Using histogram representation and Earth Mover's Distance as an evaluation tool for text detection

Stefania Calarasanu
EPITA Research and Development
Laboratory (LRDE)
F-94276, Le Kremlin Bicêtre, France
Email: calarasanu@lrde.epita.fr

Jonathan Fabrizio
EPITA Research and Development
Laboratory (LRDE)
F-94276, Le Kremlin Bicêtre, France
Email: jonathan.fabrizio@lrde.epita.fr

Séverine Dubuisson
Sorbonne Universités, UPMC Univ Paris 06
CNRS, UMR 7222, F-75005, Paris, France
Email: severine.dubuisson@isir.upmc.fr

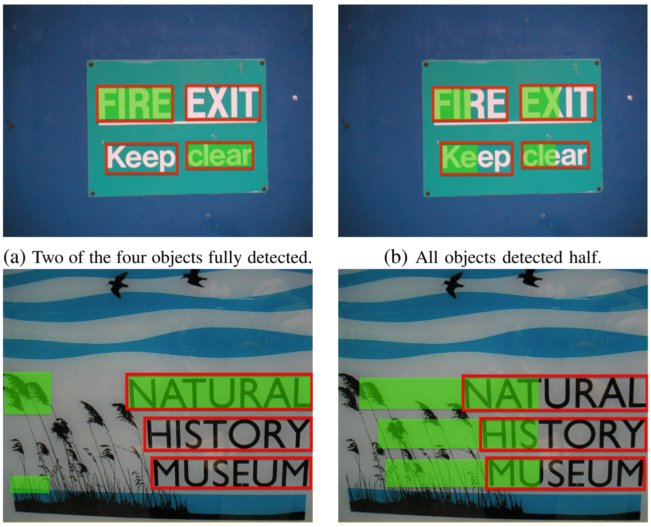
***Abstract*—In the context of text detection evaluation, it is essential to use protocols that are capable of describing both the quality and the quantity aspects of detection results. In this paper we propose a novel visual representation and evaluation tool that captures the whole nature of a detector by using histograms. First, two histograms (coverage and accuracy) are generated to visualize the different characteristics of a detector. Secondly, we compare these two histograms to a so called optimal one to compute representative and comparable scores. To do so, we introduce the usage of the Earth Mover's Distance as a reliable evaluation tool to estimate recall and precision scores. Results obtained on the ICDAR 2013 dataset show that this method intuitively characterizes the accuracy of a text detector and gives at a glance various useful characteristics of the analyzed algorithm.**

I. Introduction

Text detection applications have become very popular in the last years. Due to the growing number of approaches in the literature, the need of relying on viable ranking systems has increased considerably. Evaluation protocols are essential tools for researchers to compare their results with those provided by state-of-the art methods but also to quantify the possible improvements of their text detectors. During text detection evaluations, the output results are compared to a ground truth (GT) throughout a matching procedure. Final scores are then generated using performance metrics. Recall and precision metrics are generally used in the literature due to their ability to characterize different aspects of a detection: recall represents the proportion of detected texts in the GT, whereas precision describes the proportion of accurate detections with respect to the GT.

However, these two indicators, individually, do not provide sufficient information about a detection. As first stated by Wolf and Jolion in [1], it is important to differentiate the quantity aspect of a detection ("how many GT objects/false alarms have been detected?") from its quality aspect ("how accurate is the detection of the objects?"). Fig. 1 illustrates the importance of this distinction when using these two metrics. One can observe that the same recall and precision scores (computed with the evaluation protocol of Sec. II) can correspond to different detection outputs. Intuitively, it is then hard to correctly interpret a detection through one value, hence the need to highlight separately the quantity and quality characteristics.

To globally evaluate a detection at a dataset level (an image or multiple images), one first needs to evaluate each detection individually (at object level). This could consist in assigning quality measurements to each GT object and detection with respect to some predefined matching rules. For example, in Fig. 1b all four GT objects have been detected with a quality score of 0.5, corresponding to the coverage area.



