Image and Vision Computing 46 (2016) 1–17

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Review article

What is a good evaluation protocol for text localization systems? Concerns, arguments, comparisons and solutions☆

Stefania Calarasanu[a,*], Jonathan Fabrizio[a], Severine Dubuisson[b]

[a] EPITA Research and Development Laboratory (LRDE) 14-16, rue Voltaire, F-94276 Le Kremlin Bicêtre, France
[b] Sorbonne Universités, UPMC Univ Paris 06, CNRS, UMR 7222, ISIR, F-75005 Paris, France

ARTICLE INFO

Article history:
Received 15 July 2015
Received in revised form 9 October 2015
Accepted 3 December 2015
Available online 4 January 2016

Keywords:
Evaluation protocol
Text detection

ABSTRACT

A trustworthy protocol is essential to evaluate a text detection algorithm in order to, first measure its efficiency and adjust its parameters and, second to compare its performances with those of other algorithms. However, current protocols do not give precise enough evaluations because they use coarse evaluation metrics, and deal with inconsistent matchings between the output of detection algorithms and the ground truth, both often limited to rectangular shapes. In this paper, we propose a new evaluation protocol, named EvaLTex, that solves some of the current problems associated with classical metrics and matching strategies. Our system deals with different kinds of annotations and detection shapes. It also considers different kinds of granularity between detections and ground truth objects and hence provides more realistic and accurate evaluation measures. We use this protocol to evaluate text detection algorithms and highlight some key examples that show that the provided scores are more relevant than those of currently used evaluation protocols.

Contents

☆ This paper has been recommended for acceptance by Cheng-Lin Liu, PhD.

* Corresponding author.
E-mail address: calarasanu@lrde.epita.fr (S. Calarasanu).

http://dx.doi.org/10.1016/j.imavis.2015.12.001
0262-8856/

1. Introduction

Text detection is an important task in image processing, and many algorithms have been proposed since the last two decades [1]. Hence, text detection systems require a reliable evaluation scheme that provides a ground truth (GT) as precise as possible and a protocol that can evaluate the precision and the accuracy of a text detector with regard to this GT. A solid evaluation protocol should also be able to fairly compare different algorithms. A text detection algorithm can be evaluated differently depending on its output, that can be either boxes surrounding the detected texts, or masks of detected texts after their binarization. One can also directly evaluate the output of an O.C.R.: in such case, the detection algorithm integrates a recognition module and provides as output the text transcription, which is then compared to the true text.

While the output provided by the O.C.R. seems to be the ultimate way to evaluate text detection algorithms, the computed scores do not always correctly reflect the detection accuracy: the transcription can fail because of the distortions of the detected text or its fonts. Furthermore, text transcription is not always necessary, especially in applications for which only the text detection is needed (such as text enhancement or license plate blurring). The evaluation of a text mask is a difficult task as well, mainly because it requires the *true* binarization of the text, that can vary depending on the text properties (stroke thickness for example). Here again, the evaluation does not focus on the detection results but evaluates both the detection and the binarization (in practice, this binarization is also not necessarily needed).

The simplest and most common way to evaluate a text detection algorithm is then to compare its detection bounding boxes to those that have been manually annotated (*i.e.* from the GT). This is the common strategy used in most text detection challenges (ImageEval, ICDAR) to evaluate and compare algorithms. However, we have noticed that these evaluation protocols are not reliable. This is due, both to the metrics used for the evaluation, and to the GT annotations [2,3], that can lead to irrelevant evaluation and comparison of text detection algorithms.

An annotation is sometimes subjective, and therefore it can be difficult to choose how text should be annotated [2]. It is yet possible to construct a dataset only composed of images in which there is no ambiguity for the annotation. However, there is still the problem of tilted or curved texts for which a bounding rectangular box is not appropriate because it can contain a lot of non-text areas. It is then important to define rules for labeling and defining the granularity, *i.e.* the minimal text entity to include into a bounding box. Different levels of granularity can be defined for the GT annotation, depending on the text to detect: the *line*, *word* and *character* levels. The line level is not well suited for tilted text. The character level provides a tedious annotation and promotes connected component approaches. The best granularity level seems to be the word level, even if it is still not the best choice for multi-oriented text.

