

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 describes a set of Performance Evaluation Tools (PETS) for document image 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 zones proposed by Randriamasy and Vincent [1]. PETS extends the approach by providing a set of metrics for overlap analysis, RLE and polygonal representation of zones and introduces type-matching to evaluate zone classification. The software is available for research use.

Keywords—evaluation document image segmentation detection classification

I. INTRODUCTION

Recent advances in search technology and decreases in storage costs has rekindled the interest in the digitization of various diverse documents. These heterogeneous collections of documents are posing new challenges in document layout analysis, a fundamental component of modern document analysis systems that attempts to produce a hierarchical representation of a document's geometric structure. Furthermore with competing algorithms for tasks such as zone segmentation [2], [3], text-line detection [4], [5] and zone classification [6], it is becoming increasingly important to have quantitative evaluation metrics to (a) evaluate the state of the art, (b) measure progress in algorithm development and (c) judge applicability of a given algorithm to a defined task.

Evaluations for text-line detection and zone segmentation algorithms suggest spatial comparisons while those for zone classification require type or label comparisons. One of the early methods to evaluate zone segmentation, however, was based on a text-only metric which analyzed errors in recognized text after page segmentation. Kanai et. al. [7], [8] 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. Since this technique requires only ASCII text ground-truth, it can not specify the error location in the image and is 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 [3]. 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.

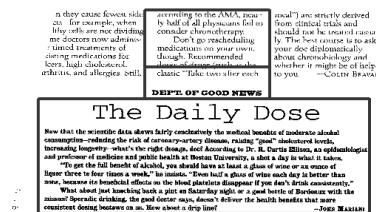


Figure 1. Result zone covering two distinct zones is not penalized using line-based evaluation

Zone-based evaluation schemes were also proposed by Vincent et. al. [1], [10] and Haralick et. al. [11]. With these approaches, a document is represented as a hierarchy of layout structure and content. Evaluation is performed at different levels of the hierarchy and the metrics are based on area overlap of segmentation and ground-truth zones. These metrics can evaluate both textual and non-textual zones based on the set of pixels they contain. One difficulty with these approaches is the precise definition of the metrics for varied segmentation scenarios like line detection, zone segmentation and noise detection is very rigid.

We have developed a set of Performance Evaluation Tools (PETS) for zone segmentation and classification algorithms using a zone-based evaluation scheme which builds on Randriamasy and Vincent [1] and uses the structure and content of zones to establish a match. It extends their approach by providing an algorithm-specific set of metrics for overlap analysis, RLE and polygonal representation of zones and introduces type-matching to evaluate zone classification. In Section 2

we give an overview of PETS capabilities for zone classification, segmentation, matching and detection. This is followed by a description of implementation details and configurable options in Section 3 and conclusions and future work in Section 4.

II. PETS OVERVIEW

Evaluation of detection, segmentation and zone-classification algorithms typically require ground-truth to provide spatial boundaries with optional labels. Correctness can be based on overlap of ground-truth and result zones, label correspondence or both. PETS allows algorithm-specific metrics for detailed evaluation.

In order to allow PETS to adapt to the various needs of its users, we distinguish between four cases of evaluation: 1) classification - where the goal of the algorithm is to label existing spatial zones, 2) segmentation - where the goal is to evaluate the correctness of the spatial partitioning produced by the algorithm, 3) matching - which includes both classification and segmentation, and 4) detection - which requires only the identification and localization of specific elements on the page, not necessarily a complete partitioning.

One challenge that must be dealt with is when there is a difference in ‘scope’ between the ground truth and the results. For example, a dataset may be annotated with only the content the user is interested in, such as content zones, but leave all other background (noise, etc) unmarked. This is fine for *detection*, where we expect identified zones 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 zones. 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.

A. Classification

Zone classification algorithms assume a pre-established spatial correspondence. Evaluation penalizes only when the result type is not equivalent to the ground-truth zone type. It does not consider location.

PETS provides an option of zone-filtering where only the zones of a particular type or types are evaluated. This helps in evaluating subsets of zone-types when datasets are significantly unbalanced [12].

B. Segmentation

Segmentation evaluation is an important but controversial 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 may be evident that two different style zones should be returned as distinct zones, splitting a text-zone along the direction of text, for example, at paragraph or line breaks, is arguably acceptable. In order to avoid this confusion, the accuracy of page segmentation algorithms is often calculated as the percentage of ground-truth text-lines contained correctly within result zones without split, merge or miss errors [3], [9]. The drawback of this approach was illustrated in Figure 1.

PETS evaluates zone segmentation algorithms based on ground-truth and result zones overlap. In order to tolerate any *valid* over-segmentations (at paragraph or line breaks), PETS also allows result zones belonging to the same ground-truth zone to be merged using a *merge* option. While it remains a zone-overlap 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.

C. Matching

Zone matching evaluation, the default mode of PETS, validates both the spatial overlaps (segmentation) and the label correspondence (classification) between ground-truth and result zones. With *merge* option, it requires an additional constraint of zone-type to be matched before a merge can be established.