(a) Two of the four objects fully detected. (b) All objects detected half.

(c) One of the three objects fully detected. Two false positives. (d) All objects detected one third.

Fig. 1. Four examples illustrating the GT objects with red rectangles and the detections with green plain rectangles: (a)-(b) two examples for which recall $R = 0.5$; (c)-(d) two examples for which precision $P = 0.33$.

Once the quality scores are produced at object level, it is necessary to quantify them to characterize the detection at dataset level. In the literature, the general way to quantify these object level scores consists in averaging them. This mean is computed depending, either on the number of objects [2], [3], or on the total object area [4]. In [5], Hua *et al.* proposed to compute an overall detection rate by averaging the detection qualities of all GT text boxes with respect to the sum of their detection importance levels. However, none of these methods provides a visual representation of the detection evaluation. Conversely, Wolf and Jolion proposed performance graphs in [1] to illustrate the quality and quantity detection nature of an algorithm. The method generates two graphs by varying two quality area constraints (for recall and precision) over a wide range of values. The area under the curve (AUC) obtained by varying these constraints is then used to represent the overall recall and precision measures. This is equivalent to averaging the sum of all object level measurements computed over all

IEEE ICDAR 2015
DOI: 10.1109/ICDAR.2015.7333756
http://ieeexplore.ieee.org/document/7333756/

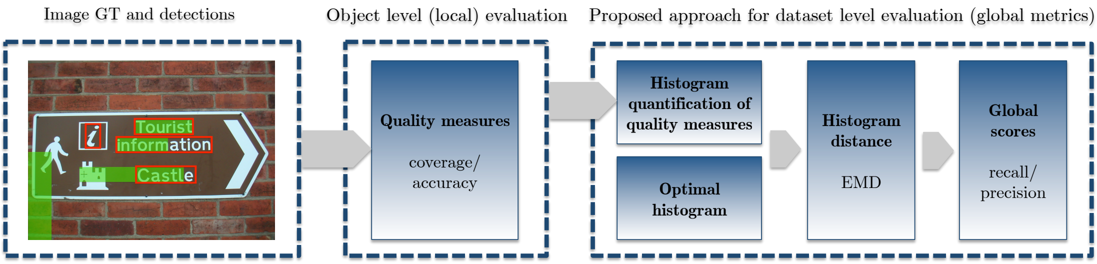


Fig. 2. Workflow of the proposed approach; left image: four GT objects (red rectangles) and 4 detections (green plain rectangles).

possible constraint values. It is not always sufficient to only consider global recall and precision scores when comparing two detectors. For example, one might be interested to know which algorithm produced more false positives (Figs. 1c, 1d), or which one detected more GT objects entirely, instead of only partially (Figs. 1a, 1b). This kind of information cannot be retrieved by only interpreting the global scores.

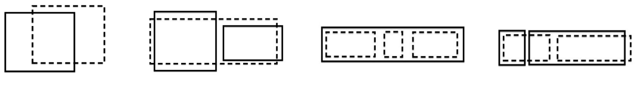
The contribution of this paper is twofold and comes as a smart alternative solution to the existing methods: first, we propose a new way to visually represent text detection results using histograms, that can, as we will show, capture both their quality and quantity aspects; secondly, we use a histogram distance to derive global recall and precision scores. For that, we rely on a "local" evaluation that produces quality scores at object level. However, this local evaluation is not in the scope of this paper, since the quality features might vary depending on the targeted characteristics of a detection. This makes our approach independent of the used quality features.

The organization of this paper is as follows. In Sec. II we give an overview of the "local" evaluation (quality measurements) used to evaluate the text detection outputs. We then describe in Sec. III our proposed approach which introduces the histogram representation of text detections and histogram distance derived metrics to evaluate a full text detection output. Results and experiments are presented in Sec. IV. Finally, concluding remarks and perspectives are given in Sec. V.

II. Introduction on Quality Measurements

In this section, we describe the procedure used to compute qualitative scores at object level. As mentioned before, this protocol is an independent module and could be replaced by any other quality evaluation protocol.

