Performance Evaluation Tools for Zone Segmentation and Classification (PETS)

Wontaek Seo, Mudit Agrawal and David Doermann
Institute of Advanced Computer Studies
University of Maryland
College Park, MD, USA
{wtseo, mudit, doermann}@umd.edu

Abstract

This paper overviews a set of Performance Evaluation Tools (PETS) for zone segmentation and classification. The tools allow researchers and developers to evaluate, optimize and compare their algorithms by providing a variety of quantitative performance metrics. The evaluation of segmentation quality is based on the pixel-based overlaps between two sets of regions proposed by Randriamasy and Vincent [12]. PETS extends the approach by providing a set of metrics for overlap analysis, RLE and polygonal representation of regions and introduces type-matching to evaluate zone classification. The software is available for research use.

1. Introduction

Document image layout analysis is a fundamental component of modern document analysis systems which aims at producing a hierarchical representation of the document embedding its geometric structure. The recent advances in search technology and decreases in storage costs has rekindled the interest in digitization of various diverse documents. These heterogeneous collections of documents are posing new challenges in document layout analysis. With the various geometric layout analysis algorithms such as zone segmentation [10, 2], text-line detection [7, 5] and zone classification [11], it is becoming increasingly important to have a quantitative evaluation metric to compare, evaluate and analyze the various approaches against different datasets.

Evaluations for text-line detection and zone segmentation algorithms suggest overlap comparisons while those for zone classification require some type or label comparisons. One of the early methods to evaluate zone segmentation, however, was based on text-based metric by analyzing the errors in recognized text after page segmentation. Kanai et. al. [4, 3] proposed a metric that is a weighted sum of the number of edit operations (insertions, deletions and moves) required to convert the generated text string into the ground-truth string. Though this technique requires only ASCII text ground-truth, it could not specify the error location in the image and was dependent on the OCR engine’s availability and accuracy. To overcome these limitations, Mao and Kanungo [9] proposed a text-line based zone segmentation scheme which uses the percentage of ground-truth text-lines contained correctly within result zones without split, merge or miss errors [10]. The drawback of this approach, however, is that in case of zone segmentation, if the segmentation algorithm outputs the whole page as one zone, the split and missed errors would disappear. As shown in Figure 1, result zones containing complete text-lines from different zones are not penalized. Region-based evaluation schemes were proposed by Vincent et. al. [12, 13] and Haralick et. al. [6]. According to this approach, a document is represented as a hierarchy of layout structure and content. The evaluation is performed at different levels of the hierarchy and the metrics are based on area overlaps of segmentation and ground-truth zones. These metrics can evaluate both textual and non-textual regions based on the set of valued pixels they contain. Another difficulty with these approaches

Figure 1: Result zone covering two distinct zones is not penalized using line-based evaluation
is the precise definition of the distance metrics for varied segmentation scenarios like line detection, zone segmentation and noise detection is very rigid.

We propose a region-based evaluation scheme which builds on Randriamasy and Vincent [12] and uses the structure and content of zones to establish a match. It extends their approach by providing algorithm-specific set of metrics for overlap analysis, RLE and polygonal representation of regions and introduces type-matching to evaluate zone classification. In Section 2 we give an overview of PETS capabilities. This is followed by the implementation details of zone matching, the merge capability to analyze substitution errors and the evaluation schemes in Section 3 and conclusions and future work in Section 4.

2. PETS Overview

The evaluation of various detection, segmentation and zone-classification algorithms like page segmentation, line-detection, stamp and logo detection, noise detection and removal, ruled-line removal and zone classification work on a basic premise of overlap detection of ground-truth zone (-type) and result zone (-type). However, each of these algorithms require a different evaluation strategy for detailed analysis.

2.1 Zone Classification

Evaluation of zone classification algorithms require consistent zone types assuming a pre-established correspondence. This is the simplest evaluation scheme where result zone is penalized only when its recognized type is not equivalent to the ground-truth zone type and does not consider location.

PETS provides an option of zone-filtering where only the regions of interest are evaluated. This helps in evaluating a subset of region-types when the datasets are significantly unbalanced [1].

2.2 Zone Segmentation

Segmentation evaluation is typically the most arguable step in a document processing pipeline. The primary reason is that there may be no deterministic and consistent way to ground-truth a page into zones especially for complex documents. While it is evident that two different style zones should always be returned as distinct zones, splitting a text-zone along the direction of text, for example at paragraph or line breaks, is often acceptable. In order to avoid this confusion, accuracy of page segmentation algorithms has been calculated as the percentage of ground-truth text-lines contained correctly within result zones without split, merge or miss errors [10, 9]. The drawback of this approach has been shown in Figure 1. Our zone-based evaluation method is a much stricter evaluation scheme in terms of zone detection. However, as stated earlier, this method will not tolerate any valid over-segmentations (at paragraph or line breaks).

PETS allows a more sophisticated evaluation scheme in which result zones belonging to the same ground-truth zone can be merged using an optional merge option. While it remains a zone-based evaluation strategy, valid (along-text) zone segments can be merged to avoid penalty. This merge-option is configurable and evaluation tool can be used with both merge (lenient) and no-merge (stricter) options.

2.3 Zone Matching

Zone matching is the default mode of PETS evaluation. It works like zone segmentation with both merge and no-merge options, but requires an additional constraint of zone-type (zone classification) to be matched before a merge or an overlap can be established.

2.4 Zone Detection

Line, stamp-logo or noise detection falls under the category of zone detection of particular types of content. Ground-truthing all regions in a document is not expected and any result zones overlapping with pixels not marked as regions under detection, should be reported as false alarms.