Choosing good metrics to compare detections that do not correctly match the GT objects is also a complex task. Most of the metrics can not efficiently deal with the difference of granularity levels between the GT and the detections. For example, if the GT is at word level and the detection at line level, the score will be most of the time over-penalized. Moreover, as pointed by Wolf and Jolion in [4], a single metric cannot truly describe the complex behavior of a localization algorithm, namely separating the quantity nature ("*how many GT boxes were detected*") from the quality aspect ("*how well the GT boxes were detected*") of a detection. Although these issues were addressed in the literature (see Section 2), the proposed solutions are still not satisfactory.

Because of all these limitations, researchers do not have any robust tool to get a representative evaluation of their algorithm and a fair comparison with other algorithms. For example, the authors in [5] claim that their scores are too low because the ICDAR2013 protocol does not correctly evaluate line level detections. Hence, some other works that provide detections at line level [6,7] have proposed to change the GT annotation of ICDAR2005 dataset from word to line level to be less penalized. However, this does not permit a correct comparison with other scores obtained using the same database with the word level annotation. Sun et al. [8] manually split their line level detections in order to use the ICDAR2013 protocol and compare their results. Manual splitting is also a problem because it makes the comparison irrelevant with other detectors integrating an automatic splitting step (or ever no splitting). Du et al. [9] have also split their line level detections into words, however, no detail about the splitting procedure is given. Due to the lack of a fair evaluation protocol, many works [10,11] evaluate their algorithm by using others protocols. However, this gives an inconsistent comparison to other algorithms.

Only few interest has been given to the evaluation protocol of text detection algorithms. Some works [12,13,14] do not mention at all what protocols are used for the evaluation, while others [15,16,17,18,19,20,21,22] limited their explanations to "*standard recall, precision and F-Score*" without any further details concerning their computation or matching strategies. DetEval is probably the most frequently used evaluation protocol. Its framework is tunable and hence its configuration should always be specified when used. However, only few works [23,24] specify the used parameters, while many do not mention them [25,26,27,28,29,30]. All these examples prove a need of revising the current evaluation protocols.

In this article, we propose a new evaluation protocol providing many advantages compared to the most common used, listed below.

- It can handle different detection granularities. For that, we propose a two-level rectangular GT annotation, which allows an equitable comparison between algorithms having different granularity outputs.
- It provides a clear identification of the matching strategy between a GT object and a detection (one-to-one, one-to-many, many-to-one and many-to-many cases) and adapts the two quality metrics (coverage and accuracy) to each type of matching.
- It computes both quantity and quality recall and precision scores to give a full comprehension of a detector's behavior.
- It can be easily adapted to manage any irregular text representation, such as polygonal, elliptic or free-form ones.

This article is organized as follows. Section 2 first gives a short survey of the existing metrics and evaluation protocols for text detection algorithm evaluation and comparison. Section 3 presents our evaluation procedure called EvaLTex. We first define our two-level annotation that permits to deal with different detector's output granularities (Section 3.1). Then we detail our matching procedures to avoid over or under penalizations while matching detections and ground truth objects (Section 3.2). We also propose a generalization of our protocol to evaluate a set of images and derive quality and quantity scores for the detection (Section 3.3). Finally, we show how EvaLTex can also manage free form annotations (Section 3.4). Section 4 is dedicated to the validation of our evaluation framework in the context of text detection and its comparison to other evaluation protocols. In particular, we show that the currently used evaluation protocols can not efficiently manage many detection scenarios and that our method provides more logical scores. Finally, concluding remarks and perspectives are given in Section 5.