Consider $G = (G_1, G_2, ..., G_m)$ a set of GT text boxes and $D = (D_1, D_2, ..., D_n)$ a set of detection boxes, with m and n the number of objects in G, resp. in D. Based on the nature of a detection, we can identify four types of matchings, as illustrated in Fig. 3: (a) one-to-one: one text box in G matches one text box in D; (b) one-to-many: one text box in G matches multiple text boxes in D; (c) many-to-one: multiple text boxes in G match one text box in D; (d) many-to-many: conditions (b) and (c) are simultaneously satisfied. We compute a coverage (the capacity to detect text) and an accuracy (the precision of the detection) score for each GT object separately, based on the matching type, as it can be seen in the following subsections.



(a) One-to-one (b) One-to-many (c) Many-to-one (d) Many-to-many

Fig. 3. Matching cases (GT is represented by dashed rectangles and detections by plain line rectangles).

a) One-to-one match: For each GT-detection pair of objects (G_i, D_j) involved in a one-to-one match, we define the coverage Cov_i, and the accuracy Acc_i, based on the true overlap area between the two objects:

$$Cov_i = \frac{Area(G_i \bigcap D_j)}{Area(G_i)}, \qquad Acc_i = \frac{Area(G_i \bigcap D_j)}{Area(D_j)}. \quad (1)$$

b) One-to-many match: During one-to-many matches, the coverage and accuracy scores are given by:

$$Cov_i = \frac{\bigcup_{j=1}^{s_i} Area(G_i \bigcap D_j)}{Area(G_i)} \cdot F_i, \quad Acc_i = \frac{\bigcup_{j=1}^{s_i} Area(G_i \bigcap D_j)}{\bigcup_{j=1}^{s_i} Area(D_j)}$$

where F_i is a fragmentation penalty, defined in [4] by $F_i = \frac{1}{1+\ln(s_i)}$, and s_i is the number of detections associated to G_i.

c) Many-to-one match: This corresponds to "many" one-to-one cases. To compute the accuracy rate for each GT G_i, we first assign it a detection area. We split the detection area between its corresponding m_j (merge level of the detection box D_j) GT objects, with respect to their areas. We define $TextArea_{D_j} = Area(\bigcup_{i=1}^{m_j}(G_i \bigcap D_j))$ as the area resulting from the union of all intersections between the GT text boxes and the detection box, and $nonTextArea_{D_j} = Area(D_j) - TextArea_{D_j}$ as the detection area excluding $TextArea_{D_j}$. Hence, the coverage and accuracy for each GT text box G_i, $i \in [1, m_j]$, are defined by:

$$Cov_i = \frac{Area(G_i \bigcap D_j)}{Area(G_i)}, \quad Acc_i = \frac{Area(G_i \bigcap D_j)}{Area(D_{j,i})}, \quad (2)$$

where $Area(D_{j,i}) = \frac{Area(G_i)}{TextArea_{D_j}} \cdot nonTextArea_{D_j}$ represents the corresponding detection area for each G_i.

d) Many-to-many match: This occurs when one-to-many and many-to-one detections overlap the same GT objects and are evaluated accordingly to coverage and accuracy equations given in Sec. II-b and II-c.

III. Proposed Approach

In this paper we propose to use histograms as an efficient way to represent and evaluate a detection. Because histograms are graphical representations of frequency distributions over a set of data, they can be also seen as convenient tools to represent simultaneously the quality and quantity aspects of a detection. Here, the quality aspect is described by the histogram's bin intervals, while the detection quantity feature

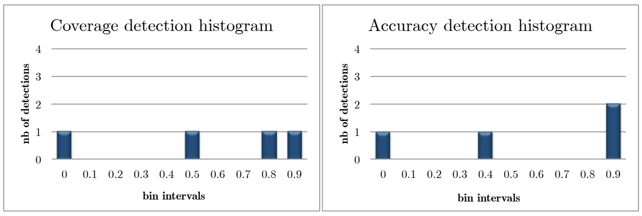


Fig. 4. Corresponding non-normalized coverage and accuracy detection histograms (respectively h_{Cov} and h_{Acc}) for the example in Figure 2.

is represented by the bin values. The quality results, given by the protocol described in Sec. II, are then quantified and represented throughout two detection histograms for coverage and accuracy values. Finally, global recall and precision scores are generated by computing the distances between the two detection histograms and an "optimal" histogram. The overview of the proposed method is illustrated in Fig. 2.