D. Detection

Line, stamp-logo or noise identification fall under the category of zone detection. Detection is essentially matching with *ignore* option turned off and zones of interest selected through zone-filtering. Ground-truthing all zones in a document is not expected and any result zone overlapping with pixels not marked or of different type, is reported as a false alarm.

III. IMPLEMENTATION

A. Input

PETS requires three sets of files as inputs: images, ground-truth files and result files. The ground-truth and result files follow the GEDI XML format specification [13] produced by GEDI, 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 and different types of zones can be created and visualized using a custom set of ‘attributes’. All segmentation and classification algorithms must produce results in GEDI XML format in order to be evaluated using PETS.

B. Output

PETS produces an evaluation XML file for each image/ground-truth/result file evaluated and an overall summary file for the set of images.

- 1) The evaluation XML file identifies false, missed, detected and/or matched zones and can be visualized in GEDI along with the corresponding image file to analyze the segmentation or classification results.
- 2) The summary file contains the matching scores of all zones, confusion matrices and a summarized result with precision, recall and F1 scores.

C. Metrics

Ground-truth and result zones are represented either as rectangles (with an optional orientation attribute) or polygons. In either case, pixels are associated with the underlying 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 criteria (overlap above a user specified threshold), pixel-based overlap is computed.

Let $G = \{g_1, g_2, \dots, g_n\}$ and $R = \{r_1, r_2, \dots, r_m\}$ be a set of ground-truth and result zones respectively. Given a pixel p in overlapping zones g_i and r_j , the following values can be calculated for a given pair of zones:

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

$$\text{FalsePositive}(FP) = \{p \mid p \notin g_i \wedge p \in r_j\} \quad (2)$$

$$\text{FalseNegative}(FN) = \{p \mid p \in g_i \wedge p \notin r_j\} \quad (3)$$

$$\text{Precision}(\rho) = |TP| / (|TP| + |FP|) \quad (4)$$

$$\text{Recall}(\sigma) = |TP| / (|TP| + |FN|) \quad (5)$$

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

These metrics are used to construct matching score tables (MSTs) between the ground-truth and result zones. A set of MST values, above user-specified thresholds, are used to establish correspondences between the ground-truth and result zones based on the four primary modes of overlap: one-to-one, one-to-many, many-to-one and many-to-many as shown in Figure 2.

A list M , containing pairs of unions of matching zones $\{Z_g, Z_r\}$ ($= \{ \cup g_i, \cup r_j \}$) is obtained from the MST. Z_g and Z_r represent subsets of matched ground-truth and result zones respectively. Comparing the cardinality of each union, following can be defined:

- 1) Missed: Z_g has no corresponding Z_r
- 2) False Alarm: Z_r has no corresponding Z_g

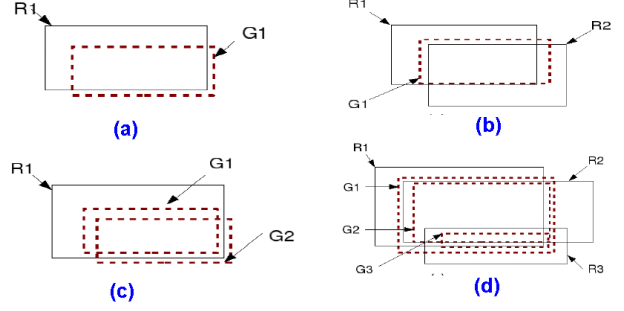


Figure 2. (a) One-to-one (b) One-to-many (c) Many-to-one (d) Many-to-many

- 3) Simple Match: Cardinality of each of Z_g and Z_r is one.
- 4) Multi-match: When $|Z_g|$ or $|Z_r|$ are non-zero and at least one is greater than 1.

D. Options

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

- 1) Type Matching: PETS reads a zone-type attribute 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.
- 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. In case of multi-match, the result zone r_j with best F1 score is chosen for each ground-truth zone to establish correspondence. Figure 3 shows the evaluation result of a document image on our voronoi based segmentation algorithm [2] using GEDI.
- 3) Merge: If used with *merge* option, the MST first calculates overlaps based on precisions and creates a result zone set $Z_r^{g_i}$ for each ground-truth zone g_i . It then calculates recall of $Z_r^{g_i}$ for each ground-truth zone. If both the overlaps satisfy the constraints, the set of merged result zones $Z_r^{g_i}$ are said to have detected the given ground-truth zone g_i .

$$Z_r^{g_i} = \{z \mid z \in R \wedge \rho_z > \tau_1\} \quad (7)$$

$$\text{Merged} - n - \text{Detected} : \sigma_{Z_r^{g_i}} > \tau_2 \quad (8)$$

where τ_1 and τ_2 are user-specified thresholds.

- 4) Ignore: Ignore flag ignores all those result zones whose precision and recall are zero, i.e. if they don't overlap with any ground-truth zone. Figure 3 explains the scenario.

