, without giving them any indication. Without context, annotators tend to favor R instead of P. If we compare the ranking given by annotator for R and ranking given for the preference, we get a score equal to 11.89%. If we do the same with P we get a score equal to 5.57%. This tends to confirm our intuition even if the estimation is coarse. This observation can imply that the F-score is not well adapted metric for that purpose. We certainly should use a F_β-score, and find the optimal β, according to the way the annotators choose the preference. Now, if we compare the preference of the annotators and the ranking deduced from the F-score (Table Ic), DETEVAL reaches the first place even if it was not the best method neither for R nor P. The underlying reason remains an open question.

D. Importance and quantity of text

A possible improvement for EPs would be to take into account the size/the importance of the text areas. Hence, on Fig. 5d, all the EPs estimate that only $2/3$ of the text is detected - they consider each textbox equally - but annotators consider that more than $2/3$ of the text has been detected. A consideration of this criteria can help to obtain a fairest evaluation.

E. Granularity

It is difficult for EPs to manage a granularity difference between the detection and the GT. Threshold-based methods are more affected by this weakness. On Fig. 5e, ICDAR13, IOU and DETEVAL fail to validate the detection of the word "up" because the GT is at word level and the detection is at line level. This kind of evaluation favors the detection that have the same granularity as the ground truth and can have troubles to differentiate detection with a different granularity and wrong detection. Among them, DETEVAL (well parametrized) was less affected by this trouble, this can explain the difference for R among these methods. On the contrary EVALTEX and EVALTEX EMD, as they do not rely on a threshold, handle well this situation even if the recall is a bit affected.

As the ground truth is usually at word level, many people add an automatic process to cut detection - introducing many artifacts in the result of the algorithm. Other tend to modify the way the evaluation is done, leading potentially to unfair comparisons [12]. An EP must manage properly granularity difference to be fair and to allow correct comparisons between methods.

DETEVAL	100.00	100.00	100.00
ICDAR13	33.33	100.00	50.00
IOU	0.00	0.00	0.00
EVALTEX	100.00	60.06	75.05
ET-EMD	100.00	61.00	75.78

(a) The oversized detection is not well handled by most of the EPs.

IOU/DE/I13	100.00	100.00	100.00
EVALTEX	72.71	66.04	69.21
ET-EMD	73.33	66.67	69.84
ET-NS	100.00	66.04	79.55
ET-EMD-NS	100.00	66.67	80.00

(b) Margin around detection is difficult to be handled by EPs. The split penalty leads to wrong results

IOU	100.00	100.00	100.00
DE/I13	50.00	50.00	50.00
EVALTEX	85.02	99.54	91.71
ET-EMD	85.50	100.00	92.18

(c) Partial detection is difficult to be handle by EPs. Note that EVALTEX has a noisy P.

IOU/DE/I13	66.67	100.00	80.00
EVALTEX	66.67	93.36	77.79
ET-EMD	67.00	93.88	78.19

(d) No one takes into account the importance of the text for R.

IOU/DE/I13	80.00	100.00	88.89
EVALTEX	95.80	99.80	97.76
ET-EMD	95.80	99.80	97.76

(e) EPs have difficulties to handle granularity differences. EVALTEX and EVALTEX EMD better handle the line detection "Washing up" (the GT is word level).

EVALTEX	99.21	78.57	87.69
ET-EMD	99.33	79.00	88.01
IOU/DE/I13	83.33	83.33	83.33

(f) Oversized detection can disturb EPs: IOU/DE/I13 validate the word "Available" but not "while".

Figure 5: Some examples with scores: R (left), P (middle), F-score (right); the abbreviations DE, I13, and ET-EMD are respectively for DETEVAL, ICDAR13, and EVALTEX EMD. ET-NS and ET-EMD-NS name for "no split" versions.

VI. CONCLUSIONS

In this article, we have proposed a protocol to evaluate text detection EPs. This evaluation is based on the comparison of rankings provided by EPs against human rankings. To do so, we have introduced a strategy to compare rankings, able to handle ties. With this scheme,

we have evaluated several classically used EPs in the text detection literature, relying on various different strategies to guarantee representativeness. To ensure fairness, we made use of the database and the results of ICDAR RR 2013 challenge. The conducted evaluation pointed out the fact that EPs suffer from many weaknesses, and are not able to properly handle some common situations happening in text detection. In addition, in various cases, they are not able to provide a fine comparison between two different (but rankable) detection results.