A. Histogram representation

Consider a 1D finite valued function f that contains values $f(j) \in [0,1]$, $j = 1, \ldots, n$. Its quantified histogram into B intervals (bins) is a 1D numerical function h defined by $h(b) = n_b$ where n_b is the number of values of f that belong to interval $[\frac{b}{B}, \frac{b+1}{B}[$, for $b = 0, \ldots, B-2$ and $[\frac{b}{B}, \frac{b+1}{B}]$ for $b = B-1$.

The evaluation protocol described in Sec. II provides two sets: f_{Cov} for coverage scores and f_{Acc} for accuracy scores. These sets can then be described by quantified histograms h_{Cov} and h_{Acc}, corresponding to coverage and accuracy histograms. The detection example in Fig. 2 (left) illustrates the case of four GT objects ("*i*", "Tourist", "information" and "Castle") and four detections, among which, one is a false positive. In this example, using the protocol described in Sec. II we get the coverage scores $\{0.0, 0.55, 0.8, 1.0\}$ and the accuracy scores $\{0.0, 0.45, 1.0, 1.0\}$. Their representation using histograms with $B = 10$ bins is given in Fig. 4. This representation consists in quantifying all coverage and accuracy scores obtained by an algorithm in all images of a dataset (in our example: 1 image and 4 detections). Next, we will consider normalized histograms, so that $\sum_{b=0}^{B-1} h(b) = 1$ (histograms of Fig. 4 are not normalized). We will call these histograms $\widetilde{h_{Cov}}$ and $\widetilde{h_{Acc}}$.

The histogram representation provides both a quantitative (*i.e.* values of bins) and a qualitative (*i.e.* number of bins) representation of the detection. A perfect algorithm should get maximal accuracy and coverage values for all detections, *e.g.* their corresponding histogram representation should have only one populated bin, the last one (for example, for $B = 10$, with all values belonging to $[0.9, 1]$). This histogram is referred to as the *optimal histogram*. We then propose to measure a detector's performance as the distance between $\widetilde{h_{Cov}}$ (and $\widetilde{h_{Acc}}$) and the optimal histogram. We describe the way we measure this distance in the next section.

B. Global metrics generation throughout histogram distances

Although histograms can be seen as powerful tools for characterizing the whole nature of a detection, their representation does not immediately conduct to an overall performance measurement. This can be achieved by computing a distance between histograms: the lower the distance, the higher the similarity between the histograms.

Let $\widetilde{h_O}$ be the normalized optimal histogram, whose all bins except the last one are empty. We then have:

$$\widetilde{h_O}(b) = \begin{cases} 1 & \text{if } b = B-1 \\ 0 & \text{otherwise} \end{cases} \quad \forall b \in [0, B-1] \tag{3}$$

By computing the distance between $\widetilde{h_{Cov}}$ and $\widetilde{h_{Acc}}$ and the optimal histogram $\widetilde{h_O}$ we get two global detection performance measures (recall and precision). There are two main families of distances between histograms [6]. Bin-by-bin distances only consider bin content (or size) and often make a linear combination of similarities measured between same bins of the two considered histograms (for example, the Euclidean distance). This assumes histograms are aligned and have the same size. Cross-bin distances also consider the topology of histograms by integrating into the computation the distance between bins.