One challenge that should be dealt with is when there is a difference in scope between the ground truth and the results. For example, a dataset may only be annotated with the content the user is interested in evaluating. A collection may be annotated with content regions, but leave all other background (noise, etc) unmarked. This is fine for detection, where we expect identified regions to be of a certain type, but is not necessarily appropriate for segmentation. A segmentation algorithm which does not try to identify content, but rather segments pages into different zones (Figure 3b), will be penalized for identifying segments with noise as different regions. To address this problem, PETS provides an optional ignore or don’t care state so that result zones that are unmatched (or matched against the background) will not be used in the evaluation.

3. Implementation

3.1 Input

PETS requires three sets of files as inputs: image, ground-truth and result files. The ground-truth and result files follow the GEDI XML format specification [8]. GEDI is a public domain ground-truth editor and document interface for scanned text documents [14]. Its
interface maintains a one to one correspondence with XML files and the corresponding image files. Different types of zones can be created and visualized using a custom set of ‘attributes’. All segmentation and classification algorithms should produce results in GEDI XML format in order to be evaluated using PETS.

3.2 Output

As a result of evaluation, PETS produces an evaluation XML file for each trio of image, ground-truth and result file and an overall result file for the set of input data.

1. Each evaluation xml file marks false, missed, detected and/or matched zones. This XML file can also be visualized in GEDI along with the corresponding image file to analyze the segmentation or classification algorithm under study. Figure 3b shows GEDI visualization of PETS evaluation file on Voronoi++ segmentation algorithm [2].

2. The overall result file contains the matching scores of all zones, confusion matrices and a summarized result with precision, recall and F1 scores.

3.3 Metrics

Zones are represented either as rectangular (with an optional orientation attribute) or polygonal. In case of overlapping zones, pixels can be associated with a zone using run-length encoding. The proposed PETS evaluation metric uses pixel-based overlaps to construct correspondences between result and ground-truth zones. Since pixel-based overlap computation is an expensive operation, especially when zones are polygonal, a bounding box overlap is computed first. If the zones satisfy the bounding box overlap criterion (overlap above a user specified threshold), pixel-based overlap is computed. Let $G = \{g_1, g_2, ..., g_n\}$ and $R = \{r_1, r_2, ..., r_m\}$ be a set of ground-truth and result zones respectively. Given a pixel $p$ in overlapping zones $g_i$ and $r_j$, the metrics are calculated as follows:

$$\text{TruePositive}(TP) = \{p | p \in g_i \land p \in r_j\} \quad (1)$$

$$\text{FalsePositive}(FP) = \{p | p \not\in g_i \land p \in r_j\} \quad (2)$$

$$\text{FalseNegative}(FN) = \{p | p \in g_i \land p \not\in r_j\} \quad (3)$$

$$\text{Precision}(p) = \frac{|TP|}{|TP| + |FP|} \quad (4)$$

$$\text{Recall}(a) = \frac{|TP|}{|TP| + |FN|} \quad (5)$$

$$F1 = 2 \times \text{Precision} \times \text{Recall} / (\text{Precision} + \text{Recall}) \quad (6)$$

These metrics are used to construct matching score tables (MST) between the ground-truth and result zones. Regions which have a set of MST values above user-specified thresholds are said to have overlapped, otherwise the values are reset to zero. This is used to associate correspondences between each region of one set to the region(s) of the other set based on the four primary modes of overlaps: one-to-one, one-to-many, many-to-one and many-to-many. The overlaps are shown in Figure 2.

3.4 Options

PETS evaluates various algorithms depending on the usage options described below:

1. Type Matching: PETS reads a zone-type attribute ($\psi$) for each ground-truth and result zone from their respective xml files. Assuming zone correspondence is set either through their ids or overlaps, type-matching can hence be performed ($\psi_r = \psi_g$).

2. Detection: A ground-truth zone $g_i$ is said to be detected by a result zone $r_j$ only if its F1 score in the corresponding MST cell is above a user-specified threshold.

$$F1_{i,j} > \tau_f \quad (7)$$

In case of multi-match, the result zone $r$ with best F1 score is chosen for each ground-truth zone to establish correspondence.

$$r = \max_{j = 1}^{J} F1_{i,j} \quad (8)$$

Figure 3 shows the evaluation result of a document image on our voronoi based segmentation algorithm [2] using GEDI.
algorithms against one of the best algorithms using zone segmentation, line detection and noise removal.

We evaluated our zone classification, toolkit for all types of zone segmentation and classification scenarios. We evaluated our zone classification, evaluation being performed, making PETS a suitable environment for document embedding its geometric structure.

4. Conclusions and Future Work

We have presented performance evaluation tool for zone detection and classification algorithms called PETS. PETS is being researched for use by the document analysis research community. PETS is a polygonal region based matching tool based on one-to-one, one-to-many, many-to-one and many-to-many overlaps. It uses a pixel-based scheme to establish correspondence between a set of regions. Result zones can be merged, filtered or ignored based on the type of evaluation being performed, making PETS a suitable toolkit for all types of zone segmentation and classification scenarios. We evaluated our zone classification, zone segmentation, line detection and noise removal algorithms against one of the best algorithms using

PETS and showed how PETS aids in analyzing various perspectives of evaluation. Currently PETS uses a flat-zone structure and does not embed hierarchy of zones. This is our next goal which will enable PETS to evaluate the complete hierarchical representation of a document embedding its geometric structure.

References