This comparison is the opportunity to start a reflection on the reliability of EPs:

- the *R* and *P* criteria have several hidden drawbacks, as their definitions are intrinsically ill-posed. In particular, the definition of *P* is too vague to be relevant. In the case where there is no detection for example, *P* usually scores 0%, but it is then impossible to distinguish between a method that detects nothing and a method that only gives wrong detections;
- overall, there is a lack of separate criteria allowing to characterize the relevance of any detection (split, overlap...). *P* and *R* are not sufficient to characterize these situations and can not handle them yet;
- threshold-based methods provide binary decisions only. It is thus impossible to finely compare two different results. In addition, finding the optimal setting of operated thresholds is an arduous task;
- granularity differences (one-to-many, many-to-one and many-to-mane) between GT and detection are usually not well handled by EPs.

We should deal with these points in depth to improve our evaluation schemes. There are also still many unanswered questions such as "Should all texts or all false positives contribute equally?". We notice that human annotators act differently according to the size of the false positives for example. Furthermore, for annotators, missing a text is most of the time more serious than having a false positive. One can wonder if the final global score provided by EPs should take care about this. It is important for the users of the protocol to be aware of these weaknesses to adapt the protocol according to the final application of the evaluated method.

Many other things should be evaluated. EVALTEX provides also two others interesting indicators but we have not studied them in this comparison: the quality and the quantity. These two indicators allow to differentiate, for *R* equal to 50% for example, whether only half of the text has been detected or all text detected half.

This comparison is then a starting point and there is still a lot of work. We should validate and improve our protocol of evaluation as it is (too) far from perfect. A weak point of our evaluation is the number of annotators. We should increase the number of annotators to improve the representativeness of the results (and also evaluate the disparity among annotators). A second weak point is that only ranking is taken into account and not the actual score provided by the EPs. We have to refine the scoring by taking into account the actual value of *P* and *R*. Notice also that our way of scoring favors EPs that generate equal ranks. We could also go deeper in the analysis: we should understand why the ranking of the preference is so different than the ranking of *P* and *R*.

We hope that with these elements, we will be able to give new clues to develop a robust universal EP.

ACKNOWLEDGMENT

The authors would like to thank Timothée Evain (Télécom ParisTech, Univ. Paris-Saclay) for his help, and Joseph Chazalon (L3I, Univ. La Rochelle) and Thierry Géraud (LRDE, EPITA) for their valuable comments.

REFERENCES

[1] D. Karatzas, L. Gomez-Bigorda *et al.*, "ICDAR 2015 competition on robust reading," in *IAPR Intl. Conf. on Document Analysis and Recognition*, 2015, pp. 1156–1160.

[2] N. Nayef, F. Yin *et al.*, "ICDAR2017 robust reading challenge on multi-lingual scene text detection and script identification – RRC-MLT," in *IAPR Intl. Conf. on Document Analysis and Recognition*, 2017, pp. 1454–1459.

[3] R. Gomez, B. Shi *et al.*, "ICDAR2017 robust reading challenge on COCO-text," in *IAPR Intl. Conf. on Document Analysis and Recognition*, 2017, pp. 1435–1443.

[4] Q. Ye and D. S. Doermann, "Text detection and recognition in imagery: A survey," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 37, pp. 1480–1500, 2015.

[5] C. Wolf and J.-M. Jolion, "Object count/area graphs for the evaluation of object detection and segmentation algorithms," *International Journal on Document Analysis and Recognition*, vol. 8, no. 4, pp. 280–296, 2006.

[6] D. Karatzas, F. Shafait *et al.*, "ICDAR 2013 robust reading competition," in *IAPR Intl. Conf. on Document Analysis and Recognition*, 2013, pp. 1484–1493.

[7] S. Calarasanu, J. Fabrizio, and S. Dubuisson, "What is a good evaluation protocol for text localization systems? Concerns, arguments, comparisons and solutions," *Image and Vision Computing*, vol. 46, pp. 1–17, 2016.

[8] ——, "From text detection to text segmentation: A unified evaluation scheme," in *Computer Vision – ECCV 2016 Workshops*, ser. LNCS, vol. 9913. Springer, 2016, pp. 378–394.

[9] ——, "Using histogram representation and earth mover's distance as an evaluation tool for text detection," in *IAPR Intl. Conf. on Document Analysis and Recognition*, 2015, pp. 221–225.

[10] S. Calarasanu, "Improvement of a text detection chain and the proposition of a new evaluation protocol for text detection algorithms," Ph.D. dissertation, Université Pierre et Marie Curie – Paris 6, France, Dec. 2015.

[11] D. Karatzas, F. Shafait, S. Uchida, M. Iwamura, L. G. Bigorda, S. R. Mestre, J. Mas, D. F. Mota, J. A. Almazan, and L. P. de las Heras, "ICDAR 2013 robust reading competition," in *IAPR International Conference on Document Analysis and Recognition*, 2013, pp. 1484–1493.

[12] Y. Du and G. Duan, "Context-based text detection in natural scenes," in *International Conference on Image Processing*, 2012, pp. 1857–1860.