Taking into account the topology of histograms is very important in our case. For example, if we consider the case where all bins of $\widetilde{h_{Cov}}$ but one are empty (same reasoning for $\widetilde{h_{Acc}}$), then the Euclidean distance between $\widetilde{h_{Cov}}$ and $\widetilde{h_O}$ will give the value 0 if bin $\widetilde{h_{Cov}}(B-1) = 1$ (case of a perfect match), 1 otherwise (any case where $\widetilde{h_{Cov}}(b) = 1, b \neq B-1$). However, we would like the distance to be lower when the only populated bin of $\widetilde{h_{Cov}}$ is close to the last bin $B-1$, because this corresponds to better recall scores on all the database. That is why it is required to both consider the bin content and the distance between bins (as a kind of relationship between bins). Hence, a cross-bin distance is a better choice for computing the histogram dissimilarity in the given context. Although many cross-bin distances were proposed in the literature (see [7] for a review), we have chosen to use the Earth Mover's Distance (EMD) for two reasons: it captures the perceptual dissimilarity better than other cross-bin distances [8]; and it can be used as a true metric [8]. A brief description of the EMD is given in the next paragraph.

The EMD, first introduced by Rubner *et al.* [8], is a cross-bin distance function that computes the dissimilarity between two signatures. Let $P = \{(p_i, w_{p_i})\}_{i=1}^{m}$ and $Q = \{(q_j, w_{q_j})\}_{j=1}^{n}$ be two signatures of sizes m and n, where p_i and q_j represent the position of ith, respectively jth element and w_{p_i} and w_{q_j} their weight. The EMD searches for a flow $F = [f_{ij}]$ between p_i and q_j, that minimizes the cost to transform P into Q:

$$COST(P, Q, F) = \sum_{i=1}^{m} \sum_{j=1}^{n} d_{ij} f_{ij}, \tag{4}$$

where d_{ij} is the ground distance between clusters p_i and q_j; the cost minimization is done under the following constraints:

$$f_{ij} \geq 0, \quad \sum_{j=1}^{n} f_{ij} \leq w_{p_i}, \quad \sum_{i=1}^{m} f_{ij} \leq w_{q_j}, \quad i \in [1, m],\ j \in [1, n]$$

$$\sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} = min(\sum_{i=1}^{m} w_{p_i}, \sum_{j=1}^{n} w_{q_j}), \quad i \in [1, m], \quad j \in [1, n]$$

The EMD distance is then defined as:

$$EMD(P, Q) = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} d_{ij} f_{ij}}{\sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij}} \tag{5}$$

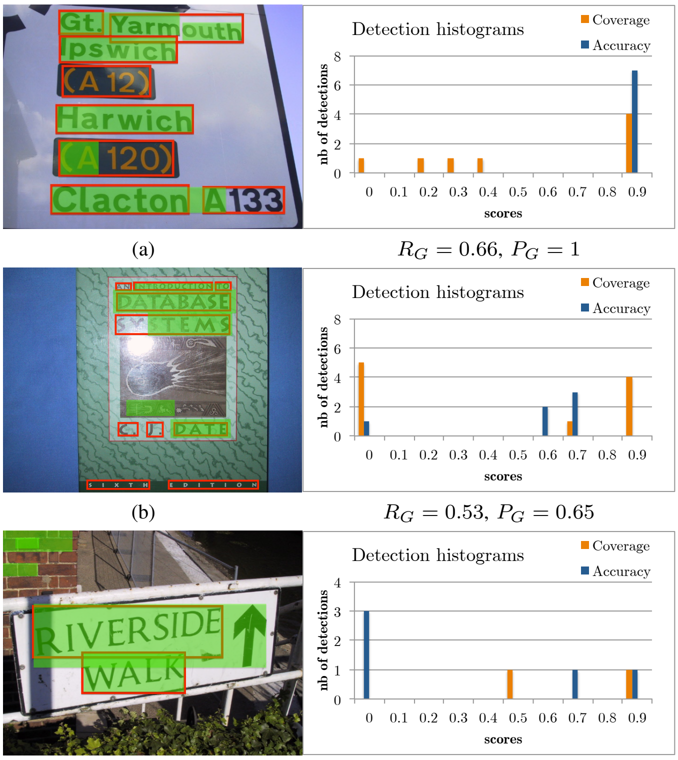
Rubner *et. al* proved in [8] that when the ground distance is a metric and the total weights of the two signatures are equal, the EMD is a true metric. Therefore, by considering d as the Euclidean distance and $\widetilde{h_{Cov}}$ and $\widetilde{h_{Acc}}$ as signatures [9], we can use the EMD as a valid dissimilarity measure. In such cases, a bin is a cluster (p and q) and its value is a weight (w). For example, if we consider the right histogram of Fig. 4b after its normalization, then its corresponding signature is $\{(0, 0.25), (0.1, 0), (0.2, 0), (0.3, 0.25), (0.4, 0), (0.5, 0), (0.6, 0), (0.7, 0), (0.8, 0), (0.9, 0.5)\}$.