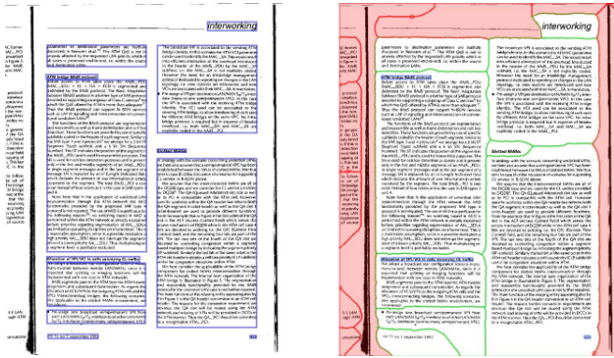


Figure 3. (a) visualization of ground-truth zones (b) visualization of evaluation file from PETS. Polygons, shaded polygons and shaded rectangles depict matched, false alarms and missed zones respectively. In case of 'ignore' option, zones not marked in ground-truth (polygonal shaded regions in evaluation file) are not counted as false alarms.

IV. CONCLUSIONS AND FUTURE WORK

We have presented PETS (Performance Evaluation Tools) for zone detection and classification algorithms, developed for and being used by the document analysis research community. PETS is a polygonal zone 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 zones. 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 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. The PETS software is available at [15] or by contacting the authors.

ACKNOWLEDGMENT

The partial support of this research by DARPA through BBN/DARPA Award HR0011-08-C-0004 under subcontract 9500009235, the US Government through NSF Award IIS-0812111 and the Korean Ministry of Information and Communication is gratefully acknowledged.

REFERENCES

[1] B. Yanikoglu and L. Vincent, "Ground-truthing and benchmarking document page segmentation," *In Proc. 3rd Int'l Conf. on Doc. Analysis and Reco.*, vol. 2, p. 601, 1995.

[2] M. Agrawal and D. Doermann, "Voronoi++: A dynamic page segmentation approach based on Voronoi and Docstrum features," in *Proc. 10th Int'l Conf. on Doc. Analysis and Reco.*, 2009, pp. 1011–1015.

[3] F. Shafait, D. Keysers, and T. M. Breuel, "Performance comparison of six algorithms for page segmentation," in *7th IAPR Workshop on Document Analysis Systems*. Springer, 2006, pp. 368–379.

[4] Y. Li, Y. Zheng, D. Doermann, S. Jaeger, and Y. Li, "Script-independent text line segmentation in freestyle handwritten documents," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 8, pp. 1313–1329, 2008.

[5] L. Likforman-Sulem, A. Zahour, and B. Taconet, "Text line segmentation of historical documents: A survey," *Int. J. Doc. Anal. Reco.*, vol. 9, no. 2, pp. 123–138, 2007.

[6] Y. Wang, I. T. Phillips, and R. M. Haralick, "Document zone content classification and its performance evaluation," *Patt. Reco.*, vol. 39, no. 1, pp. 57–73, 2006.

[7] J. Kanai, T. Nartker, S. Rice, and G. Nagy, "Performance metrics for document understanding systems," in *Ann. Symp. on Doc. Anal. and Info. Retrieval*, 1993, pp. 424–427.

[8] J. Kanai, S. V. Rice, T. A. Nartker, and G. Nagy, "Automated evaluation of OCR zoning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 17, no. 1, pp. 86–90, 1995.

[9] S. Mao and T. Kanungo, "Automatic training of page segmentation algorithms: An optimization approach," in *Proc. of Int'l Conf. on Patt. Reco.*, 2000, pp. 531–534.

[10] B. Yanikoglu and L. Vincent, "Pink Panther: A complete environment for ground truthing and benchmarking document page segmentation," *Patt. Reco.*, vol. 31, no. 9, pp. 1191–1204, September 1998.

[11] J. Liang, R. Rogers, R. M. Haralick, and I. T. Phillips, "UW-ISL document image analysis toolbox: An experimental environment," in *Proc. 4th Int'l Conf. on Doc. Anal. and Reco.*, 1997, pp. 984–988.

[12] W. Abd Almageed, M. Agrawal, W. Seo, and D. Doermann, "Document-zone classification using partial least squares and hybrid classifiers," in *Proc. Int'l Conf. on Patt. Reco.*, 2008, pp. 1–4.

[13] S. J. M. Roth and D. Doermann, "GEDI: Ground truth editor and document interface," in *Summit on Arabic and Chinese Handwriting Reco.*, 2006.

[14] E. Zotkina, H. Suri, and D. Doermann. GEDI: Groundtruthing Environment for Document Images (software). [Online]. Available: [http://lamp.cfar.umd.edu \(Media Group/Research/GEDI\)](http://lamp.cfar.umd.edu (Media Group/Research/GEDI))

[15] W. Seo, M. Agrawal, and D. Doermann. PETS: Performance Evaluation Tools (software). [Online]. Available: [http://lamp.cfar.umd.edu \(Media Group/Research/PETS\)](http://lamp.cfar.umd.edu (Media Group/Research/PETS))