We then derive the two global similarity metrics [10], recall R_G and precision P_G:

$$R_G = 1 - EMD(\widetilde{h_{Cov}}, \widetilde{h_O}) \tag{6}$$
$$P_G = 1 - EMD(\widetilde{h_{Acc}}, \widetilde{h_O}) \tag{7}$$

IV. Results and Discussion

The dataset used during our experiments is the one proposed during the ICDAR 2013 Robust Reading (*Challenge 2*) competition [11]. It contains 233 images of natural scene texts and a word level annotation. Fig. 5 illustrates three examples of detections, their corresponding non-normalized coverage and accuracy histograms with $B = 10$ bins and the resulting global recall and precision scores. The interpretation of these two histograms is straightforward. For example, the first bin of h_{Cov} (orange) encloses the total number of non-detected (or poorly detected, coverage ≤ 0.1) GT objects, while the first bin of h_{Acc} (blue) encloses the number of false positives (or detections with poor precision, accuracy ≤ 0.1). In Fig. 5a, the scattered coverage values of h_{Cov} indicate the presence of either partial ("A120" ($[0.3, 0.4[$) and "A133" ($[0.2, 0.3[$)) or one-to-many ("Yarmouth" ($[0.4, 0.5[$)) detections. On the other hand, all accuracy values are accumulated into the last bin of h_{Acc} which means that all detections were truthful with respect to the GT. By analyzing the histograms of Fig. 5b, we observe that the first bin value of h_{Cov} equals the sum of values of the other bins. This shows that only half of the GT objects were detected ("INTRODUCTION", "TO", "DATABASE", "SYSTEMS", "DATE"), while the other half was missed or poorly detected ("AN", "C.", "J.", "SIXTH", "EDITION"). h_{Acc} of Fig. 5c, suggests there are three possible false positives. The values 1 of bin intervals $[0.7, 0.8[$ and $[0.9, 1]$ correspond to one detection that exceeds its corresponding GT boundary object ("RIVERSIDE") and one accurate detection ("WALK") respectively. More results are given here [1].



(a) $R_G = 0.66,\ P_G = 1$

(b) $R_G = 0.53,\ P_G = 0.65$

(c) $R_G = 0.8,\ P_G = 0.42$

Fig. 5. Three examples of GT (red rectangles) and detections (green plain rectangles) and their corresponding coverage/accuracy histograms (resp. h_{Cov} (orange) and h_{Acc} (blue)) and R_G/P_G scores.

A. Comparison of two algorithms

A good advantage of this representation is that, used on a dataset, it allows to characterize and compare at a glance text detectors. In Fig. 6 we illustrate the overall detection behavior of two algorithms, *detector 1* (Inkam) and *detector 2* (TextSpotter), based on the detection results submitted to ICDAR 2013 Robust Reading competition [11]. The left plot shows coverage values ($\widetilde{h_{Cov}}$) of both algorithms. Both coverage normalized histograms illustrate a similar tendency: two high peaks on the first and last bins and a lower peak around the value 0.5. This means that, for both algorithms, most of the GT objects were either missed, either accurately detected, while only approximately 6% of the GT objects were involved in partial or one-to-many detections. One can however conclude that from the coverage aspect, *detector 2* slightly outperforms *detector 1*: the number of missed GT objects (value of the first bin) is lower while the last bin's value is higher. This is confirmed by R_G scores (see caption in Fig. 6). The right plot shows accuracy values of both algorithms. Contrary to the coverage similarity behavior discussed above, the accuracy profiles of the two detectors are very different. *detector 1* produces a significantly higher number of false positives than *detector 2*. The accuracy histogram of *detector 2* has higher bin values in the quality intervals $[0.7, 0.8[$ and $[0.8, 0.9[$. This is because *detector 2* adds a large border to all its detections [11], which decreases the object-level accuracies. On the other hand, *detector 1* produces as many false positives as accurate detections (first and last bin values close to 0.4). The corresponding P_G scores, given in the caption of Fig. 6, confirm that *detector 2* outperforms *detector 1* by about 20%.

We now compare our histogram representation with the performance plots generated with *DetEval* tool [1] (see Fig. 7). The representation in [1] is obtained by varying two quality constraints for each measure (recall and precision) and counting how many objects fall into a certain interval, whereas our method implies a qualitative local evaluation from the beginning. Although both approaches capture the quality and quantity natures of a detection, we introduce a more compact representation using only two plots for depicting a detection (instead of generating four plots, two for recall and two for

[1] www.lrde.epita.fr/~calarasanu/ICDAR2015/supplementary_material.pdf

precision, as proposed in [1]). Secondly, histograms have the advantage of being more intuitive and easier to interpret in the given context of text detection. One can easily visualize the proportion of missed GT objects or false positives, as well as the amount of detections that fall into any other coverage or accuracy interval. Concerning the overall recall and precision scores obtained with the two approaches, we can observe that the results are different, which is due to the different object level evaluation used by the two methods. However, both sets of scores follow the same tendency and hence confirm the ranking in which *detector 2* outperforms both in recall and precision, the performance of *detector 1*.

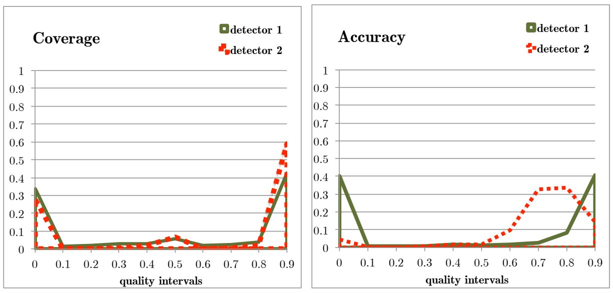
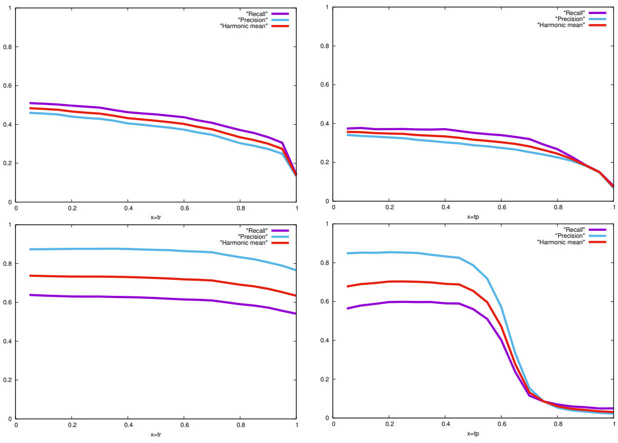


Fig. 6. Coverage and accuracy normalized histograms associated to *detector 1* ($R_G = 0.60$, $P_G = 0.58$) and *detector 2* ($R_G = 0.70$, $P_G = 0.80$).



varying constraint t_r varying constraint t_p

Fig. 7. Performance plots generated with *DetEval* tool [1] (recall in purple, precision in blue); top: *detector 1* ($R_{OV} = 0.37$, $P_{OV} = 0.32$); bottom: *detector 2* ($R_{OV} = 0.49$, $P_{OV} = 0.69$).

B. Impact of tuning the number of bins

By using histograms to represent detections, the generated global scores will depend on the chosen number of bins (B). While a value of 10 bins is mostly appropriate for graphical illustration purposes, when computing final scores, one should however choose a higher number of bins to produce a more precise evaluation result. Fig. 8 illustrates the variation of R_G and P_G scores when B varies from 10 to 100 bins. The natural tendency of these two metrics is to decrease when B increases. When B exceeds 50 intervals, one can observe the stabilization of these two global scores.

V. Conclusion

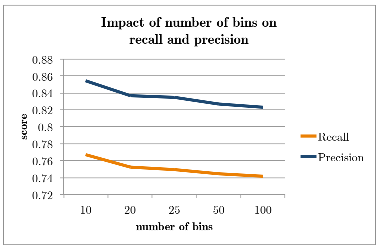


Fig. 8. Variation of R_G and P_G scores depending on the number of bins B (detection results provided by [12] on the ICDAR 2013 dataset).

In this article we have presented a new approach for visually representing and evaluating text detection results using histograms. It consists of firstly generating detection histograms based on a "local" evaluation and secondly, employing the Earth Mover's Distance as a reliable evaluation tool for computing global scores. In this paper, we used coverage and accuracy features to illustrate the quality nature of a detection. Depending on the targeted detection characteristics, other quality features can be equally exploited (e.g. fragmentation feature derived from one-to-many detections). As described in Sec. IV, the histogram dynamics permits to intuitively observe both the quality and the quantity aspects of a detection. Compared to other methods, the proposed approach offers a compact graphical visualization, a clear understanding of a detector's output, an easier comparison between different detection behaviors at precise quality intervals and finally a powerful similarity measure, based on the cross-bin Earth Mover's Distance, used to compute global detection scores.

Acknowledgment

This work was partially supported by ANR 12-ASTR-0019-01 "Describe".

References

[1] C. Wolf and J.-M. Jolion, "Object count/area graphs for the evaluation of object detection and segmentation algorithms," *IJDAR*, vol. 8, no. 4, pp. 280–296, 2006.

[2] Y. Ma, C. Wang, B. Xiao, and R. Dai, "Usage-oriented performance evaluation for text localization algorithms," in *ICDAR*, 2007, pp. 1033–1037.

[3] A. Clavelli, D. Karatzas, and J. Llados, "A framework for the assessment of text extraction algorithms on complex color images," in *DAS*, 2010, pp. 19–26.

[4] V. Mariano, J. Min, J.-H. Park, R. Kasturi, D. Mihalcik, H. Li, D. Doermann, and T. Drayer, "Performance evaluation of object detection algorithms," in *ICPR*, 2002, pp. 965–969.

[5] X.-S. Hua, L. Wenyin, and H.-J. Zhang, "Automatic performance evaluation for video text detection," in *ICDAR*, 2001, pp. 545–550.

[6] S. Dubuisson, "Tree-structured image difference for fast histogram and distance between histograms computation," *PRL*, vol. 32, no. 3, pp. 411–422, 2011.

[7] W. Yan, Q. Wang, Q. Liu, H. Lu, and S. Ma, "Topology-Preserved Diffusion Distance for Histogram Comparison." *BMVC*, pp. 1–10, 2007.

[8] Y. Rubner, C. Tomasi, and L. Guibas, "The earth mover's distance as a metric for image retrieval," *IJCV*, vol. 40, no. 2, pp. 99–121, 2000.

[9] H. Ling and K. Okada, "Emd-l1: An efficient and robust algorithm for comparing histogram-based descriptors," in *ECCV*, 2006, pp. 330–343.

[10] X. Wan, "A novel document similarity measure based on earth movers distance," *Information Sciences*, vol. 177, no. 18, pp. 3718 – 3730, 2007.

[11] ICDAR, "Robust reading competition results," http://dag.cvc.uab.es/icdar2013competition/, 2013.

[12] J. Fabrizio, B. Marcotegui, and M. Cord, "Text detection in street level image," *PAA*, vol. 16, no. 4, pp. 519–533, 2013.